了解您的 Novus 系统

Novus

NetApp
November 03, 2026

This PDF was generated from https://docs.netapp.com/zh-cn/novus/get-started/learn-novus.html on November 03, 2026. Always check docs.netapp.com for the latest.

目录

了解您的 Novus 系统 . . . 1
- 了解 Novus 存储系统 . . . 1
  - 概述 . . . 1
  - 系统设计原则 . . . 1
  - 网络文件系统（NFS）实施 . . . 2
  - 其他技术和概念 . . . 2
  - 相关信息 . . . 3
- Novus 存储系统架构 . . . 3
  - ONTAP 数据服务器 . . . 3
  - Novus Data Director . . . 4
  - Novus 部署 . . . 4
- Novus 存储系统环境和处理 . . . 5
  - Novus Data Director 和 ONTAP 数据服务器内部通信 . . . 5
  - NDD 集群的高可用性和弹性 . . . 6

了解您的 Novus 系统

了解 Novus 存储系统

NetApp Novus 是一款功能强大的存储系统，专为满足超大规模人工智能的需求而构建。它将 NetApp 在企业存储方面的经验扩展到了训练和推理所需的超高性能基础架构中。了解其管理结构和配置选项对于实现有效管理和最佳性能至关重要。

概述

NetApp Novus 是一款横向扩展文件存储系统解决方案，专为现代 AI 工作负载所需的吞吐量和规模而设计。它将许多独立的 ONTAP 存储集群连接到一个单一的命名空间中，客户端将其视为一个文件系统。Novus 构建在并行 NFS （pNFS）上，它将元数据与文件数据分开。专用元数据服务管理命名空间，而存储集群直接向客户端并行提供文件数据。这消除了单个存储集群的吞吐量上限，不再限制系统传输数据的速度。

Novus 部署只需添加存储集群即可实现增长。每个新集群都将其全部带宽和容量添加到命名空间，因此随着系统的增长，总吞吐量会增加，而无需客户端重新挂载或进行中断性重新平衡。元数据服务与数据服务器分开扩展，因此您可以扩展命名空间大小和元数据性能以适应工作负载。Novus 旨在全速运行每个存储节点，并在大规模下将组件故障视为常规且无中断的事件。主要目标是即使系统扩展到非常大的规模，也能保持数据流向计算。

客户端使用标准 NFS 客户端，使 Novus 能够轻松适应大型 GPU 和计算集群。而且 Novus 是多租户的，因此一个部署可以托管多个隔离的客户或工作负载，每个客户或工作负载都有自己独立的命名空间。它还建立在 ONTAP 经过验证的数据管理和保护基础之上。对于客户而言，这意味着存储可以使 GPU 保持繁忙，通过从每个节点获得更高的性能来降低基础架构成本，并在单个命名空间中支持非常大的数据集。

Novus 可以帮助大规模运行 AI 和其他数据密集型工作负载的组织。这通常包括新云、GPU 即服务提供商、主权云、超大规模云服务提供商，以及运行 AI 训练和推理的大型企业和研究机构。Novus 可以让任何使用高性能计算或媒体处理的客户受益，这些客户需要每秒数 TB 的吞吐量，并希望与计算资源同步扩展存储。

系统设计原则

Novus 的设计和构建基于几项核心原则。这些基础架构选择可提高其性能、可扩展性和弹性。这可以帮助需要高性能存储的客户保持 GPU 忙碌并降低基础架构成本。

Exascale 就绪型存储

Novus 旨在成为一个支持百亿亿次级的软件定义的存储基础堆栈。它将水平扩展的 pNFS 元数据服务器与独立且解耦的数据服务器相结合。

超出传统 **ONTAP** 集群限制的吞吐量

传统上，单个 ONTAP 集群会达到由其控制器数量和互连带宽设定的固定吞吐量上限。AI 训练和推理需要单个 ONTAP 集群无法达到的吞吐量。

通过添加集群进行扩展

Novus 将许多独立的 ONTAP 集群（每个集群小到一个 HA 对）合并到一个命名空间中。每个新集群都贡献了其全部带宽，因此总吞吐量呈线性增长。客户端不需要重新挂载，设备群也不需要重新平衡。

