StarGFS 并行文件系统白皮书

上海星实信息技术有限公司

摘要

本文介绍上海星实信息技术有限公司自主研发的 StarGFS 并行文件系统，内容涵盖 StarGFS 的架构原理、系统部署、访问接口及功能特性，并对其进行了优势分析及典型应用举例。

如有疑问，请拨打咨询电话 021-31200121，

或邮寄您的问题至 support@xinfostor.com

其他更新内容，请访问：www.stargfs.com

目录

一、 并行文件系统需求背景和趋势 5
1.1 需求背景 5
1.2 并行文件系统发展趋势 5
二、 StarGFS 文件系统介绍 7
2.1 StarGFS 简介 7
2.2 StarGFS 组成 7
2.3 StarGFS 设计基础理念 8
2.3.1 文件系统的数据与元数据分离 8
2.3.2 文件系统元数据的存储和管理 8
2.3.3 文件系统真实数据的存储 8
2.3.4 文件系统的管理和监控 8
2.4 StarGFS 容量和性能方案配置 9
2.4.1 存储节点的配置和性能 9
2.4.2 元数据节点的配置 9
2.4.3 硬件平台的兼容能力 9
2.5 高可靠以及动态在线升级能力 10
2.6 文件的存储和访问 10
2.6.1 元数据指向对象 10
2.6.2 数据条带化存储 11
2.6.3 文件 I/O 过程 12
2.7 元数据节点服务 12
2.7.1 可扩展架构和配置说明 12
2.7.2 动态分布元数据管理 12
2.7.3 元数据存储 13
2.7.4 元数据服务高可靠设计 13
2.8 存储节点服务 14
2.9 客户端服务 14
2.10 存储池管理 15
2.11 块存储接口 15
2.12 对象存储接口 16
2.13 StarOND: StarGFS On Demand 17
2.14 其他高级功能 18
2.14.1 硬件和操作系统兼容性 18
2.14.2 访问控制列表（ACL） 18
2.14.3 配额–用户和组配额 18
2.14.4 快照克隆 19
2.14.5 备份工具 19
2.14.6 文件操作日志追溯 19
2.14.7 访问质量控制 QOS 19
2.14.8 数据透明加密存储 20
2.14.9 文件名称过滤 20
三、 StarGFS 核心优势 21

3.1 支持单节点、双节点、集群化部署 21
3.2 StarRAID 提高数据安全性和空间利用率 21
3.3 千亿级文件存储能力和聚合存储功能 22
3.4 文件系统和对象存储融合互通共享 23
3.5 超高的读写性能 23
四、 关于星实 25
五、 版权申明 26
六、 联系星实 26

一、 并行文件系统需求背景和趋势

1.1 需求背景

根据 IDC 的预测，未来十年数据的年增长速度将超过 40%。数据存储系统作为信息基础架构的重要组成部分，已经成为了业务发展的关键。目前，集群/分布式架构以其良好的扩展性和性价比优势，已经被广泛应用于各行各业。企业的相关业务已经从传统的昂贵且专用的“大型机”、“小型机”转换到了使用 Linux 和 Windows 操作系统的通用集群化架构，极大地提升了整体计算处理能力，并且能够良好地支持业务不断发展带来的扩展需求。

虽然集群/分布式架构给企业带来了灵活的计算处理能力，但同时也带来了新的问题。当计算集群中的业务高并发地对数据进行访问时，如果数据不能被快速地获取和共享，将会严重影响业务运转效率，并耗费大量的 CPU/GPU 计算资源。从集群化的趋势可以看出，企业也需要一个具备灵活扩展和高效率数据 IO 能力的大规模存储平台来匹配强大的集群计算处理能力。

目前主要的存储技术包括以服务器为中心的 DAS、以数据为中心的 NAS 和以网络为中心的 SAN。从现有的存储方案来看，NAS 和 SAN 主要是为含有小型文件和高级别事务的结构化数据设计的。然而，云计算时代的诸多应用场景具有传统存储方案在设计时并未考虑到的独特特征，包括大型文件、海量零碎小文件、高吞吐量、高 IOPS 要求、读取密集型访问方式以及大量并发、容量增长迅速的访问。特别是对于互联网、流媒体、高性能计算、能源、生物、金融等数据密集型、高并发访问的行业用户来说，容量与性能的线性扩展需求是传统存储方案无法跨越的鸿沟。

