

HAMMERSPACE

如何利用并行NFS v4.2 改造传统基础设施 以支持基于GPU的AI/DL工作负载

技术综述

By Floyd Christofferson

引言

如今，许多行业正竭力探索如何释放人工智能与深度学习（AI/DL）技术的潜力——从非结构化数据中挖掘隐藏价值以优化业务。然而核心挑战在于：**如何在不将数据复制至昂贵新型高性能存储孤岛的前提下，满足基于GPU的AI/DL工作负载的性能需求。**

本文将揭示当今全球数据中心如何利用已部署的标准化组件（内置于Linux系统），通过支持Flex Files的并行NFS v4.2（pNFS）协议，在已有IT基础设施上构建高性能的并行文件系统。

您将看到，相较于传统的pNFS v4.1“文件”布局方案，以及目前在高性能计算（HPC）数据中心中使用的其他商业和开源并行文件系统，pNFS v4.2结合Flex Files布局在AI/DL工作负载上的性能实现了显著提升。

此外，您还可以了解到，Hammerspace软件如何独特运用pNFS v4.2并行机制，在极端性能水平实现GPU数据自动化编排。例如满足Meta线性扩展需求——为其AI研究超级集群24,000块GPU提供超高吞吐性能（支撑着Llama 2 & 3大语言模型）。Hammerspace通过标准pNFS v4.2实现了这一目标，仅仅基于Meta现有存储服务器基础设施及标准网络，无需专属客户端软件或对存储进行改造。

技术变革：pNFS v4.2为何成为AI/DL工作负载首选

NFS （网络文件系统）自1980年代诞生以来，已经成为了存储行业的普适标准。数十年来，所有厂商的网络附加存储（NAS）平台均原生支持NFS。而Linux发行版也全面集成了NFS客户端与服务端组件，使其成为业界最可靠的标准之一。

然而尽管使用者广泛遍布于全球，但NFS从未被视为与高性能相关联。2011年推出的初代并行实现方案：pNFS v4.1加深了这种认知局限。虽然优于标准NFS，但pNFS v4.1存在架构缺陷，严重限制性能并阻碍了其在HPC环境的高性能场景得到更广泛的应用。取而代之，Lustre、GPFS等并行文件系统因满足HPC场景极致性能需求而被广泛采用。





企业中高性能AI/DL场景面临的问题在于，HPC文件系统需要特定硬件和昂贵的InfinBand网络基础设施，而且除非通过文件网关，否则无法兼容标准NFS和SMB文件访问。此外，这些并行文件系统极其复杂，需要定制客户端软件，并且需要专业技术团队进行维护。

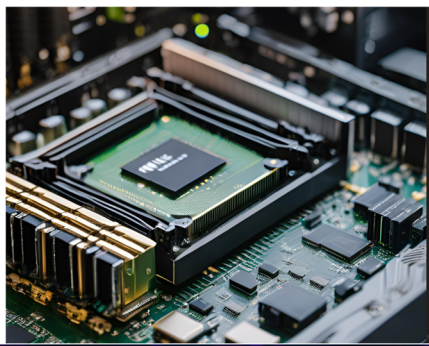
这一局面在2019年8月发生了根本性改变，当时pNFS v4.2与Flex Files被引入RHEL 7.7及其他标准Linux发行版，彻底改变了协议的整个架构。这是首次，标准Linux中包含了一个性能达到或超越全球最快HPC文件系统的并行文件系统。pNFS v4.2结合Flex Files——通过基于标准的架构实现了可谓是极端的线性扩展能力，这意味着它兼容通用服务器和标准以太网，并能够整合来自任何厂商的任意组合的通用NFSv3存储，构建高性能并行文件系统。

同样重要的是，由于pNFS v4.2的客户端和服务器组件均已内置在企业所使用的每款Linux发行版中，IT人员已具备管理该文件系统的技能，无需在现有系统上安装任何专有客户端或服务器端软件。除了性能显著提升外，用户仍可通过标准文件协议访问文件系统，甚至有可能他们都并未意识到：现存的IT环境已因pNFS v4.2而发生根本性变革。

pNFS v4.2与Flex Files：基于标准的并行文件系统

为满足基于GPU的AI/DL工作负载所需的极致性能，pNFS v4.2不仅包含了其他HPC并行文件系统中的关键特性，还具备诸多有区别于这些文件系统的其他能力。它也是对2011年发布以来pNFS v4.1的重大改进，后者至今仍被部分厂商产品所采用。