可单独扩展的元数据

元数据服务层和分片元数据存储都可以从一个实例增加到 1,000 多个实例。元数据容量和请求率不取决于使用的数据节点数量。

线速数据节点

每个数据节点在读取时应以网卡全速运行，在写入时应以 NVMe 全速运行。任何未使用的每节点吞吐量仅表现为额外的硬件和成本。

弹性是主要的设计制约因素

在 1,000 个或更多存储节点的规模下，故障并不罕见。弹性是主要的设计约束，而不是事后的运维考量。

网络文件系统（NFS）实施

Novus 是网络文件系统（NFS）协议的软件定义实现。它依赖于 NFSv4（特别是并行 NFS（pNFS）扩展）来拆分元数据和数据路径。客户端可以使用标准 NFS 客户端使用 NFSv4 访问 Novus Data Director（NDD），同时使用 NFSv3 直接在 ONTAP 数据服务器上访问数据。

并行网络文件系统（pNFS）

并行 NFS (pNFS) 是 NFSv4.1 引入的一项功能，它将文件元数据管理与文件数据访问分开。NFS 服务器处理元数据操作，并为客户端提供有关文件数据存储位置的信息。然后，客户端可以直接在专用存储设备和数据服务器上访问数据，而不是通过单个服务器发送所有读取和写入。

NFS 服务器提供了一个布局，该布局描述了每个文件的数据如何排列以及客户端如何访问它。根据布局和存储配置，客户端可以与一个或多个数据服务器并行通信。这提高了数据传输性能，并减少了元数据服务器上的数据处理负载。

pNFS 支持多种布局类型，包括文件、块、对象和灵活文件。每种类型都描述了将文件数据映射到底层存储的不同方式，以及客户端访问这些数据的相应方式。Novus 使用灵活的文件布局。

pNFS 灵活的文件布局是布局选项的最新成员。它提供了一个灵活的基于文件的布局，可以描述存储在多个数据服务器上的数据，还可以支持镜像副本。客户端使用布局信息来选择适当的数据服务器路径进行访问。这使部署能够根据存储配置和工作负载要求调整数据放置和访问行为。

主要组件

有两个主要的 Novus 组件，它们建立在 pNFS 标准和 NetApp 的企业存储专业知识之上。

ONTAP 数据服务器

ONTAP 数据服务器实现 Novus 数据平面。它存储文件数据并将其提供给 NFS 客户端。Novus 命名空间可以将其文件数据分布在多个 ONTAP 集群中，一个 ONTAP 集群可以为多个命名空间提供服务。

Novus Data Director (NDD)

Novus Data Director 提供 Novus 元数据平面。它管理命名空间元数据，处理客户端对元数据操作的请求，并为客户端提供描述文件数据在 ONTAP 集群中存储位置的布局。客户端使用带有 pNFS 的 NFSv4.2 与 NDD 交互，而实际文件数据则直接从 ONTAP 访问。

查看"Novus 架构"以了解有关这些组件以及它们如何协同工作的更多详细信息。

其他技术和概念

还有其他几个与 Novus 和底层技术相关的概念。

命名空间

命名空间是一种组织和标识资源的逻辑结构，允许在定义的范围内一致地引用它们。使用 NFSv4，命名空间

是呈现给客户端的文件和目录的层次结构；它提供了导出文件系统的统一视图，客户端可以使用路径导航。每个 NDD 命名空间都有自己的元数据服务器实例。

联邦

联合是一种连接独立系统的技术，以便它们可以在保持单独管理的情况下协同工作；它可以应用于身份、应用程序服务、网络、数据库和其他资源。当与数据提供者一起使用时，联合使用户或应用程序能够通过统一界面访问来自多个源的数据，同时提供者保留对其数据的控制——从而支持互操作性并减少集中或复制信息的需要。

