如何满足现今高性能计算工作负载的数据存储与移动需求

解决方案综述

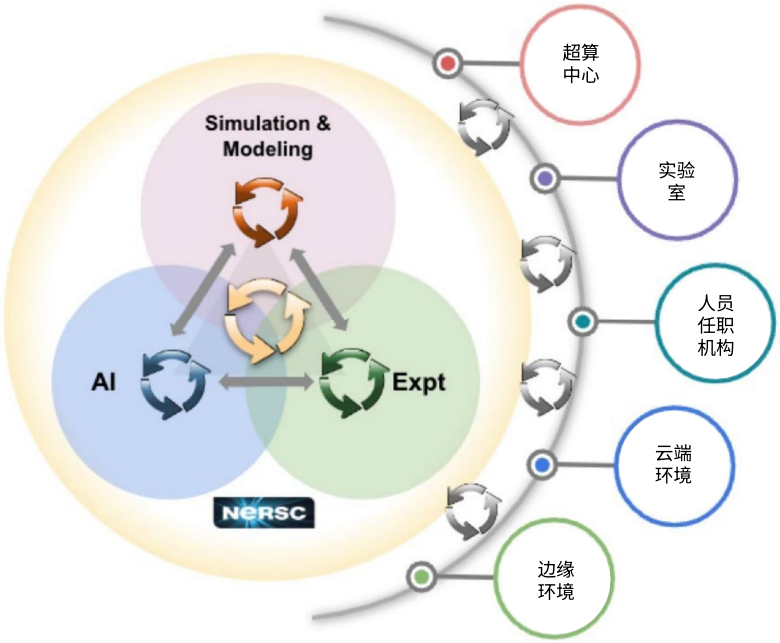
综述

高性能计算（HPC）的用例、应用场景及工作流模式正在快速演变。HPC的应用范围已从政府与学术界扩展至了工业领域。当前的HPC系统需处理**异构工作负载**，包括实时任务与事件触发型用例。用户、数据及算力资源分布于不同地理位置，数据必须满足以下全程要求：

- **可流动性**：从仿真建模到可视化、分析与归档的全阶段无缝迁移
- **可检索性**：支持跨地域的数据搜索与定位
- **可共享性**：打破物理边界实现协同访问

人工智能与机器学习（AI/ML）与工作流的深度集成，正在催生新的洞察能力。容器化与自动化技术使工作负载具备**跨平台可移植性**及**资源突发扩展能力**。

这些HPC领域的变革对**数据存储**与**数据移动性**提出了更高要求。Hammerspace通过将**高性能全局并行文件系统**与**数据编排能力**相结合，构建出独特的全局数据环境，可以有效应对上述挑战。