正如您将在下文关键能力中看到的，pNFS v4.2所有改进的设计目标，是为了实现HPC并行文件系统的极致性能，且同时仍然利用企业数据中心中已广泛存在的标准硬件和网络。**换句话说，我们的目标并非为：创建另一个被厂商锁定而需要特殊网络、专用硬件或专有客户端软件的存储孤岛，而是实现在现有数据中心架构中，高性能与标准兼容性的结合，并将必要的关键能力引入NFS标准，以消除阻碍其在高性能场景中部署的性能瓶颈。**



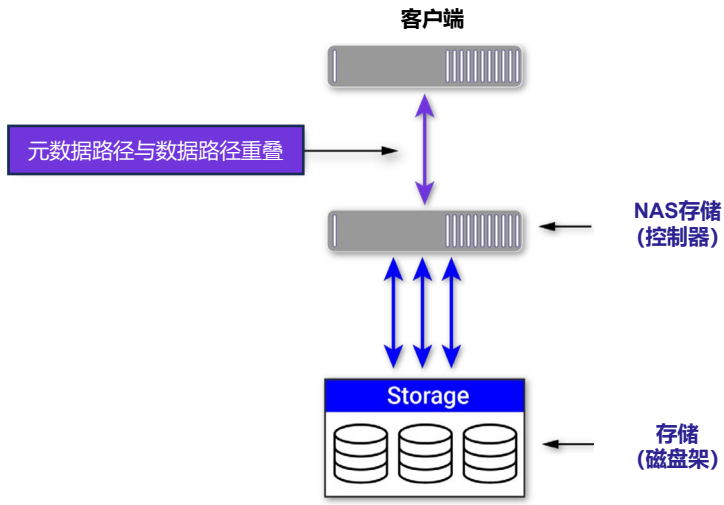
如需了解更多-可联系
ralphie.yu@hammerspace.asia
或 18810311799
手机・微信同号
HAMMERSPACE

pNFS v4.2核心优势及解决的问题：

1. 问题：客户端-服务器I/O瓶颈

限制所有传统NAS架构性能的根本症结在于：元数据与数据必须挤在一条传输路径上。

这在高负载条件下就会形成瓶颈，导致延迟增加、性能下降。同时也严重制约存储节点的扩展数量（无法线形扩展）。
即便是宣称支持AI场景所需的最先进横向扩展NAS架构，也存在同样致命缺陷。因此当存储节点达到数十个时，最先进的横向扩展NAS平台性能都将逐步恶化。

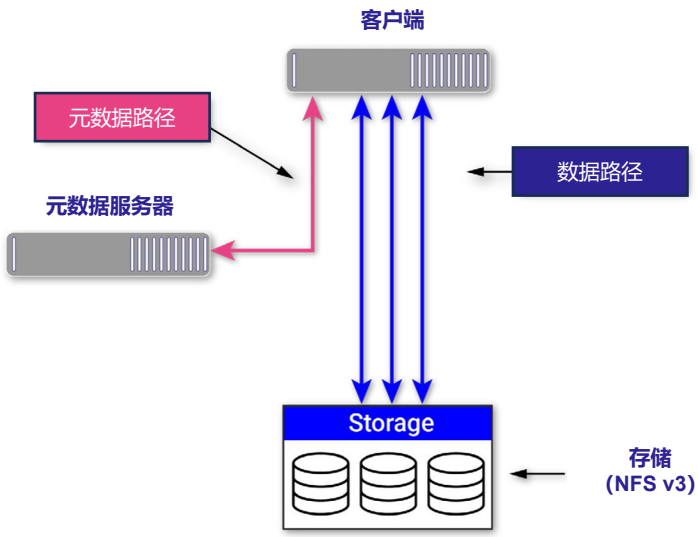


右图：传统NAS架构在单一客户端中造成元数据与数据操作瓶颈，限制了性能和扩展性。

解决方案：分离元数据路径与数据路径

各类并行文件系统——均通过创建完全独立于数据路径的元数据控制层，来解决此性能瓶颈。通过分离元数据操作与数据I/O，客户端可直接从存储设备访问数据，无需经元数据服务器中转，从而提升数据访问速度，并减少资源争用。

此架构同时增强了扩展性，使系统在pNFS v4.2支持下可扩展至数千个存储节点，且不会压垮元数据服务器。下文将通过Meta等场景案例，具体展示该特性及其他改进如何实现AI/DL工作负载的极致扩展能力。



右图：pNFS v4.2分离了元数据路径与数据路径，消除I/O性能瓶颈，实现并行I/O与极限的线性扩展。

pNFS v4.2可突破前述瓶颈结构限制，实现对商用横向扩展NAS存储的加速。Hammerspace现有客户正利用这一协议并行机制，使所有客户端通过直接I/O路径，连接所有存储节点。

