Hammerspace 超大规模NAS 配置示例

技术综述

介绍

本文档旨在帮助Hammerspace的合作伙伴、用户和潜在客户理解Hammerspace超大规模NAS裸机部署的示例。本文并非完整的参考架构指南，也不能替代专业咨询。每个系统设计必须由Hammerspace解决方案架构师参与，以确保满足客户需求。

超大规模NAS（Hyperscale NAS）综述

超大规模NAS是一种共享存储架构（而非独立产品），具有两大核心特征：

- 通过标准pNFS协议（采用灵活文件布局类型Flex Files）提供与专业并行文件系统相当的高性能。其中元数据与数据路径的分离是该架构的关键特性
- 具备企业级横向扩展NAS设备常见的易用性和功能，例如快照、复制、审计等

构成组件

客户端（Clients）是存储服务的使用终端，通常为工作站、应用服务器或集群。

存储服务器（Storage Servers）承载存储设备。任何支持NFSv3协议的存储系统均可通过Flex Files技术成为存储服务器。

元数据服务器（Metadata Servers）作为系统核心，负责管理数据组织结构、访问控制以及所有文件系统元数据。

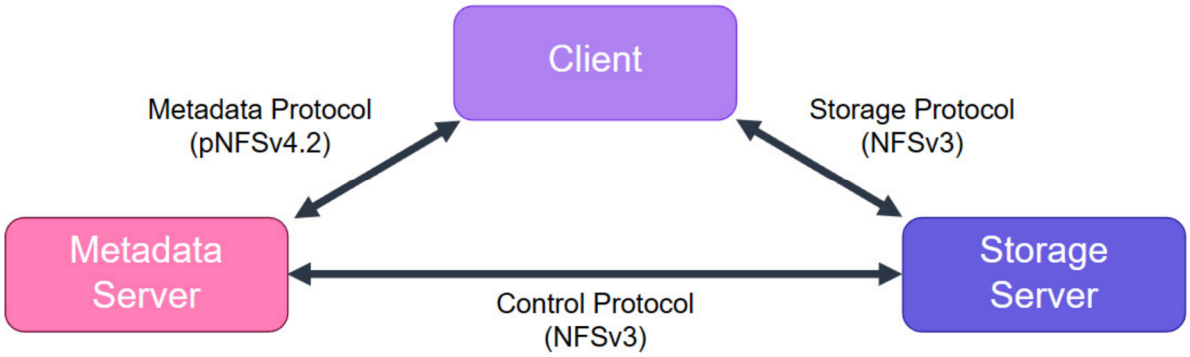
目录

- 超大规模NAS：兼具高性能与易用性
- 组件与通信类型
- 通过Dell构建模块化的部署架构示例
- 约1PB和5PB容量系统的示例机架布局图
- 组件规模考量
- 容量扩展选项
- 性能说明





HAMMERSPACE



上图 - 超大规模NAS的组件与协议

协议

上述三类组件通过pNFS Flex Files定义的协议进行通信。

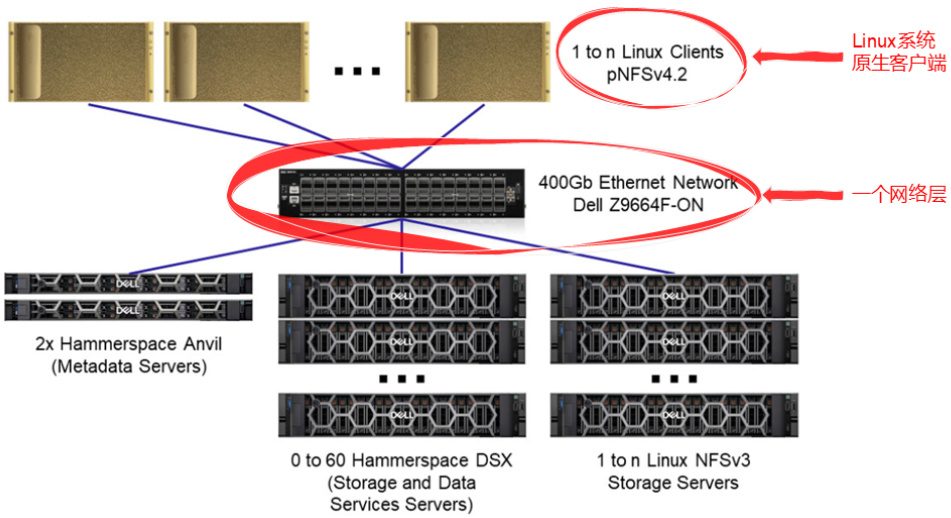
客户端通过NFSv4.2协议与元数据服务器通信，挂载共享并执行其他元数据操作。在从元数据服务器获取布局信息后，客户端使用NFSv3协议直接向存储服务器读写数据。而元数据服务器也通过NFSv3协议与存储服务器通信以管理数据访问。