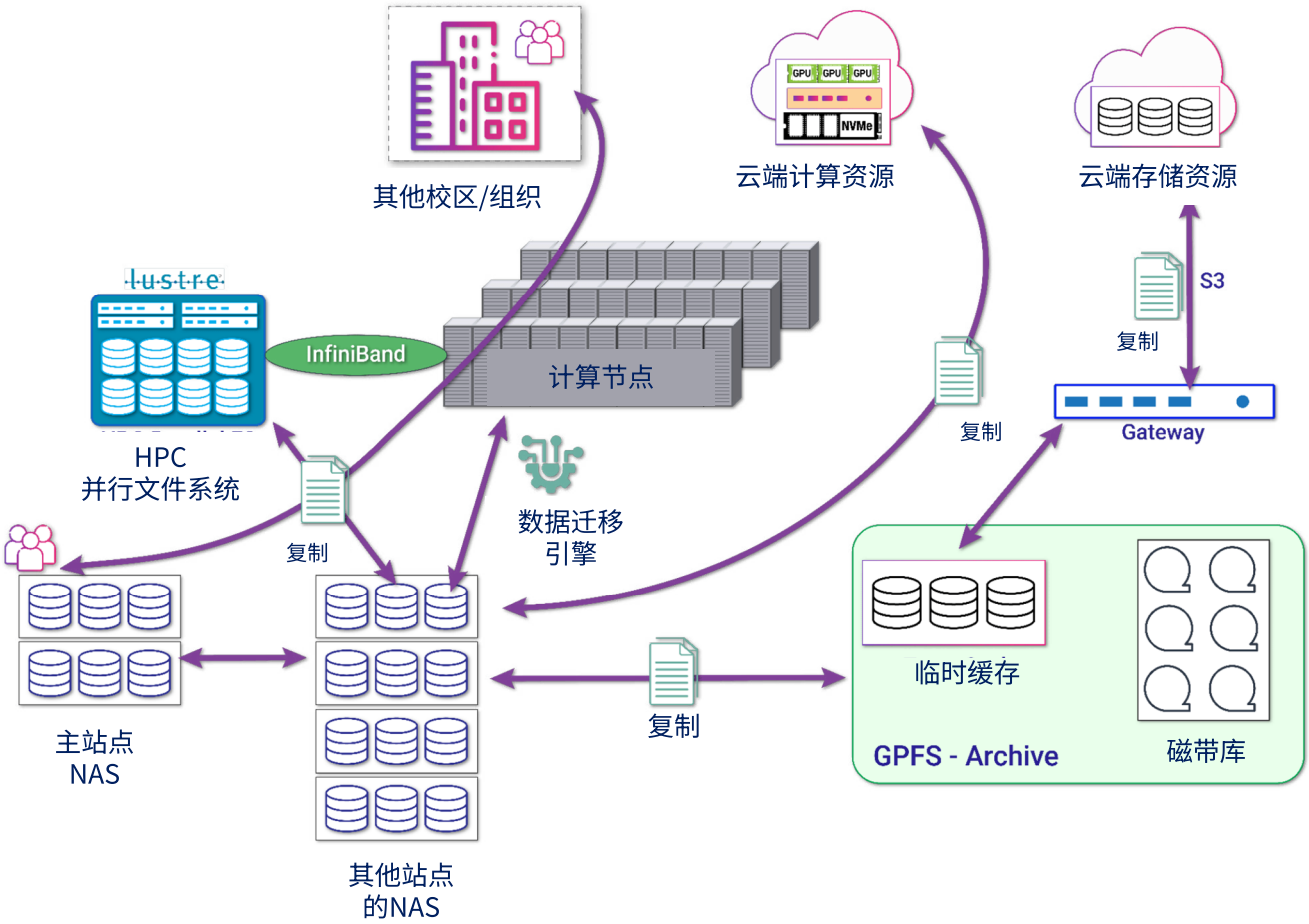


上图 – 融合化的科研工作流示例，NERSC

解决方案概述

传统高性能计算（HPC）环境中存在来自不同厂商的多重存储孤岛。**专用并行文件系统服务于高性能临时存储，横向扩展NAS处理作业工作流的前后阶段，云存储用于云爆发和协作，长期归档数据则驻留在对象存储或磁带。**管理数据访问和保护需要为每个孤岛配备单独的工具，而协调孤岛间的数据流动又需额外增加软件层。数据经常在不同文件系统间复制，既浪费时间又消耗存储容量。