1.2 并行文件系统发展趋势

传统的解决方案，以 NFS 为例，提供了一个简单的模型，其中服务器导出一个文件系统层次结构，客户端可以映射到其本地名称空间。尽管被广泛使用，但客户机/服务器模型中固有的集中化已被证明是可扩展性能的一个重大障碍。在强烈的市场需求背景下，并行文件系统得到了快速发展，被广泛视为新一代的海量数据存储架构。集群技术应用于数据中心的服务器设备已经有几年的历史了，效果显著。作为集群技术的发展和延伸，并行文件系统具备了灵活的扩展性和极高的性价比。通过由若干个存储系统组成的集群，并行文件系统能够让企业在兼顾性价比的同时轻松实现性能与容量的在线无缝扩展，还有效解决了文件共享的效率问题。

系统设计师长期以来一直在寻求提高文件系统的性能，事实证明，文件系统对极其广泛的应用程序的总体性能至关重要。尤其是科学和高性能计算社区推动了分布式存储系统的性能和可扩展性的进步预计几年后会有更多的通用需求。

未来体系结构基于这样一个假设：PB 级系统是动态的、大型系统不可避免地以增量的方式构建、节点故障是常态而不是例外、工作负载和特性随着时间的推移不断变化。该体系结构的主要目标是可伸缩性（达到数百 PB 甚至更高）、性能和可靠性。可伸缩性可以从多个方面考虑，包括系统的总体存储容量和吞吐量，以及单个客户端、目录或文件的性能。目标工作负载可能包括这样的极端情况，即数万或数十万个主机同时读取或写入同一文件，或在同一目录中创建文件。这种场景在超级计算集群上运行的科学应用程序中很常见，越来越显

示出未来的常规工作负载。更重要的是，我们认识到分布式文件系统工作负载本质上是动态的，随着活动应用程序和数据类型的变化，数据和元数据访问会发生显著变化。并行文件系统直接解决了可伸缩性问题，同时通过三个基本设计特性实现了高性能、可靠性和可用性：解耦数据和元数据、动态分布式元数据管理和可靠的智能分布式存储。

二、 StarGFS 文件系统介绍

2.1 StarGFS 简介

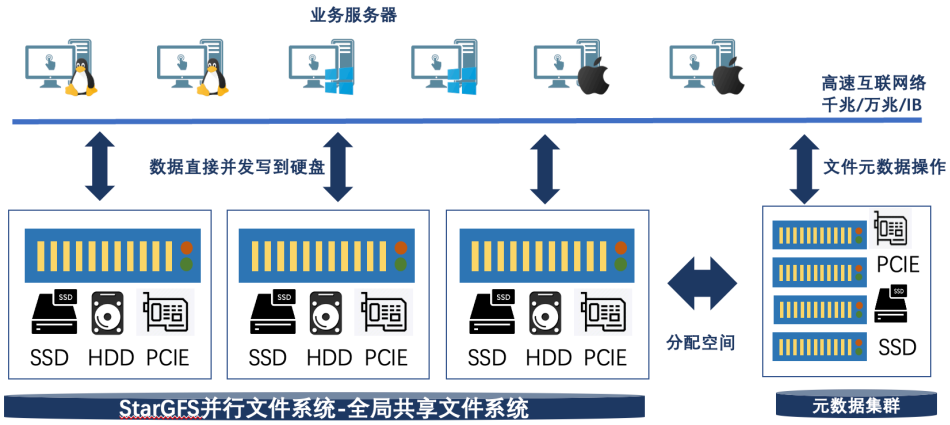
StarGFS，作为领先的基于 POSIX 文件系统接口的并行集群文件系统，允许业务无需重写或修改应用程序即可利用其功能。我们在开发 StarGFS 时，非常注重实现最大的性能和可扩展性，同时保持高度的灵活性和设计考虑了健壮性与易用性。

StarGFS 客户端通过高速网络连接与存储服务器进行通信，以访问文件系统内的数据。StarGFS 通过在多个存储服务器上透明地分布用户数据，使得在增加系统中服务器和硬盘的数量时，所有硬盘和所有服务器的容量和性能都聚合在一个文件系统名字空间中。这样，文件系统的性能和容量可以轻松地扩展到满足特定业务需求的水平，并在以后系统投入生产时实现在线扩展。

StarGFS 文件系统的可伸缩性有助于减少部署许多独立文件系统的需求，例如每个集群部署一个文件系统，或更糟糕的是每个 NFS 文件服务器部署一个文件系统。StarGFS 的统一名字空间带来了强大的存储管理优势，例如避免维护在多个文件系统上暂存的数据的多个副本。由于这个原因，与使用其他解决方案相比，使用 StarGFS 文件系统所需的聚合存储要少得多。通过将文件系统容量与许多服务器聚合在一起，还可以聚合 I/O 吞吐量，并随其他服务器扩展。此外，通过动态添加服务器，可以在安装集群后轻松地调整吞吐量或容量。