Hammerspace与超大规模NAS

Hammerspace软件可创建跨越多个站点和多个云环境的全局数据环境。在超大规模NAS部署中，由冗余配对的Hammerspace Anvil服务器充当元数据服务器。存储服务器可采用Hammerspace DSX数据服务节点或普通Linux NFSv3服务器。当需要数据编排能力，或除NFSv4.2外还需通过NFSv3、SMB、S3或Kubernetes CSI协议导出文件系统时，则使用DSX节点。有关Hammerspace架构与功能的详细信息，请参阅《Hammerspace技术白皮书》

部署架构

戴尔PowerEdge R660和R760的1U及 2U机架式服务器是本示例的基础构建单元，但来自其他厂商的设备同样适用。大多数主流服务器 厂商都提供类似产品。NVMe闪存可为超大规模NAS提供高性能（这是其核心的设计标准），但该架构并不专属于NVMe。对于DSX或Linux存储服务器中的大容量 存储，HDD或SSD同样适用。

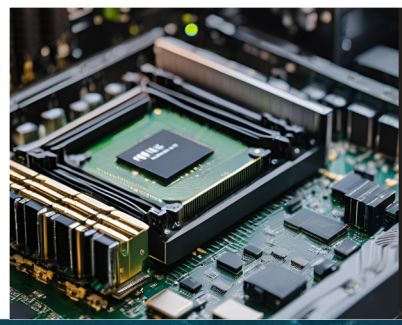


上图 – 超大规模NAS部署示例

服务器规格

Anvil 元数据服务器	服务器	Dell R660 1U 2.5” Chassis Up to 10 front NVMe drives, 2CPU, no PERC (RAID) Controller Riser Config 3, Full Height, 2x16 FH Slots (Gen5)
	启动盘	BOSS-N1 Controller card w/2 M.2 480GB Drives, RAID1
	CPU	2x Intel Xeon Gold 6548Y 2.5GHz, 32 Core / 64 Thread, 60M Cache, 250W, DDR5-5200
	内存	16x 64GB RDIMM, 5600MT/s (1TB Total)
	NVMe	4x 6.4TB Enterprise Mixed Use AG Drive U.2 Gen4 Metadata storage, mirrored (12.8TB Usable)
	网络	1x Nvidia ConnectX-7 Single Port NDR OSFP PCIe, Full Height
DSX 或 Linux 存储服务器	服务器	Dell R760 2U 2.5” Chassis Up to 24 front NVMe drives, 2 CPU, no PERC (RAID) Controller Riser Config 3, Half Length, 2x8 FH Slots (Gen4), 2x16 FH Slots (Gen5), 2x16 LP Slots (Gen4)
	启动盘	BOSS-N1 Controller card w/2 M.2 480GB Drives, RAID1
	CPU	2x Intel Xeon Gold 6542Y 2.9GHz, 24 Core / 48 Thread, 60M Cache, 250W, DDR5-5200
	内存	16x 64GB RDIMM, 5600MT/s (1TB Total)
	NVMe	24x Solidigm D7-P5520 15.36TB Enterprise Read-Intensive Drive U.2 Gen4 (368TB Total Usable)
	网络	2x Nvidia ConnectX-7 Single Port NDR OSFP PCIe, Full Height, installed in x16 Gen5 slots