无需改造已有底层NAS架构，pNFS v4.2即可消除服务器与存储节点间串扰及缓存争用（上页所示），使已有NAS存储性能提升高达原生能力的3倍。详见： https://hammerspace.com/hyperscale-nas

2. 问题：NFS及pNFS v4.1通信冗余瓶颈

标准NFS及pNFS v4.1的严重性能制约，是源于高开销的元数据交互——即使常规操作也需通过大量getattr（获取属性）等请求在元数据服务器与存储节点间同步状态。

例如NFS基础文件写入操作，至少需要3次与存储节点的元数据往返通信，文件读取则另需2次客户端-服务器往返通信。更严重的是这5次通信严格按照串行执行，系统必须等待前序操作完成确认才能启动后续操作。

由于协议的无状态特性，客户端必须通过持续与服务器往复通信获取状态属性，无法主动感知文件状态。pNFS v4.1的"文件"布局方案也未解决此问题：元数据服务器需在客户端和数据服务器间多次往返更新状态标识（lock/opensateid等），确保通信有效性。

少量大文件场景下该问题尚可容忍，但在海量小文件和混合I/O的AI/DL工作负载中，该缺陷升级为致命瓶颈：系统有超过30%网络带宽消耗于此频繁通信，彻底扼杀高性能与扩展性。

解决方案：pNFS v4.2智能客户端架构

2019版pNFS v4.2通过三项革新根除该问题：

首先实现元数据批量聚合操作，允许多请求并行执行；

后续请求可自动引用前序结果；

构建智能指令链体系，彻底消除客户端-服务器间的往复确认。

此外，pNFS v4.2不再要求客户端持续查询元数据服务器获取文件状态信息，而是将这些状态信息缓存在客户端。实际上，客户端具备了智能感知能力，可持续掌握文件状态而无需反复同步。这一根本性变革消除了NFS及pNFS v4.1系统中，80%的客户端与服务器通信开销，显著缓解元数据服务器瓶颈与延迟，大幅提升性能。

但需注意的是：客户端状态缓存并非简单的页缓存。根据NFSv4.2规范，元数据服务器实质将智能决策权委托至客户端，使其获得自主执行能力。这使pNFS 4.2客户端能够：

- 即时执行任务（无需向元数据服务器申请许可）
- 智能处理错误状态及其他潜在问题（如下文所示）

客户端智能体系的变革意义，在需要并行文件系统扩展能力极高性能的AI/DL环境中尤为显著——它确保了可以持续、高速向GPU集群输送数据的通路。

3. 问题：TCP/IP连接瓶颈

传统IT基础设施适配高性能I/O的另一障碍在于TCP/IP通信协议的单连接通道限制：标准TCP/IP在客户端与服务器间仅建立单连接通道，这意味着数据传输速率被限制在单网络接口带宽内。此限制实质扼杀了pNFS v4.1及其他系统在标准IT环境中的扩展能力：

- 无法在横向扩展系统中实现跨节点负载均衡
- 因缺乏内置冗余机制，网络路径中断将直接导致I/O阻塞

这也是HPC环境采用InfiniBand或通过RDMA改造以太网的原因之一——两者均需专用网卡与交换机支持多连接架构，但此类升级成本高昂且极少应用于企业数据中心。