2.2 StarGFS 组成

StarGFS 并行文件系统包含三种系统模块：可用于访问文件系统的文件系统客户端、提供文件 I/O 服务的智能存储节点和管理文件系统中的名称和目录的元数据节点。下图显示了一个具有 StarGFS 文件系统的集群。



StarGFS 通过将用户文件的元数据与数据块分离，为用户提供更高效、智能的存储服务。具体来说，用户文件的数据块是由智能存储服务器负责存储和管理的，这些数据块包含了用户希望保存的数据内容。而元数据，即“关于数据的数据”，如文件的访问权限、大小以及如何在存储服务器上分布用户文件块的信息，则由 StarGFS 的元数据模块进行管理。当客户端获取了特定文件或目录的元数据时，它可以直接与智能存储服务器进行通信，以便存储或检

索文件的数据块。这样一来，元数据服务就不再参与读或写操作，从而提高了整体性能和可扩展性。

StarGFS 适用于所有需要处理大型和/或快速文件存储的应用场景。尽管 StarGFS 最初是为高性能计算而开发的，但如今它已广泛应用于工业和研究的各个领域，包括但不限于**人工智能、生命科学、石油天然气、金融、音视频媒体以及影像档案**。其无缝可伸缩性的特点使得用户能够轻松适应其随着时间不断变化的容量和性能需求。

2.3 StarGFS 设计基础理念

2.3.1 文件系统的数据与元数据分离

StarGFS 并行文件系统采用了基于对象存储理念的体系结构，其中传统硬盘被集成了 CPU、网络和数据的智能存储节点所取代。智能存储节点将传统的块级接口替换为一种新型接口，在这种接口中，客户端可以读写更大（通常大小可变）的命名对象，将低级块分配决策交由存储节点设备自身处理。客户端主要与元数据节点交互执行元数据操作（如打开、重命名），同时直接与智能存储节点通信执行文件 I/O 操作（读、写），显著提高了总体可扩展性。这种设计使得 StarGFS 能够更好地适应大规模数据存储的需求。

2.3.2 文件系统元数据的存储和管理

StarGFS 文件系统实现了元数据信息的任意写入，从而实现了无缝在线扩容。与传统的固定位图 bitmap 设计不同，StarGFS 采用了基于独立 SSD 存储来提高性能，并将文件系统的元数据信息任意写入到高速 SSD 存储中。这种设计不仅可以提高元数据的读写访问效率，而且可以根据实际文件系统中的文件数量动态地创建新的元数据信息。由于 StarGFS 的文件系统扩容不需要停止数据读写访问，因此可以确保更高的业务访问连续性。

2.3.3 文件系统真实数据的存储

此外，为了提高灵活性和管理便利性，StarGFS 采用了间接方式来访问硬盘。具体而言，StarGFS 将数据存储在本地 Linux POSIX 文件系统中，例如 ext4、xfs 或 zfs。这种设计使得用户可以根据不同的业务需求、数据特性、访问模式以及硬件特征等因素来选择最合适的本地文件系统。

2.3.4 文件系统的管理和监控

管理监控系统是 StarGFS 中元数据、存储和客户端节点配置及运行状态信息的集中展示和统计平台，为日常管理和维护提供了极大的便利。其设计简洁高效，通常无需专用机器，也不会存储用户数据。它负责监控所有已注册的服务并检查其状态。因此，在新环境中部署时，管理监控服务通常是首先启动的服务，且通常与管理元数据服务融合部署。

2.4 StarGFS 容量和性能方案配置

2.4.1 存储节点的配置和性能

StarGFS 使用服务器内置硬盘，并采用硬盘直通模式，无需配置传统的 RAID。用户可以选择将单个硬盘作为独立的分区，并格式化为本地文件系统（如 Ext4、XFS 等）。StarGFS 的智能存储节点和元数据节点会按照这些文件系统规定的格式来读取、写入和修改数据。每个智能存储节点可以管理 1 到 120 个数据存储硬盘，并根据服务器和目标进行 I/O 流量的负载平衡。StarGFS 文件系统的裸容量是所有数据硬盘提供的裸容量之和。此外，智能存储节点还应平衡业务网络和连接存储之间的网络带宽，以避免出现网络瓶颈。