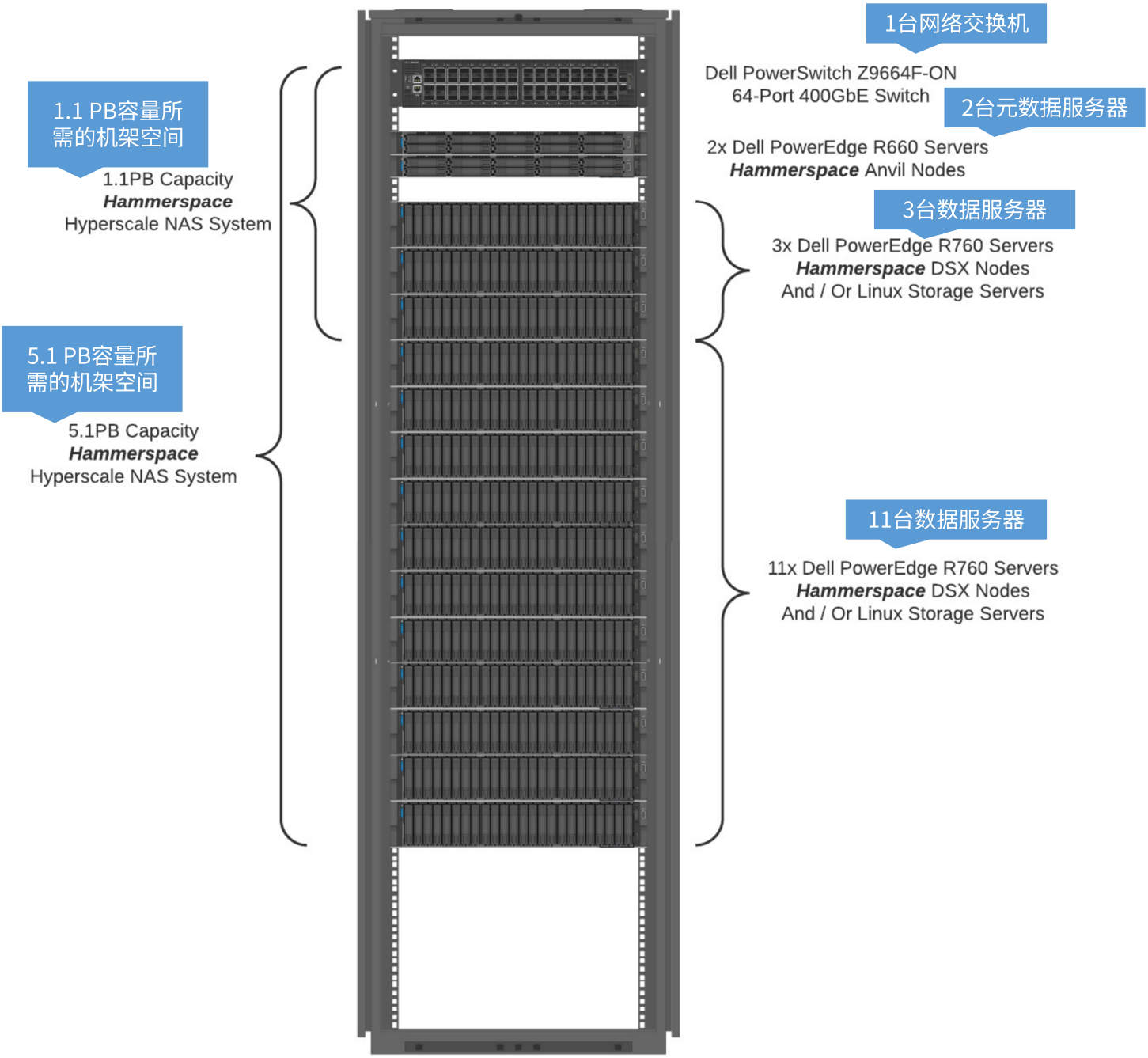
上图- 服务器规格



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

部署示例

下方的机架立面图展示了基于上述部署架构构建的Hammerspace横向扩展NAS系统容量示例：1.1PB与5.1PB



上图 – 部署示例

容量规划与配置注意事项

所有服务器组件必须由标准Linux发行版支持（例如RHEL）。

Anvil服务器

主板存储接口：尽管支持"三模"或"通用"存储控制器，但选择"直连"NVMe选项以最小化延迟。不同厂商的术语可能有所差异。

CPU：元数据操作涉及大量小规模随机I/O，建议选用多核、高主频（2.5GHz+）CPU以避免CPU性能瓶颈。

内存：为所选CPU配置最快内存的均衡组合，以最大化内存带宽并降低延迟。对于第5代Intel CPU，建议每颗CPU配置8或16根DIMM。

元数据存储NVMe：使用“混合用途”级别、3DWPD及以上耐久级别的NVMe驱动器。每管理十亿文件约需8TB元数据存储空间。

DSX及Linux存储服务器

主板存储接口：尽管支持“三模”或“通用”存储控制器，但选择“直连”NVMe选项可以实现最小化延迟。不同厂商的术语可能有所差异。

CPU：高核心数和高主频对最大化IOPS及提升小块I/O性能至关重要。若工作负载以大块流式I/O为主，可选用核心数较少、成本更低的CPU（例如Intel至强金牌6526Y 2.8GHz 16核37.5M缓存）。

内存：为所选CPU配置最快内存的均衡组合，以最大化内存带宽并降低延迟。对于第5代Intel CPU，建议每颗CPU配置8或16根DIMM。

NVMe：若工作负载以读为主，或容量与耐久性可妥协，可选用高容量低耐久的"读密集型"驱动器；若写负载较重或需要更高闪存耐久性，建议选用低容量"混合用途"驱动器。

本配置选用的是15.36TB驱动器，只因其在戴尔在线配置工具中，为该平台可提供的最高容量。只要服务器平台支持，亦可使用更小或更大容量驱动器。

数据环境

影响服务器规模的数据量和使用模式的相关因素包括：

All Installations	Multi Site Installations Only
• File Count - 文件数量 • Share Count - 共享计数 • Concurrent Users -并发用户数 • Concurrent Open Files -并发打开文件数 • Peak Aggregate Client Throughput - 聚合客户端吞吐量峰值 • Policy Objectives Activity –策略目标细节	• Site Count -站点数量 • Shared Object Count -共享对象数量 • Shared Object TB -共享对象存储容量(TB) • Shared Object Throughput -共享对象吞吐量数值 • Data Compression -数据压缩比例