https://www.nersc.gov/assets/NERSC-10/Workflows-Archetypes-White-Paper-v1.0.pdf

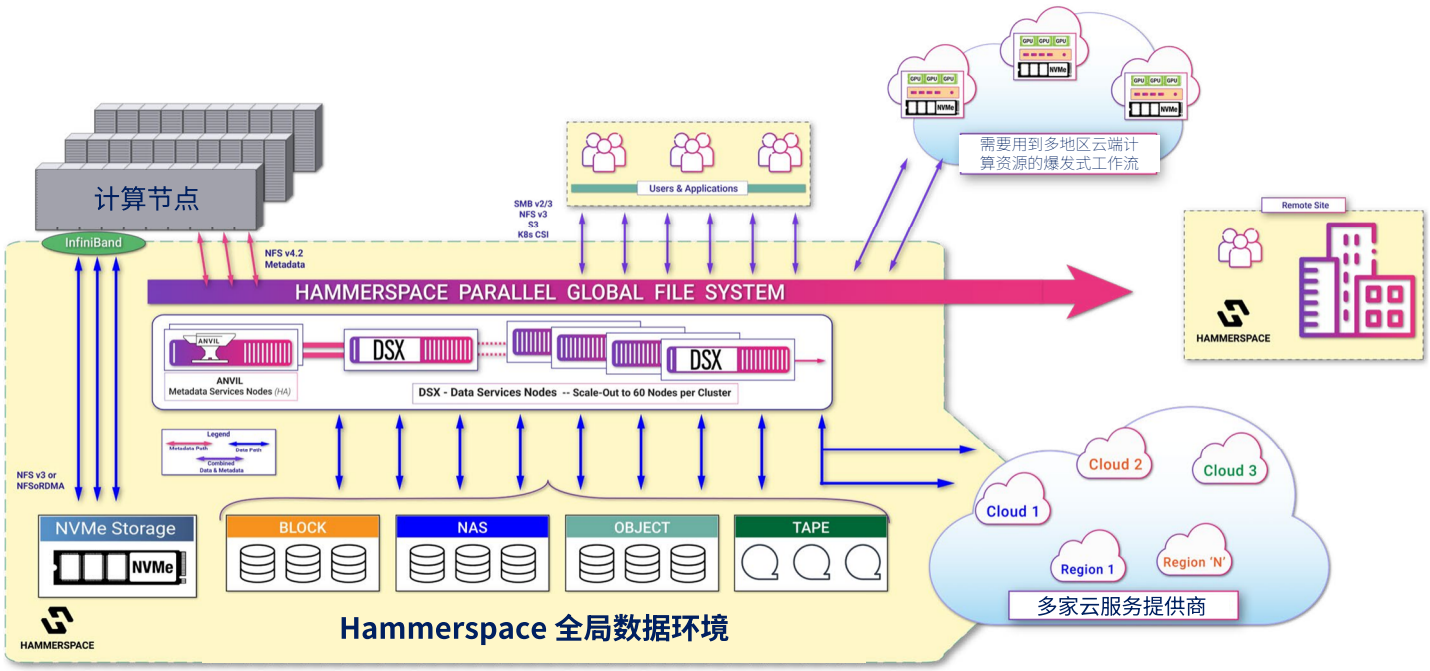


上图-Hammerspace问世前，高性能计算存储环境的普遍架构

而Hammerspace软件可以帮助创建一个全局数据环境，通过以下关键方式简化和优化HPC场景下的数据管理：

- 整合多个孤岛、站点和云
- 在异构存储系统和站点间，实现数据服务标准化
- 通过基于目标的策略，自动化完成数据编排
- 通过基于pNFS的超大规模NAS创建高性能并行文件系统