例如，当有 100 台智能存储节点，每台配置了 36 个 10 TB 数据硬盘时，整体可提供将近 36PB 的文件系统容量。假设每台存储节点可以从每个数据硬盘获取到 100 MB/sec 的读取速度，那么每台存储节点可以提供高达 3600 MB/sec 的硬盘带宽。进一步而言，如果整个系统配置了万兆网络，每台智能存储节点配置了四个万兆网口绑定，或者采用了两个 25Gbps 及更高带宽的网络连接，那么每个智能存储节点都可以提供 3600 MB/sec 的端到端 I/O 吞吐量。从整个 StarGFS 文件系统层面来看，我们可以看到 100 台存储节点合计可以提供大约 360 GB/sec 的总吞吐量。在实际的生产系统规划中，我们还需要考虑更多的因素，如硬件性能、基准测试方案以及软件优化等参数，以确保系统的稳定性和性能。

智能存储节点通常使用服务器内置的硬盘，避免了 RAID 5 或 RAID 6 条带模式，从而降低了硬件或 CPU 的资源消耗。为了实现数据冗余保护，智能存储节点之间采用了 StarRAID N+M 或多副本的方式。智能存储节点的内存主要用于缓存只读文件，或者用于处理从写入中产生的脏数据。这样的设计提高了系统的可靠性和性能，同时降低了硬件成本和维护成本。

2.4.2 元数据节点的配置

元数据节点存储的数据访问模式与智能存储节点的存储方式有很大的区别。元数据节点主要处理少量数据的多次寻道和读写操作。而智能存储节点则通常涉及大量数据的网络传输和硬盘读写。由于元数据的存储对高吞吐量的要求不高，反而更注重随机读写性能和低延迟。因此，建议使用高速的 SSD 闪存介质来存储元数据。此外，元数据处理通常需要消耗 CPU 资源，StarGFS 建议至少使用十个处理器内核来确保系统的处理能力和稳定性。这样的优化配置能够更好地满足实际应用的需求。

2.4.3 硬件平台的兼容能力

StarGFS 的另一个重要特性是对多种硬件平台的广泛支持，包括 Intel 和 AMD 的 x86 平台，以及国产化的鲲鹏、飞腾、申威、龙芯、兆芯和海光平台。由于 StarGFS 的网络协议是独立于硬件平台的，不同平台的主机可以混合部署在同一个文件系统实例中。这确保了系统管理员能够在系统的整个生命周期中方便地添加新平台的系统，从而提高了系统的灵活性和可扩展性。这种跨平台的支持也使得 StarGFS 在处理大规模数据和提供高可用性方面具有显著优势。

StarGFS 典型配置如下：

名称	配置	数量
HDD 智能存储节点	CPU：2*INTEL 4314 内存：128GB 硬盘：1~120 块 HDD 机械盘（根据实际需求配置） 2 x 480GB SATA SSD(系统盘) 2 x 960GB SATA SSD（元数据盘） 网络：4 x 10Gb Ethernet Port/INFINIBAND	1~10000 台
全闪 智能存储节点	CPU：2*AMD 7513 / 2*intel 8368 内存：256GB 硬盘：1~120 块 SSD（根据实际需求配置） 2x 480GB SATA SSD(系统盘) 2 x 960GB SSD（元数据盘） 网络：4 x 10Gb Ethernet Port/INFINIBAND	1~10000 台

2.5 高可靠以及动态在线升级能力

StarGFS 服务器通常配备大量的硬盘，为数十到数万个客户机提供服务。并行集群文件系统应该通过高可用性机制（如故障转移）来完全透明地处理存储服务器节点重新启动或故障。当存储服务器节点发生故障时，应用程序应该只感受到访问文件系统的系统调用执行时产生的延迟。缺乏健壮的故障转移机制可能会导致作业挂起或失败，需要重新启动和集群重新启动，这将产生严重的后果。而 StarGFS 的故障转移机制提供了完全透明的文件访问保障。

StarGFS 的元数据节点配置为 Active-Active 的配对方式，元数据在元数据节点之间实现了相互备份。假设某一台元数据节点发生宕机，其服务将自动切换至与其配对的元数据节点，以确保元数据服务访问的连续性。而智能存储节点之间通常配置了多副本或 StarRAID（N+M）机制实现了跨机柜、存储节点、硬盘之间的可按需配置的冗余方式。这实现了单机最多容错四块硬盘、双机容错一个节点、集群容错四个存储节点/四个机柜等不同安全等级，并且存储空间利用率最高可达 90%以上。