网络

网络技术、接口速度和接口数量的选择将高度依赖于客户的具体需求、偏好及现有基础设施。在本示例中，我们选择了Dell PowerSwitch Z9664F-ON交换机以提供最大数量的400Gb端口。选择400GbE网络旨在实现最大存储带宽，但需要说明的是，400GbE并非Hammerspace的强制要求。

- 建议使用nconnect挂载选项，使NFS能在每个客户端与服务器之间建立最多16条并行连接。当客户端和服务器配备多网口时，此配置可提升整体性能。建议通过测试确定最优连接数，因为并非连接数越多越好。
- 超大规模NAS系统无需依赖RDMA即可实现高性能，但RDMA作为可选技术可进一步提升吞吐量并降低延迟。使用RDMA功能时，Hammerspace DSX服务器需配备Mellanox ConnectX-4或更高版本的RNIC网卡。
- 支持TCP/IP协议的任何网络技术均可使用。如需RDMA支持，可选择InfiniBand（通过IP over IB）或RoCEv2（基于融合以太网的RDMA第二版）。
- Hammerspace已通过GPUDirect认证，可兼容配备NVIDIA GPU的客户端。关于Hammerspace与GPUDirect的协同使用，请参阅《GPUDirect技术简报》。

数据编排

在Hammerspace全局数据环境中，不同数据集可能因项目工作流程需要，在不同时间点产生差异化的存储需求。数据编排技术可根据存储目标（包括可用性、持久性、性能和地理位置等要素），将数据智能分配至合适的存储位置。编排操作由Anvil节点发起、监控并生成报告，但实际数据迁移工作由DSX服务器执行。根据数据迁移的规模及频率，可能需要升级DSX存储服务器的资源配置，或部署专用于数据编排的额外DSX服务器。当前系统最多支持60台DSX服务器。

数据保护

文档中的容量数据未包含数据保护需求。例如，若需在本地保留两个文件副本，实际可用存储容量将显示为标称容量的一半。若需在异地或云端存储数据保护副本，则需要相应增加数据迁移量，这可能会影响服务器规格及网络（局域网与广域网）的配置与容量规划。

容量扩展

存储容量扩展有两种方式：纵向扩展（Scale Up）和横向扩展（Scale Out）。纵向扩展指在现有DSX或Linux存储服务器上增加更多驱动器；横向扩展则是通过添加更多存储服务器实现。这两种方式可单独或结合使用。