分片

分片是一种用于在多个服务器之间拆分数据库数据的技术，每个服务器都存储和服务一部分数据。它通过分布存储和查询负载来提高容量和性能。

边界网关协议

边界网关协议（BGP）是一种路由协议，用于在自治系统（构成互联网的网络）之间交换可达性信息，并使用路由策略和属性选择路径。Novus 依靠 BGP 来支持其多租户架构。

可扩展的分布式元数据存储区

可扩展的分布式元数据存储是跨多个节点管理元数据的系统，可确保高可用性、容错和高效访问。在 Novus 中，Novus Data Director (NDD) 实现了可扩展的分布式元数据存储（或简称 *metadata store*），以处理命名空间元数据并为客户端提供用于访问文件数据的布局。

相关信息

- "RFC 8435：并行 NFS（pNFS）灵活的文件布局"
- "RFC 8178：NFSv4 扩展和次要版本规则"
- "RFC 8881：网络文件系统（NFS）版本 4 次要版本 1 协议"
- "RFC 7862：网络文件系统 (NFS) 版本 4 次要版本 2 协议"

Novus 存储系统架构

Novus 架构由 Novus Data Director（NDD）和 ONTAP 数据服务器的实例组成。ONTAP 数据服务器支持数据路径，而 NDD 通过并行 NFS（pNFS）管理元数据和客户端访问。

ONTAP 数据服务器

数据服务器是专门为 Novus 构建的 ONTAP 集群。它们实现 Novus 数据平面，存储文件数据并将其提供给 NFS 客户端。一个 Novus 命名空间可以将其文件数据分布在多个 ONTAP 集群中；一个集群可以为许多不同的命名空间提供服务。这种灵活性允许根据需要添加数据服务器，从而实现容量和性能的增长。

客户端使用标准 NFS 客户端访问 ONTAP 数据服务器，无需自定义 NetApp 软件。客户端从 Novus Data Director 获取文件布局后，可以通过 NFSv3 和 RDMA（可选）直接在 ONTAP 数据服务器上读取和写入文件数据。批量数据从不通过 Novus Data Director，因此数据流量在数据服务器之间并行传输。

数据服务器不参与管理命名空间。它们将每个文件的数据作为后备文件存储在 ONTAP 卷中。ONTAP 数据服务器不维护 Novus 命名空间元数据，例如目录树或文件在数据服务器上的条带化方式。

适用于 NetApp Novus 的 AFF A90 是专为运行 ONTAP 的 Novus 打造的 AFF A90 版本。其特点包括：

- 额外的双 NVRAM 卡
- 高带宽网络适配器
- 配置为标准 ONTAP HA 对
- 使用 RAID-DP 或 RAID-TEC 保护数据。

标准 ONTAP 集群不能用作 Novus 数据服务器；仅支持专门为 Novus 构建的平台。

Novus Data Director

Novus Data Director (NDD) 是 pNFS 架构中元数据服务器 (MDS) 角色的实现，并提供 Novus 元数据平面。_MDS 主机_是一个运行 Linux 以及 MDS 协议服务和元数据存储的物理服务器。