解决方案：pNFS v4.2多连接协议

pNFS v4.2标准通过引入n-connect特性，支持单存储节点建立多重网络连接通道。该功能可实现：

- **将I/O负载分散至多个标准网络接口**
- **优化可用网络带宽利用率**
- **降低单连接通道阻塞风险**

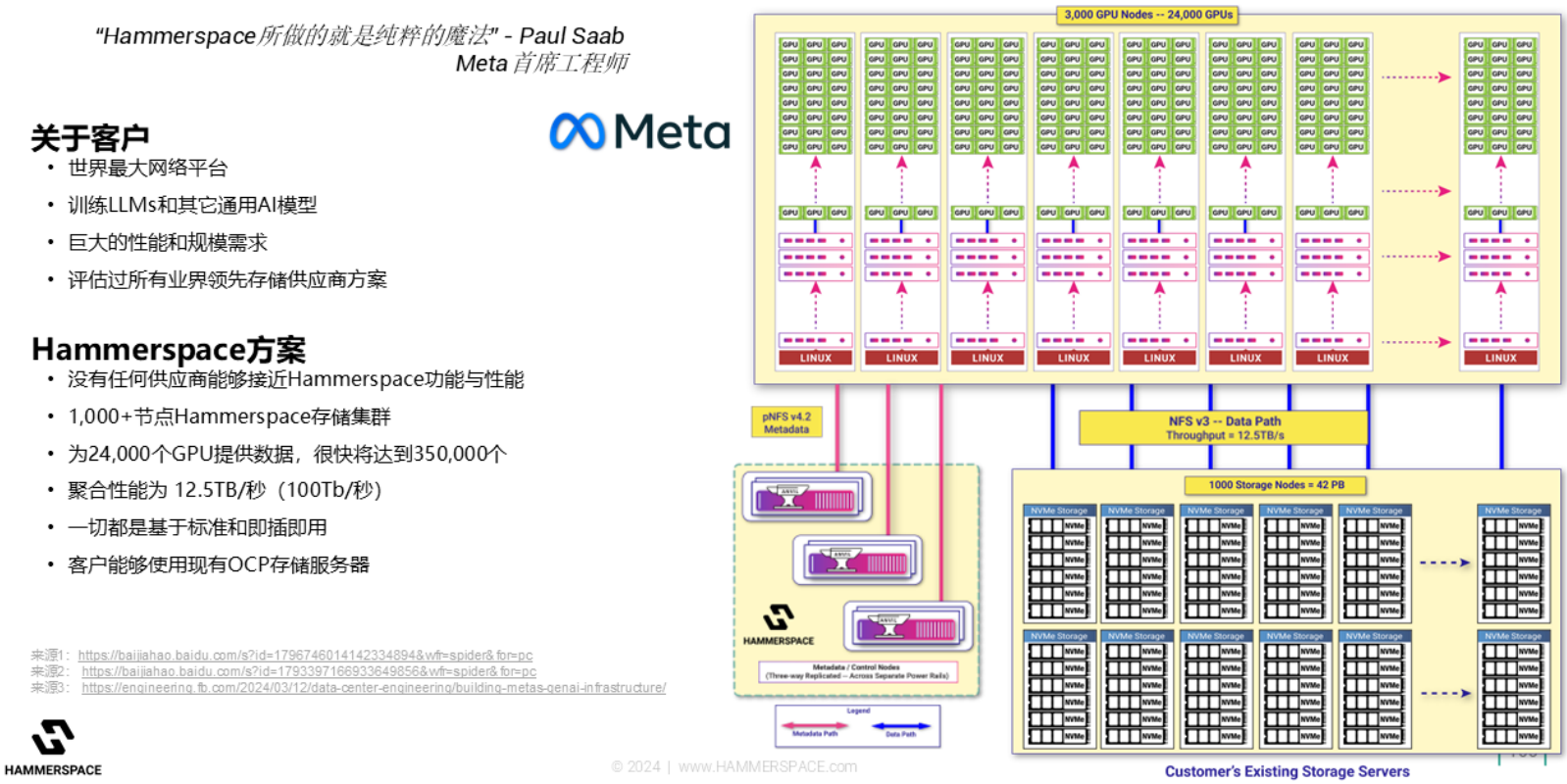
该功能无需改动网卡、交换机或现有标准网络基础设施，即可显著提升系统性能——这些设施普遍存在于各企业数据中心。

更重要的是，结合前文提及的pNFS v4.2客户端智能架构，n-connect特性使得客户端能够动态扩展至极限规模而避免遭遇瓶颈。

而pNFS v4.2进一步强化了n-connect能力：支持以条带化或镜像模式，同时写入多个存储节点，实现I/O负载分流。通过此类同步镜像机制，可在非高可用存储节点间建立冗余。目前社区正推进pNFS规范升级提案，拟引入客户端侧纠删码作为该功能的另一选项。

基于这些能力，智能pNFS v4.2客户端可协同元数据服务器将文件冗余写入多个目标。这意味着即使是由通用服务器组成的阵列也能实现高可靠性，且无需客户端额外配置。

Hammerspace正是使用该技术在Meta AI研究超算集群实现线性扩展：**跨越1,000个存储节点+并行I/O吞吐高达12.5TB/秒（标准以太网环境）+持续满足24,000个GPU数据供给需求**



上图 - Hammerspace通过内置于Linux内核的pNFS v4.2客户端，在通用硬件和标准以太网上为Meta的AI研究超级集群提供极致性能，无需改造现有基础设施或安装专有客户端软件。

4. 在现有NFSv3存储平台上实现并行性能

为实现与任意厂商现有存储的极致兼容性，pNFS v4.2仅要求存储节点支持标准NFSv3连接——该协议已嵌入当今所有NAS平台。这意味着企业可像Meta AI研究超算集群那样，将任意厂商的存储用于高性能并行文件系统架构。