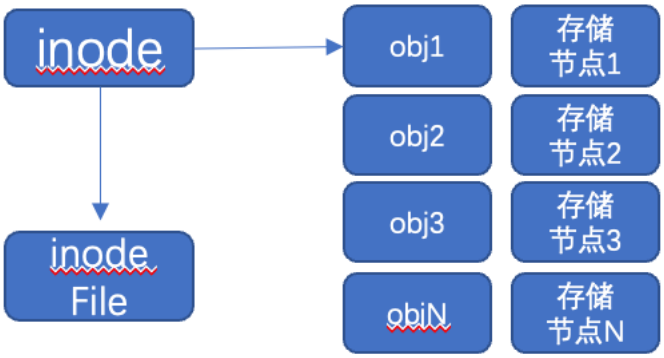
为了支持并行集群文件系统软件的滚动升级，需要一种健壮的故障切换机制以及提供版本间互操作性的软件。StarGFS 允许逐个升级存储服务器节点而无需关闭整个文件系统。存储服务器只需单独脱机、升级并重新启动。所有数据访问任务将无缝继续运行，不会出现故障，应用软件在访问数据存储时仅会经历几秒的延迟。

2.6 文件的存储和访问

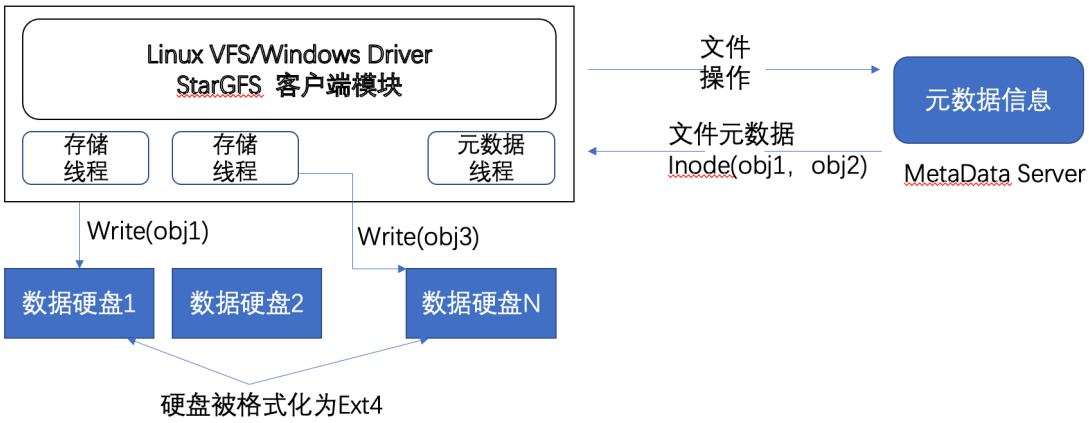
2.6.1 元数据指向对象

传统的 UNIX 硬盘文件系统使用 inode，inode 包含存储 inode 文件数据的块号列表。类似地，对于 StarGFS 文件系统中的每个文件，元数据节点上都存在一个 inode。但是，在 StarGFS 文件系统中，StarGFS 的 inode 并不指向数据块，而是指向与文件关联的一个或多个对象。这些对象实际上是存储在分布式存储系统中的数据块，而不是传统的硬盘数据块。

因此，StarGFS 的 inode 提供了一种将文件映射到分布式存储系统中相应数据块的机制。



这些对象在文件系统上以文件的形式实现，并包含文件的数据。下图展示了当客户端打开文件时，文件打开操作如何将对象指针从元数据节点传输到客户端，以及客户端如何利用这些信息直接与存储对象的智能存储节点进行交互，执行 I/O 操作。



2.6.2 数据条带化存储

在 StarGFS 文件系统中，如果一个文件只与一个对象相关联，那么该对象将包含文件中的所有数据。然而，当一个文件与多个对象关联时，文件中的数据将被“分条”，跨不同的对象存储。这种分条机制类似于使用 RAID 0 模式，其中数据被划分成多个条带，并分布在多个对象中。每个对象都包含一部分数据，称为“块”。当向某个对象的某个块写入数据时，如果数据量超过了条带的限制，则文件的下一个块数据将被存储到下一个目标上。

StarGFS 文件系统的容量等于所有数据硬盘的容量之和。聚合带宽则由智能存储节点所能提供的聚合带宽决定。系统的容量和聚合 I/O 带宽都可以通过增加智能存储节点的数量来扩展。