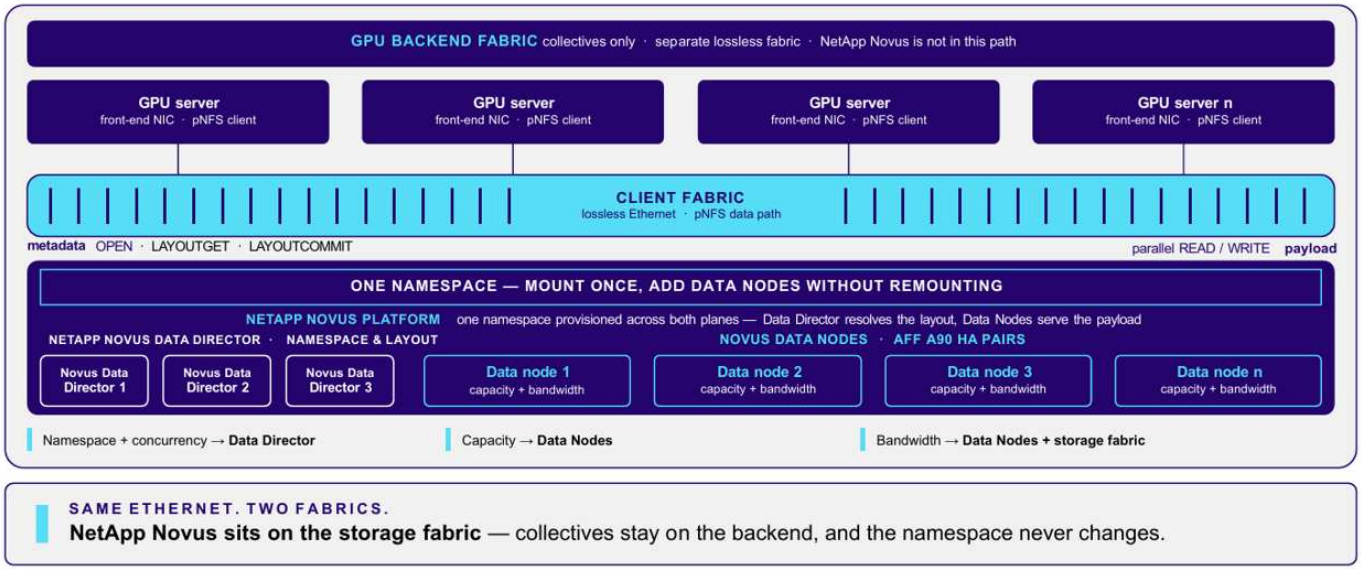
NDD 将每个命名空间作为单个文件系统呈现给客户端。它拥有命名空间元数据：文件和目录、文件属性、每个文件的数据存储位置的映射以及协调客户端访问所需的协议状态。客户端通过其虚拟 IP 地址挂载命名空间，并使用带有并行 NFS (pNFS) 的 NFSv4.2 与 NDD 进行通信。NDD 处理所有元数据操作，例如查找、文件创建和属性更改。当客户端打开文件时，NDD 会为其提供 pNFS 灵活文件布局，该布局会告诉客户端哪些数据服务器保存了文件数据。然后，客户端可以直接与 ONTAP 数据服务器进行读取和写入；NDD 位于元数据路径中，但不在数据路径中。

NDD 被实现为在至少三个专用服务器的集群上运行的软件。它将元数据服务器进程与复制的分布式数据库相结合，该数据库将命名空间元数据存储在加密的本地存储中。每个命名空间都在其中几个服务器上运行自己的元数据服务器进程。所有正常运行的服务器都会通知命名空间的虚拟 IP 地址，从而分散客户端负载，让客户端无需重新挂载即可进行故障切换。由于权威元数据存在于复制的数据库中，因此服务器可以发生故障并重新启动，而不会丢失命名空间状态。

ONTAP 集群上的数据放置由 NDD 控制。它保留数据服务器的注册表，使用轻量级 NFS 探测器确保每个数据服务器都可用，并跟踪每个服务器上的可用容量。创建文件时，NDD 会选择数据服务器，通过 NFS 创建后备文件，并记录其位置。如果数据服务器不可用，NDD 会停止在该服务器上放置新数据，并撤回指向该服务器的客户端布局。

Novus 部署

下图从宏观层面展示了典型的 Novus 部署。值得注意的方面包括具有多个 Novus Data Director 和 ONTAP 数据服务器节点的单一命名空间。此架构提供高可用性和低延迟，以支持后端 GPU 服务器。