其他并行文件系统不具备这种跨平台使用现有存储的能力，通常需专用服务器和后端存储。

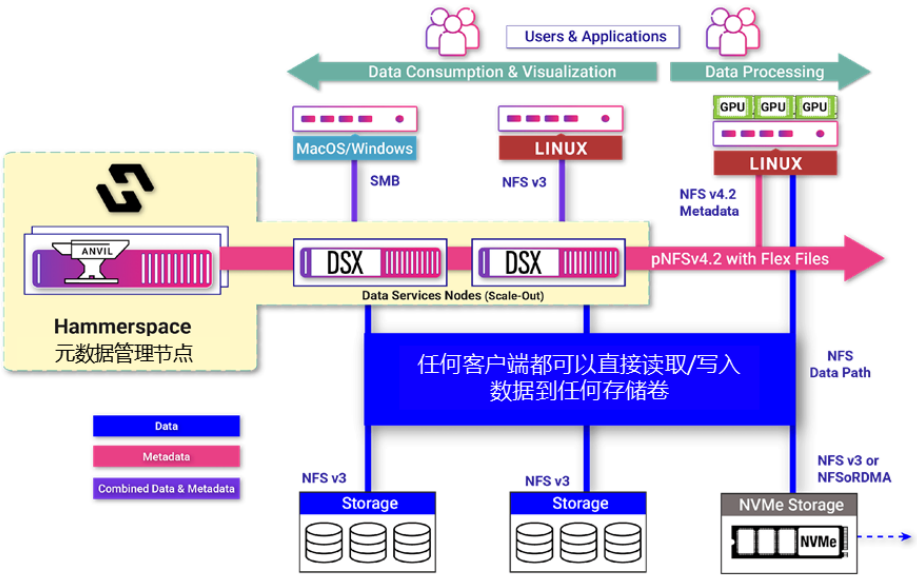
结合前述消除NFS与pNFS v4.1瓶颈的多项能力，pNFS v4.2规范的兼容标准，确保了企业可改造现有基础设施来支撑AI/DL流程，无需重复购置新存储。尤其在AI建设初期，当非结构化数据分散在各类现有存储资源中时，企业可直接利用现有存储中的数据投入工作流，无需额外成本和停机时间将其迁移至新建存储设备中。

5. 灵活文件布局 （Flexible Files Layouts）——pNFS v4.2的核心优势.

pNFS v4.2实现卓越灵活性与可扩展性的关键——在于灵活文件布局的引入。该能力深度利用pNFS v4.2内置的客户端智能架构（详见下图说明）。

Flex Files功能虽强大，实际操作却极为简便：当客户端需要特定数据子集（例如执行分析、供给GPU集群等场景），它会向元数据服务器请求文件布局映射图，精准定位数据所在的存储位置。这些数据子集常分散在多个孤立存储系统中，布局图提供精确访问信息，使pNFS v4.2客户端能以文件级颗粒度跨所有存储系统直接读取数据。

右图 - pNFS v4.2结合Flex Files可将并行文件系统架构连接至任意NFSv3后端存储。Hammerspace进一步扩展了这一能力，使其支持更多类型的后端存储，例如对象存储和磁带存储。



文件布局映射图即时下发给客户端后，客户端可以直接与存储服务器进行读写操作。由于客户端可以智能缓存布局信息，同时消除了pNFS v4.1中反复获取属性（getattr）等状态信息的通信流量，这意味着pNFS v4.2实现了客户端与存储间的直接I/O访问——完全绕过元数据服务器。与v4.1及其他文件系统相比，这可以显著降低延迟并大幅提升性能。

对于必须持续为昂贵GPU集群供给数据的AI/DL场景，该能力消除了延迟并确保GPU高利用率：

- **前项任务完成后，需跨不同存储获取新数据子集的新任务可近乎瞬时启动**
- **pNFS v4.2客户端向元数据服务器请求新任务布局图**
- **即刻获取实时数据位置信息**
- **杜绝数据准备延迟或其他中断，避免GPU闲置等待**

强大的错误处理与容错能力

该架构的精妙之处还在于其极强的容错性：服务器端和客户端的实时遥测数据，会与布局统计持续反馈至元数据服务器。这使得系统能够做到自适应网络拥塞等不利因素，动态更新客户端布局以重定向I/O，或即时进一步并行化I/O以维持目标性能。