这种条带化文件的存储方式具有几个优点。**首先，它打破了单个目标大小的限制，使得文件系统可以支持单个文件达到 PB 级别，甚至单个文件系统达到 EB 级别。其次，通过将文件分散到多个对象中，可以有效地利用系统的 I/O 带宽。单个文件的 I/O 带宽可以随着与之相关联的对象的聚合 I/O 带宽而增加，这有可能相当于数百台服务器的带宽。**因此，StarGFS 的条带化存储机制可以显著提高文件的存储和访问性能。

2.6.3 文件 I/O 过程

当进程想要打开一个文件时，客户端会向元数据节点集群发送一个请求。元数据节点集群会遍历文件系统的层次结构，将文件名转换为对应的文件 inode，其中包括唯一的 inode 编号、文件所有者、模式、大小以及其他文件的元数据。如果文件存在并且允许访问，元数据节点将返回 inode 编号、文件大小以及关于如何将文件数据映射到对象的条带化策略的信息。此外，元数据节点还可以向客户端发出一个功能，明确允许哪些操作。在文件 I/O 过程中，元数据节点的后续参与仅限于管理功能，以确保文件的一致性并实现适当的语义。

2.7 元数据节点服务

元数据服务存储有关数据的信息，例如目录信息、文件和目录所有权以及用户文件内容在存储目标上的位置。当客户端打开文件时，它向客户端提供有关单个用户文件的位置信息。在文件关闭之前，元数据服务不会参与到实际的数据访问过程中，即文件的读写操作。

2.7.1 可扩展架构和配置说明

StarGFS 的元数据服务是一种可扩展的服务，这意味着在 StarGFS 文件系统中可以有一个或多个元数据服务。每个元数据服务负责其特定的全局命名空间的一部分，因此拥有更多的元数据服务器可以提高整体系统性能，而且未来添加更多的元数据服务器始终是可能的。

每个元数据服务实例只对应一个元数据目标来存储其数据。在元数据节点上，StarGFS 会创建多个元数据文件。这一设计决策是为了避免将所有元数据存储在一个数据库中，从而可能导致的损坏问题。

通常，元数据目标是基于闪存硬盘单盘、RAID1 或 RAID10 的存储配置，因为较低的元数据访问延迟可以提高文件系统的响应能力。StarGFS 的元数据量相对较小，并且与用户创建的文件数量呈线性增长。例如，512GB 的可用元数据容量可以容纳大约超过 1 亿个用户文件。

此外，更快的 CPU 内核将有助于降低延迟，从而提高系统整体性能。

2.7.2 动态分布元数据管理

由于文件系统元数据操作占据了典型文件系统工作负载的一半，因此有效的元数据管理对整个系统的性能至关重要。在分布式文件系统中，元数据管理也提出了一项关键的可伸缩性挑战：尽管容量和聚合 I/O 速率可以随着添加更多的存储设备而几乎任意扩展，但元数据操作涉及更大程度的相互依赖性，使得可扩展的一致性和一致性管理更加困难。

为了解决这个问题，StarGFS 采用了一种创新的基于动态子树划分的新型元数据集群体系结构。这种体系结构能够智能地将管理文件系统目录层次结构的责任动态地分配给多个元数据节点。这样，每个元数据节点的工作负载中都保持了局部性，有助于高效的更新和预取，从而在高负载下提高性能。StarGFS 能够有效地利用可用的元数据节点资源，并实现近似线

性的扩展，无论工作负载如何。

这种动态子树划分的方法使得 StarGFS 在元数据管理方面具有出色的性能和可伸缩性，从而在分布式文件系统中表现出色。

2.7.3 元数据存储

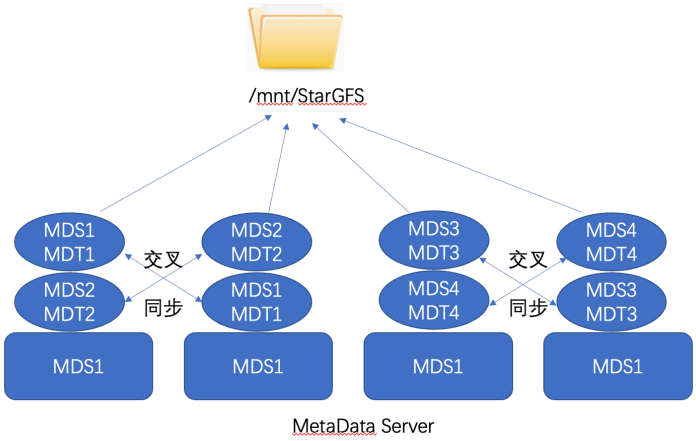
在 StarGFS 中，文件和目录的元数据信息相对较小，主要由 inode 组成，与传统的文件系统不同，它不需要分配元数据。这种设计简化了元数据的工作负载，使元数据节点能够有效地管理大量的文件，而不会受到文件大小的影响。为了进一步优化元数据相关的硬盘 I/O 操作，StarGFS 采用了两层存储策略，并通过动态子树分区来提高局部性和缓存效率。