Novus 存储系统环境和处理

您应该了解 Novus 环境及其处理的几个重要方面。

Novus Data Director 和 ONTAP 数据服务器内部通信

Novus Data Directors 与 ONTAP 数据服务器进行通信，以支持文件访问和管理。其中大部分是由 NDD 而不是 ONTAP 服务器发起的。

ONTAP 数据服务器发现

Novus Data Directors 需要发现可用的数据服务器。ONTAP 从不发起与 Novus Data Director 的通信。操作员或工具通过 `mds.conf`配置文件注册每个数据服务器，该文件在启动时使用；也可以使用命令 `mds-admin ds add`手动添加数据服务器。每个数据服务器在分布式元数据存储中表示为一行，包括其主机、导出、端口、状态和容量。

ONTAP 数据服务器可用性

有几个过程用于确定每个 ONTAP 数据服务器的可用性。

心跳

MDS 通过 TCP 向每个在线数据服务器发送 NFSv3 NULL RPC 调用。默认情况下，它每隔 5 秒使用心跳消息执行一次此操作。

正在下线

连续发生设定数量的故障后，数据服务器将被标记为离线。MDS 停止在其上放置新文件，并通知客户端归还指向它的布局。

返回

在连续进行足够多的成功探测后，它会返回到在线状态。冷却期可防止其来回切换。

客户端错误不计入

当客户端报告错误时，它只会被记录下来。ONTAP 数据服务器健康状况仅由探针决定。

ONTAP 故障转移不可见

如果 ONTAP HA 对执行接管，则 MDS 只能通过探测超时来发现。

ONTAP 数据服务器数据内容

NDD 不会询问 ONTAP 它存储了什么。它已经知道，因为它创建了每个后备文件：

- 创建文件时，MDS 会挑选数据服务器，在其上创建后备文件，并将结果保存在分布式元数据存储的条带映射中。条带映射记录哪个数据服务器和 NFS 文件句柄保存每个条带和镜像。
- ONTAP 上的后备文件在 NDD 中具有特定的名称格式。
- 容量是 NDD 在每个已挂载的导出上使用 statvfs() 从 ONTAP 读取的唯一操作值。
- 未根据 ONTAP 验证文件大小。它是由客户端的布局提交决定的。

NDD 集群的高可用性和弹性

由 Novus Data Directors 锚定的元数据平面确保了系统的高可用性和弹性。

集群设计

Novus 使用多主机集群来确保 NDD 服务的高可用性和容错。

三台具有三个故障域的主机

最小 NDD 群集是部署在单独服务器上的三个物理主机或虚拟机（VM）。每个主机运行一个 MDS 进程和一个元数据存储节点。该集群的设计可承受一台主机的丢失。

所有副本都提供

每个 MDS 副本同时处理客户端，没有任何副本被指定为待机模式。所有健康的副本都通过 BGP 宣布相同的 VIP。需要 BFD（快速链路故障检测），因为仅靠 BGP 自身的计时器并不是完全可接受的故障转移机制。只有当其 NFS 服务、网关和元数据存储连接都已准备好处理客户端时，副本才会宣布 VIP。

元数据复制

分布式元数据存储保留三个同步副本。仲裁器通过选择变为不可用而不是允许双方同时写入，来防止发生脑裂情况。

重新启动是安全的

MDS 进程不保留其自身的状态；分布式元数据存储保存权威元数据。

限制和注意事项

您还应注意以下限制和注意事项：

- 子树所有权不会自动故障转移。命名空间可以拆分为子树，每个子树由一个 MDS 拥有。如果所有者发生故障，即使 VIP 已经移动，其子树也会挂起，直到所有者恢复。
- 一个主机故障会影响该主机上的每个租户，因为每个租户每个主机都有一个副本。

- 客户端必须挂载 VIP，而不是主机 IP。否则，客户端无法进行故障转移。
- 硬挂载可能会在恢复过程中阻塞。阻塞的挂载并不意味着数据丢失。
- 始终一次对一个副本执行维护。首先将该副本从 VIP 中移出并将其排空。

版权信息

版权所有 © 2026 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。 美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 http://www.netapp.com/TM 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。