这种自愈能力使系统能在任务执行中智能适应实际环境，通过克服网络或存储服务器故障来维持所需I/O。即使单个节点因运行滞缓或性能不足，系统也能实时自动补偿。

这在传统pNFS v4.1实现中并非如此：

- pNFS v4.1虽会向元数据服务器反馈存储节点故障，但客户端遇到的网络状况等外部错误信息将丢失且无法回传
- pNFS v4.1无法将所遇故障状态传回元数据服务器以触发系统自适应，导致客户端I/O直接中断，需IT人员手动排查修复

pNFS v4.2通过始终畅通的客户端与服务器间的反馈环实现了突破：

客户端检测到问题时，元数据服务器会基于客户端/服务器综合遥测实时更新布局，自动指向文件不同副本或绕开故障重定向I/O。

这为pNFS v4.2系统提供了其他方案无法实现的弹性与自愈能力，显著减少业务中断的停机时间，确保持续向GPU输送数据。

6. pNFS v4.2实现了实时数据迁移——在使用过程中透明地进行数据迁移

pNFS v4.2规范的另一项关键特性（通过前文提及的增强功能实现）是：可在文件被读写使用时，实时在其他存储节点上创建新文件实例，且不会中断用户或应用程序。这并非创建文件副本，而是共享相同文件元数据的另一实例化过程，用户对此毫无感知。

该能力对企业IT环境的核心价值在于：数据可透明的在后台跨存储层迁移，或并行供给高性能集群以支持AI/DL流程，同时不中断用户访问，也不产生需要人工合并的多版本文件副本。即使用户在迁移发生时正在使用该文件，也完全无法察觉文件已从一种存储类型迁移至另一种。

除了前文第3条所述为提升性能跨多节点分发数据的并行化能力外，实时数据迁移还意味着：对于数据分级存储、灾备或从旧存储迁移至新存储等操作，其对用户和应用程序完全透明无中断。pNFS v4.2彻底消除了此类操作带来的业务干扰。

而消除副本扩散对控制成本的IT组织至关重要：

- 降低运维成本（IT团队无需管理副本）
- 无需处理符号链接、存根文件等分层存储管理方案中的脆弱组件
- 规避其他类似技术的管理负担

在多个用户及应用程序需访问相同数据（尤其当数据集分散在互不兼容的存储孤岛中）的场景下，实时数据迁移具有变革性意义。IT管理员和数据所有者可在需要时将数据精准部署至目标位置，完全无需担心中断用户或应用程序。

Hammerspace如何利用pNFS v4.2构建全局数据平台

Hammerspace凭借独特优势充分释放pNFS v4.2的潜能——该技术是其软件定义全局数据平台的核心组件，助力客户在跨越存储厂商孤岛、多站点和云环境的现有IT设施中实现HPC级性能。而Hammerspace的核心优势在于其深厚的Linux社区基础：

- **上文所述pNFS v4.2关键技术能力由其工程师历经多年研发**
- **近十年来持续将相关改进推进至Linux内核**
- **自2019年Linux社区正式批准pNFS v4.2规范起持续推进**

与Linux社区紧密整合是Hammerspace的核心基因。首席技术官Trond Myklebust担任Linux NFS内核维护者逾20年，始终确保Hammerspace的开发工作严格遵循社区维护标准。首席高级软件工程师Mike Schnitzer担任Linux内核中设备映射器（DM）子系统维护者。另一位Hammerspace软件工程师Tom Haynes是RFC 8435标准合著者——该标准于2018年定义pNFS弹性文件布局，次年即成为pNFS v4.2规范的核心特性。

基于对标准的专注，Hammerspace近期技术路线图包含更多pNFS规范增强：今年夏季及未来一年将向互联网工程任务组（IETF）提交多项特性，力争纳入未来pNFS版本。

这一策略的优势在于：Hammerspace可无缝扩展pNFS规范能力，为其全局数据平台增添变革性功能，同时保持与开放标准的严格兼容。这使其不仅能桥接各类厂商的存储孤岛，更借助pNFS v4.2实现跨数据中心与云存储的高性能数据编排，构建真正的高性能并行全局文件系统。

对AI/DL场景而言，该架构使Hammerspace能整合互不兼容存储资源的数据，以峰值性能供给超大规模GPU集群，客户无需迁移已有数据或采购被厂商锁定的新存储设备。

以下是Hammerspace基于pNFS v4.2构建全局数据平台解决方案的核心特性摘要，详见Hammerspace官网：

1. 元数据就地同化（Data-in-place metadata assimilation）

该能力可以让Hammerspace无需迁移已有数据至新存储设备之中，即可从存储中快速提取元数据。该强大特性是处理速度极快：即使在Meta的42PB、1,000节点超大规模存储集群中，Hammerspace也能快速整合元数据并立即启用文件操作。