这种策略带来了两个主要的优势。首先，它能够以高效（顺序）的方式将更新数据流式传输到硬盘，减少了重写工作负载，从而可以优化长期的硬盘存储布局，实现真正的读访问。其次，通过将 inode 直接嵌入到目录中，元数据节点可以一次性读取整个目录的元数据信息，利用了大多数工作负载中存在的高度目录局部性。这使得预取操作更加高效，进一步提高了系统性能。

2.7.4 元数据服务高可靠设计

在 StarGFS 文件系统中，整个文件系统的名字空间被平均分配给多个元数据服务器，每个元数据服务器负责存储和管理其分配到的部分名字空间。这些元数据服务器同时向客户端提供服务，使客户端可以访问整个文件系统的元数据信息。

在元数据集群中，所有的元数据服务器都是同时对外提供服务的，这使得 StarGFS 能够支持海量文件的高速查询。由于元数据服务器集群的规模可以轻松扩展，因此 StarGFS 能够轻松应对大量客户端并发进行查询操作。这种设计提高了系统的可伸缩性和性能，使得 StarGFS 能够高效地处理大规模的文件系统工作负载。



元数据服务保护模式图

为了确保元数据服务的可靠性和稳定性，StarGFS 采取了元数据服务器配对机制。每对配对的元数据服务器都会互相备份对方的数据和服务，并实时同步元数据的更新。这样，如

果其中一台元数据服务器出现故障或损毁，另一台元数据服务器将自动接管其服务，确保前端应用能够继续运行而不受影响。这种设计提高了系统的可靠性和可用性，使得 StarGFS 能够提供不间断的元数据服务。

2.8 存储节点服务

StarGFS 的存储服务是用于存储条带化用户文件内容(也称为数据块文件)的主要服务。它基于可扩展的设计理念，这意味着每个 StarGFS 文件系统实例可以有一个或多个存储服务，而每个存储服务都会为文件系统增加更多的容量和性能。

存储服务实例具有一个或多个存储目标，这些目标通常是配置了直通模式的以单块硬盘大小为分区的数据卷，并且可以连接内部或外部硬盘。与任何本地 Linux POSIX 文件系统一起工作，存储服务并不会在存储节点内部的数据硬盘之间配置任何 RAID。此外，StarGFS 会自动使用存储服务器上的所有可用 RAM 进行缓存，进一步提高了性能。

为了提高数据读写的性能，建议配置一个比服务器的硬盘流吞吐量快得多的网络。此外，为了分配空间并聚合多个服务器的性能，即使是对于单个大型文件，StarGFS 也使用条带化，这意味着文件被分割成固定大小的块，这些块分布在多个智能存储节点上。每个文件的块大小和目标数由负责的元数据服务在创建文件时决定，此信息称为条带模式。

在智能存储节点上，包含用户数据的文件称为块文件。为了不浪费空间，StarGFS 只在客户端实际将数据写入相应目标时创建块文件。同时，为了进一步节省空间，块大小不会静态分配。这意味着当用户只向文件中写入一个字节时，StarGFS 只会创建一个大小为 1 字节的块文件（如果大量小文件则会采用容器存储技术将小文件跟大文件一样的存储）。

默认情况下，StarGFS 会随机选择文件的存储目标，这使得在多用户环境中可以提供最佳结果。同时，这种策略也允许不同用户同时创建大小文件的随机混合。

为了防止存储目标耗尽可用空间，StarGFS 可以在服务端设置指定某些存储节点不再继续写入新数据，只提供老数据的读取操作。这种方法使 StarGFS 能够处理不同大小的存储目标。可靠的智能分布式对象存储是由数千个设备组成的大型系统，具有固有的动态性：它们以增量方式构建、增长和收缩，设备故障频繁且预计会发生，同时大量数据被创建、移动和删除。StarGFS 通过将数据迁移、复制、故障检测和故障恢复的责任委托给存储数据的智能存储节点集群，实现了更有效地利用每个智能存储节点上的智能（CPU 和内存），以实现可靠、高可用的数据存储，并具有线性伸缩性。