将AI/ML集成到高性能计算（HPC）工作流中实现了新的能力和洞察，但也会给数据存储基础设施带来新的压力。



上图- 基于Hammerspace搭建的高性能存储环境

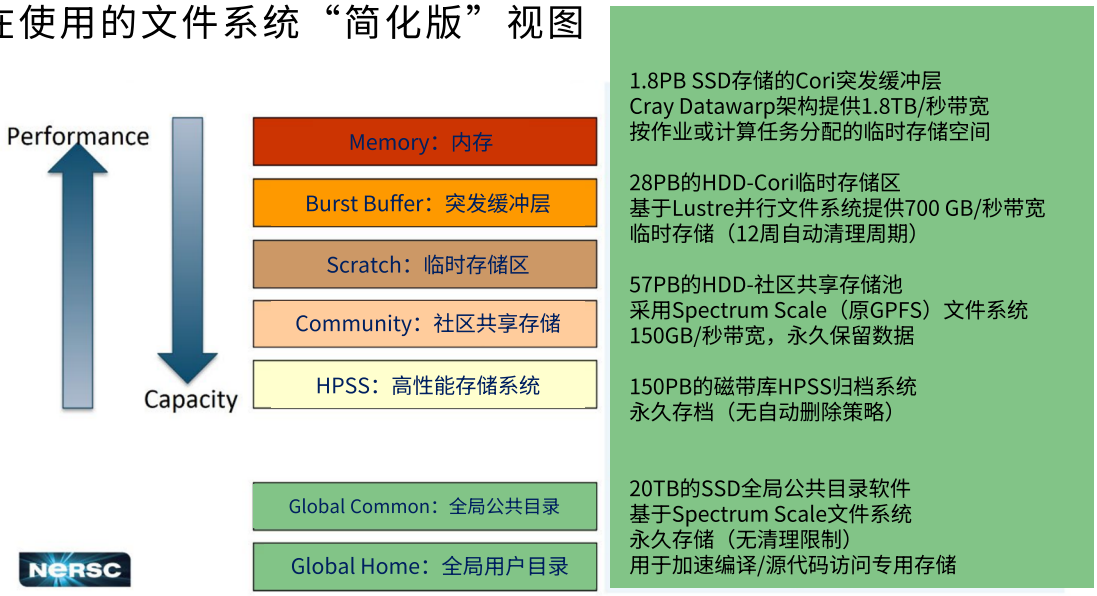
解决方案优势

并行文件系统级别性能且不存在厂商锁定	横跨存储技术、站点和云端的全局文件系统	掌控数据存放位置以满足工作流程需求
・创建高性能并行文件系统，用于数据的临时处理及预处理/后处理，支持使用任意厂商或混合厂商的存储服务器构建而成 ・用标准Linux并行网络文件系统（pNFS）替代专有文件系统 ・启用RDMA和GPUDirect Storage以实现最大性能	• 提供对所有数据的访问，无论其位于何处 • 通过云存储实现协作，或通过突发工作负载的模式，将任务扩展至云端 • 减少管理负担，为研究工作留出更多时间	• 就地存放——无需冗长的存储迁移流程 • 根据策略目标自动放置并迁移数据 • 分配自定义元数据以支持组织架构与自动化流程

存储孤岛的整合

高性能计算（HPC）环境，传统来说都需要整合来自多个厂商的文件系统，每个系统都具备独特属性。例如，下图描绘的劳伦斯伯克利国家实验室高性能运算中心（Lawrence Berkeley NERSC）正在使用的文件系统的“简化版”视图。图中展示了五套不同的文件系统（来自四个不同厂商），用于满足不同的场景需求。

NERSC正在使用的文件系统“简化版”视图



上图-劳伦斯伯克利国家实验室高性能运算中心-HPC文件系统

大多数甚至全部此类场景需求，均可通过单一Hammerspace环境而实现。系统仍会存在多个导出点及多个存储系统，但所有组件均会被纳入Hammerspace全局数据环境中。这种方式减少了管理节点的数量，同时增加了高价值功能。

传统厂商的存储整合方案通常提供一个更先进、更大规模的存储孤岛来替代多个现有孤岛。这种物理整合迫使企业每隔几年就必须执行耗时且高风险的数据迁移——需将所有现有文件复制到新存储中。在HPC环境中，此类物理存储整合方案并不适用。暂且不论迁移数PB量级数据的痛苦——根本不存在能同时满足每个HPC存储层级性能、规模与成本需求的单一存储系统。

而Hammerspace并非采用物理整合，而是对存储孤岛进行逻辑整合。**通过称为同化（Assimilation）的过程，**所有存储资源被纳入统一管理框架。整合过程将所有文件系统元数据导入Hammerspace，而数据仍保留在现有存储中。