Hammerspace兼容所有厂商存储平台，这意味着整合多源异构存储的元数据后，这些存储将统一纳入同一全局数据环境，彻底消除不同存储类型间的技术壁垒。

2. 跨孤岛、跨站点、跨云的并行全局文件系统

通过整合现有存储的元数据，Hammerspace构建起跨平台的高性能并行全局文件系统。该系统不仅能统一单一数据中心内不同厂商的存储孤岛，更能扩展到跨越多数据中心、多云服务商及多云区域。

Hammerspace全局数据环境可纳入任何存储类型，包括任意厂商的块存储、NVMe、NAS、对象存储、云存储及磁带存储。所有孤岛、站点和云资源通过统一的跨平台全局命名空间呈现。如上一页的第6项所述，Hammerspace提供跨所有存储孤岛/站点/云的实时数据迁移能力，确保文件迁移等操作对用户和应用程序完全无中断。

3. 多协议全局文件系统访问

用户可通过标准NFS、SMB和S3协议即时访问Hammerspace并行全局文件系统。如上所述，无需安装客户端软件，因为系统设计兼容现有文件访问协议。

从用户或应用程序视角看，数据访问方式与安装Hammerspace前完全一致：

- **文件与文件夹结构保持原样**
- **用户无需学习新应用程序或界面**
- **用户甚至可能察觉不到后台运行的Hammerspace**

关键区别在于：用户不再需要跨多个文件共享或存储桶访问不同存储孤岛，而是通过统一全局命名空间无缝访问所有数据——如同全部本地及云端数据已整合至单一的全局超大规模NAS系统之中

4. 自动化数据编排

依托全局覆盖能力与实时无中断数据迁移特性，Hammerspace构建了文件级颗粒度的强大数据编排体系。结合贯穿本文所述的pNFS v4.2极致性能，系统可实现：

- **后台自动迁移数据至任意位置**
- **满足数据中心内部/跨多站点/多云的海量需求**

应用场景示例：向本地或远程GPU即服务集群供给AI/DL任务时，即使数据分散在多个孤岛和位置，仍可在后台自动完成。所以，数据保护机制同样实现了革新：

- **通过同等无中断编排能力执行远程站点/云复制**
- **区别于备份软件或Robocopy/RSync等手动流程**
- **复制生成的文件副本可纳入全局自动化策略统一管理（非隔离存储）**

5. 可编程元数据

除了从现有存储整合元数据外，Hammerspace全局数据平台还支持多种自定义的元数据类型。这些可以是标签、标记或其他引用，帮助数据所有者对数据进行分类。

依托并行全局文件系统的跨平台覆盖，此类自定义元数据标签，可以在数据随时间迁移至不同位置时持续有效，实现精准跟踪和基于策略的操作。

此外，此类自定义标签可通过受控词汇表的自动应用，消除用户遗漏标记的风险。为文件夹设置引用项目ID、成本中心或其他变量的自定义标签，随后存入该文件夹的任何数据将自动继承该标签。

这使得构建数据编排工作流成为可能，可引用此类自定义元数据标签来确定特定数据类别的数据保护规则、追踪成本报告信息，或在AI/DL工作流中，仅将已清洗并正确标记的数据提取至高性能存储层供给GPU集群。