纵向扩展（Scale Up）

根据初始及未来容量需求、性能要求和预算，存储服务器可部分搭载NVMe驱动器部署。例如：1U服务器可部署5块或10块驱动器；2U服务器可部署6块、12块或24块驱动器。

以上仅为示例，实际可随时按需添加驱动器，直至机箱支持的最大数量。

横向扩展（Scale Out）

超融合NAS系统也可通过添加更多DSX或Linux存储服务器实现横向扩展。目前生产环境中已有最多1000台Linux存储服务器的配置案例。

性能说明

容量固然重要，性能亦不可忽视。存储系统的容量可通过简单计算得出，但性能预测涉及众多变量和限制条件。我们无法精确预知示例配置在客户具体环境中的性能表现。不过，基于硬件理论能力、实验室测试及实际部署经验，仍可预测接近真实场景的性能水平。

下表数据假设客户端无瓶颈限制。

吞吐量

Parameter	1.1PB 配置		5.1PB 配置	
	网速理论值	实际近似值	网速理论值	实际近似值
顺序读	300GB/s	264GB/s	1.4TB/s	1.2TB/s
顺序写	300GB/s	264GB/s	1.4TB/s	1.2TB/s

上表 - 示例配置下的吞吐量数字

本硬件配置的预期吞吐量基于以下信息计算得出：

- 所指定的Solidigm NVMe驱动器标称顺序读取速度为7.1GB/s，顺序写入速度为3.7GB/s
- 每台R760服务器搭载24块驱动器时，顺序读取总吞吐量为170.4GB/s，顺序写入总吞吐量为88.8GB/s

- 每个400GbE端口理论带宽为50GB/s
- 每台R760服务器配备2个400GbE端口，总带宽达100GB/s
- 基于TCP和RDMA高性能网络经验，应用程序可利用88%-94%的理论带宽，即每台存储服务器实际带宽为88GB/s（100GB/s × 0.88）

由于网络接口实际吞吐量88GB/s低于存储总吞吐量（顺序读170.4GB/s，顺序写88.8GB/s），每台R760存储服务器的性能将受限于网络接口的88GB/s。吞吐量呈线性扩展，因此示例配置的预期系统总性能为R760存储节点数量乘以88GB/s：

1.1 PB系统：含3台R760存储服务器，总吞吐量3 × 88GB/s = 264GB/s
5.1 PB系统：含14台R760存储服务器，总吞吐量14 × 88GB/s = 1.232TB/s

注：以上数据基于未启用RDMA的配置。启用RDMA可将吞吐量提升若干个百分点

IOPS

Parameter	读	写	70% 读/ 30%写 混合
单存储节点 IOPS 近似值 - TCP 4KB	390000	320000	365000
单存储节点 IOPS 近似值 - RDMA 4KB	470000	320000	400000

上表 – 每存储节点IOPS

上表数据基于上一代硬件实验室测试结果，对本示例配置采取保守估算：

平台: Supermicro SYS-2029BT-HNR

CPU: 2x Intel 6150

内存: 384GB (12x 32GB)

启动盘: M.2 固态硬盘

存储: 6块西数Ultrastar DC SN840 3.2TB NVMe固态硬盘，每驱动器独立卷

网络: 100GbE Mellanox ConnectX-5以太网卡，SN2100交换机，MLNX-OS操作系统，优先级流控制(PFC)、显式拥塞通知(ECN)、差分服务代码点(DSCP)配置

Parameter	1.1PB 配置			5.5PB 配置		
	读	写	70% 读/ 30%写 混合	读	写	70% 读/ 30%写 混合
预期IOPS (TCP协议/4KB)	1170k	960k	1095k	5460k	4480k	5110k
预期IOPS (RDMA协议/4KB)	1410k	960k	1200k	6580k	4480k	5600k

上表 – 示例配置的总IOPS数值

上表中的数据为上一页表格中，各存储节点的IOPS数值，乘以示例配置中的存储节点数量。客户端数量需匹配或超过存储节点数量方可达到此性能指标。

延迟

Parameter	读	写
最低延迟	45µs	105µs

上表 – 示例配置的延迟数值

本示例配置的延迟数据是基于使用上文IOPS部分所列硬件，在实验室进行测试得出的预估结果。