与通常耗时数月至数年的PB级别的数据迁移相比，同化过程仅需数小时至数天。例如，最近完成的一个40PB数据量用户的同化过程仅耗时两天。更值得注意的是，同化任务可以采用后台任务与按需执行相结合的方式最大限度降低业务中断——一旦启动任务，所有数据即可通过Hammerspace访问，如同同化任务已经完成。

Hammerspace的逻辑同化方案，在承认多类型存储必要性的同时，提供了物理存储整合层面的多种优势。诸如多站点全局命名空间、云存储支持等附加功能，通过消除对单一用途网关产品的需求进一步创造价值。在Hammerspace中可创建多个导出点或共享目录，但这些抽象实体可跨越多个存储系统。随时可被添加新存储容量，数据向新存储迁移将作为后台编排任务执行。存储系统可能会进行更替变化，但用户与应用所见的实体——导出点与共享目录——将保持稳定，确保业务连续性与用户生产力。

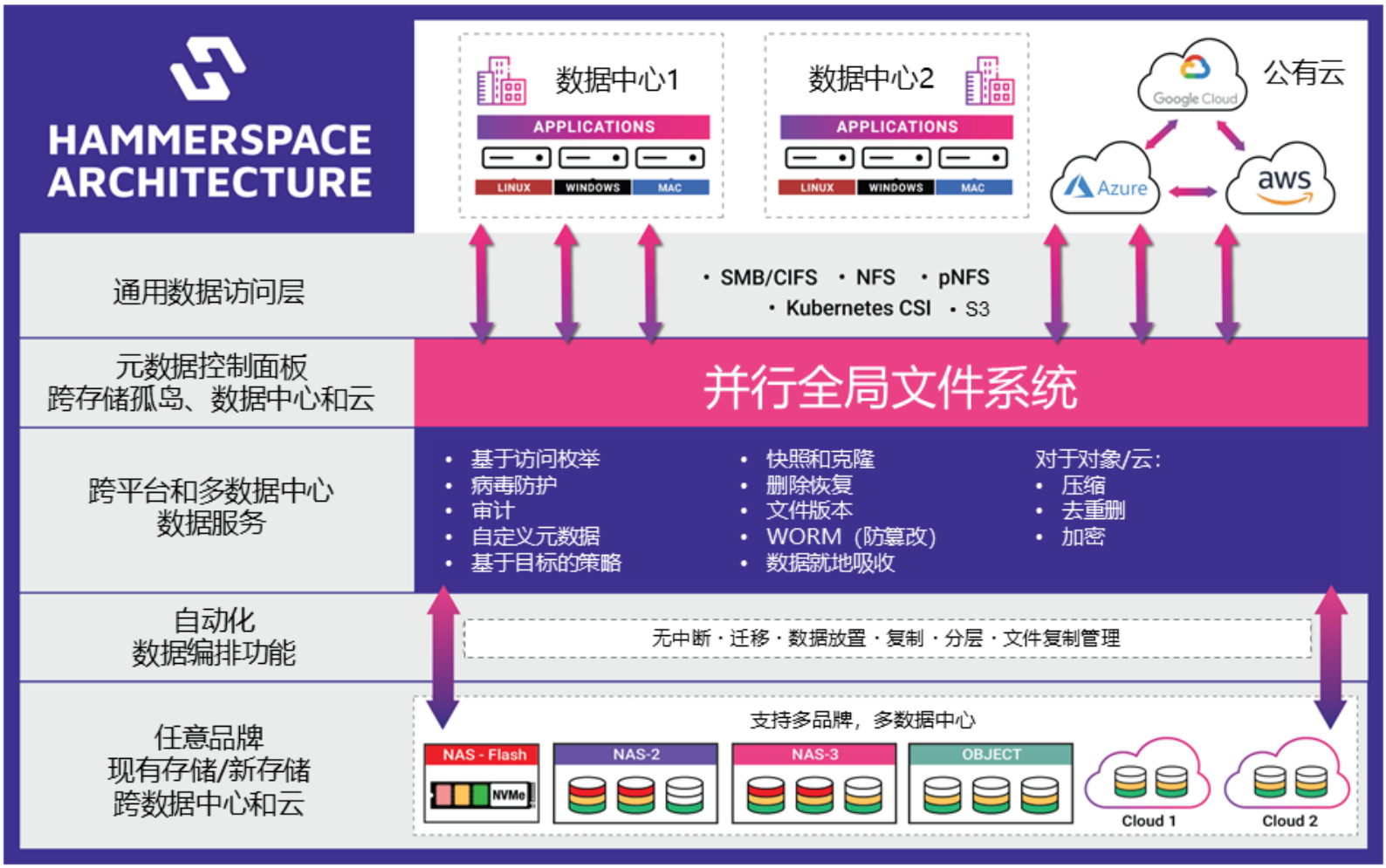
标准化的数据服务

由于Hammerspace将文件系统与底层存储解耦并提升至存储层之上，快照与克隆、审计、版本控制、数据存放位置等数据服务得以跨所有存储及所有站点统一实现。

拥有一套统一工具集来跨异构存储系统管理数据保护与安全性，可显著降低管理负担。这对学术与科研环境尤为有益——为研究活动释放更多时间。同时，通过减少错误或由于配置不当而未被察觉事件的概率，也可以进一步降低系统中数据丢失风险。

下图列举了Hammerspace能够提供的全局数据服务。

下图 - Hammerspace 架构





自动化的数据编排

Hammerspace并行全局文件系统可横跨从闪存到磁带、NAS到对象存储及云的各类存储技术。从高性能暂存存储到归档存储的所有层级均属于同一数据环境。这使得研究人员能够专注于数据需求，而非存储系统的维护工作。