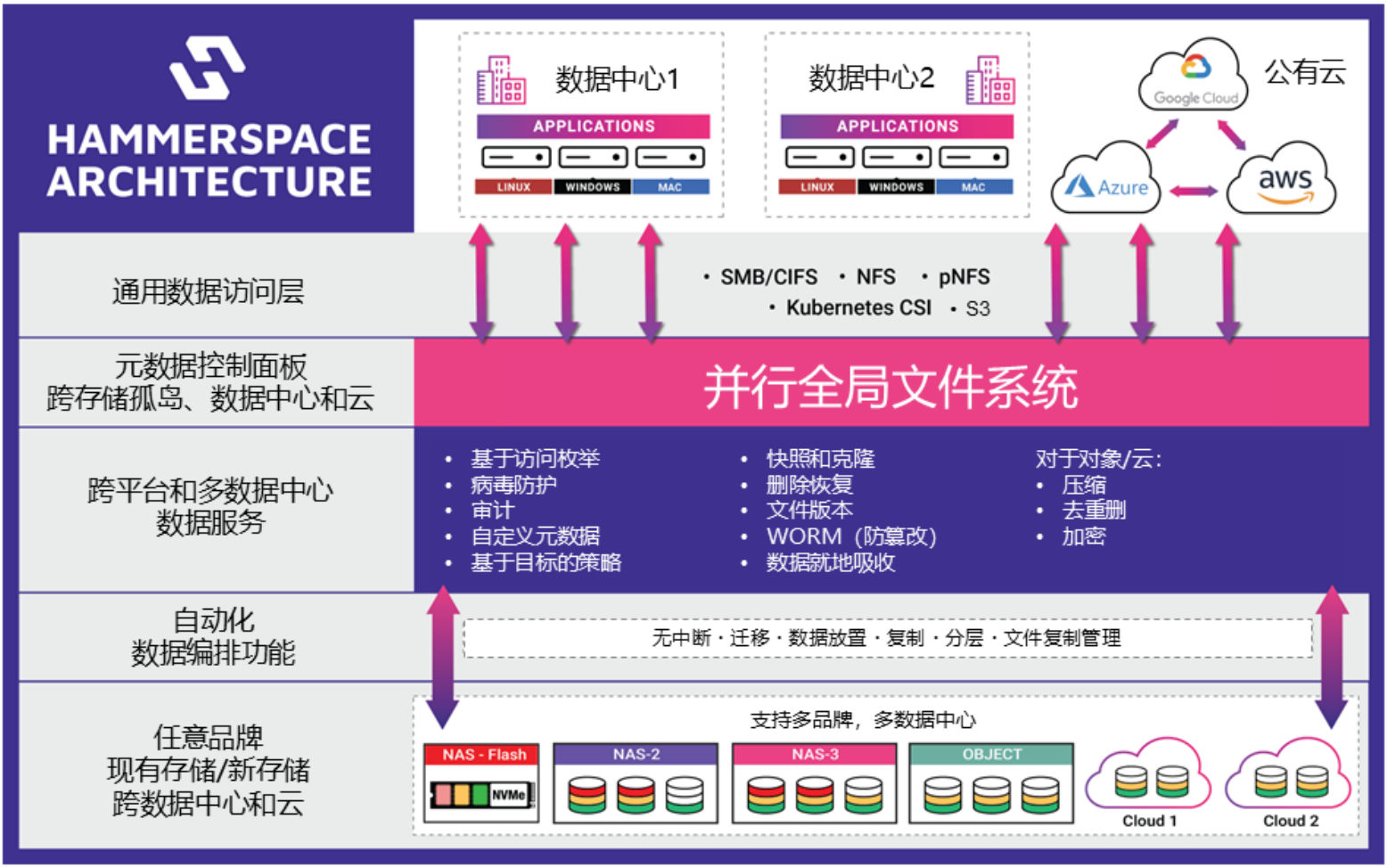
由于与Hammerspace中所有功能一样，此操作在文件级别的颗粒度执行，意味着即使深藏在独立文件系统层级中的子数据集也能被全局处理，无需其他解决方案所需的全量数据迁移。

6. 跨平台数据服务

跨孤岛、站点和云的无中断数据编排关键优势在于，Hammerspace可自动化实现全局的企业级数据服务。这些包括企业级NAS平台上非常常见但在并行文件系统中普遍缺失的数据服务。

数据保护服务，如全局快照与克隆、文件版本控制、误删除恢复、一次写入多次读取（WORM）、全局审计、反病毒防护等，均可在文件级别的颗粒度实现自动化。数据放置服务如分层存储、数据迁移和灾难恢复，也均可由元数据变量（包括自定义元数据标签）的任意组合驱动实现自动化。

Hammerspace包含基于目标的策略引擎，使管理员能够使用简单语言编写的业务规则来管理不同数据类别。从不同数据类型的数据保护级别、根据底层存储成本制定数据放置策略，到更多其他功能，均可通过Hammerspace在整个全局数据环境中实现自动化。



上图- Hammerspace全局数据平台软件功能层的逻辑视图。Hammerspace作为单一软件包，可以安装在通用服务器、虚拟机或云端实例上。

总结

直到近期人工智能/深度学习（AI/DL）爆发之前，大多数企业采用性能较低的NFS、SMB及S3协议即可满足需求。但AI/DL引发的变革，以及向GPU集群输入数据所需的高性能计算（HPC）级性能，彻底改变了这一局面。

pNFS v4.2的强大功能使各类组织能够即刻利用GPU集群。该协议支持企业直接从现有存储中整合数据来供给AI/DL工作负载。企业无需浪费预算构建定制化新存储平台来迁移数据，而是可充分利用现有系统。预算得以聚焦于成果产出，而非资源重复消耗。Hammerspace基于此标准化基础，通过数据编排等特性增强了pNFS v4.2的能力，使AI/DL工作负载可以无缝桥接多个孤岛和站点，包括将突发工作负载扩展至远程或云上GPU集群。

pNFS v4.2作为已嵌入全球各数据中心Linux内核的基础，成为了业界的根本性变革力量：它能在现有数据中心架构中提供并行文件系统性能，加速基于GPU计算在各地快速普及。