一旦数据受Hammerspace管控，即可通过自动化数据编排实现迁移、复制、分层、保护及智能放置。可配置的策略目标会以无中断的方式完成这些任务。用户和应用程序始终可访问文件——即使文件正处于迁移过程当中——使数据迁移与存放流程实现完全的透明化。

文件与目录被添加自定义的元数据标签，这些标签将跟随数据穿越全局数据环境，无论数据存储于何处。标准元数据与自定义元数据均可用于驱动策略目标和工作流自动化，或根据项目、研究组属性等追踪文件与目录。

Hammerspace内置的数据编排功能消除了对数据迁移软件、归档系统及云存储网关等独立工具的需求。

超大规模NAS架构（Hyperscale NAS）

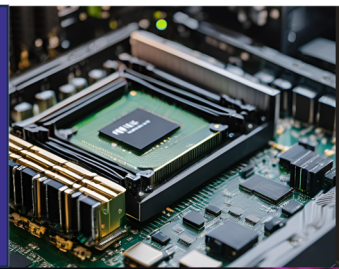
Lustre®和IBM Storage Scale™（原GPFS™）等专用HPC并行文件系统通过分离元数据与数据实现高性能与可扩展性。该架构使客户端机器可直接与存储系统交互，但都需要安装定制的客户端软件。

Hammerspace采用并行NFS（pNFS）的Flex Files布局类型，仅使用Linux内置的标准NFS客户端即实现同等架构，并获得相同性能与可扩展性优势。这种超大规模NAS架构实现了线性扩展的并行吞吐量，性能仅受存储与网络硬件限制。

超大规模NAS架构的开放性可延伸至NFS服务器，支持使用任何兼容NFSv3的存储系统构建并行文件系统——包括运行标准Linux的白牌服务器或OCP服务器。

尽管超大规模NAS架构无需RDMA即可实现高性能，但仍支持基于InfiniBand和RoCEv2的RDMA。我们的超大规模NAS实现已通过NVIDIA GPUDirect Storage认证，可在需要GPU的工作流中提供极致性能。

一旦启动了数据同化任务，所有数据即可通过Hammerspace进行访问，就好像同化任务已经完成一样。



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