2.9 客户端服务

StarGFS 提供了 POSIX 兼容的客户端模式，该客户端通过 Linux 内核的虚拟文件系统接口进行本地注册，并支持 RDMA，以实现最佳性能。加载客户端时，它将后台通过服务启动脚本启动 StarGFS 客户端实现文件系统挂载的方法，与其他 Linux 服务一样。

本机 StarGFS 客户端应该在所有需要以最大性能访问 StarGFS 的主机上使用。然而，我们还可以通过 NFSv4 或 Samba 重新导出 StarGFS 挂载点，或者使用 StarGFS 作为 Hadoop 的 HDFS 的替代品。

StarGFS 还为 Windows 和 MacOS 提供了一个原生的 StarGFS 客户端驱动。安装并设置好 Windows 的 StarGFS 客户端驱动后，业务可以像使用已经完成格式化的本地硬盘一样访

问 StarGFS 的文件系统名字空间，并且可以跟 Linux 客户端共享同一个文件系统名字空间。同时，StarGFS 客户端还能够以块设备的方式来给客户端提供裸设备的访问接口，访问 StarGFS 服务端的块存储服务。

客户端运行在操作系统内核中，以驱动的方式运行，基本没有其他依赖包，因此能更好地兼容各个客户端环境。另外，客户端节点也可以通过 NFS、CIFS 或 Hadoop HDFS 访问 StarGFS。这样的设计使得 StarGFS 成为一个灵活、可扩展、高性能的文件系统，适用于各种不同的应用场景。

2.10 存储池管理

高性能存储系统的目标应该是充分利用闪存设备，为用户当前正在处理的项目提供最佳的访问速度。为了达到这一目标，一个有效的方法是结合使用闪存硬盘和机械硬盘，即将闪存硬盘用作透明缓存。

为了使用户能够获得他们当前正在工作的项目的全部闪存性能，StarGFS 存储池功能将闪存硬盘存储池对用户开放。这意味着，用户可以请求将当前项目移至闪存硬盘，并且所有对项目文件的访问都将直接从闪存硬盘提供。只有在用户决定将项目移回机械硬盘之前，才可以访问机械硬盘。

数据的放置对应用程序完全透明。这意味着，当数据从机械硬盘移动到闪存硬盘时，数据将保留在同一目录中。无论数据当前分配给哪个池，都可以直接访问文件而无需任何隐式移动。为了防止用户尝试将其所有数据放在闪存池中，不同的池存在不同的配额级别。系统管理员可以在这样的池中对任意存储目标进行分组，而不管它们使用哪种存储介质。此外，还可以有两个以上的池，并且设置不同的容量配额。

存储池可以由一个或多个存储目标组成，也可以包含所有的存储节点。尽管只是将存储池设置了容量配额，但实际该存储池的数据还是会写入到所有存储节点的所有数据盘中，确保得到最佳的并发读写性能。这样设计的目的在于实现最佳的存储性能、可扩展性和灵活性，以满足不断变化的工作负载需求。

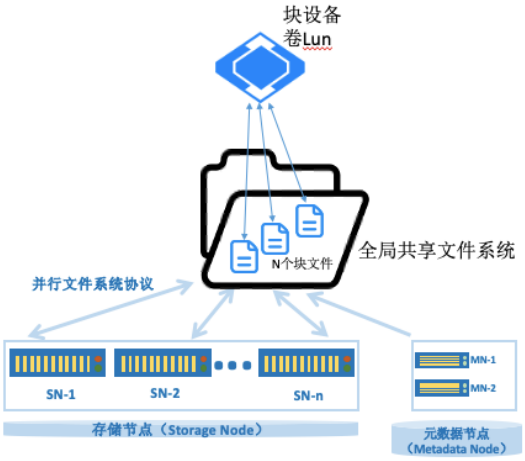
2.11 块存储接口

为了满足一些类似容器以及虚拟化等需要块存储的应用场景需求，StarGFS 并行文件系统在文件系统之上同时提供了 block 块存储的访问接口，支持 ISCSI、StarSAN 等协议。Linux 操作系统里面一切都是文件，块存储从操作系统层面看来其实也是一个文件。StarGFS 在全局共享的文件系统之上通过一个或者多个文件聚合的方式导出块存储的接口。这样的设计使得 StarGFS 不仅具有高性能的文件存储能力，还具备了提供高效块存储服务的能力。

StarGFS 块存储服务通过将文件系统中的数据块进行聚合和抽象，为上层应用提供了统一的块存储访问接口。这种设计方式不仅提高了数据的访问效率，还降低了存储管理的复杂度。由于 StarGFS 支持多种网络协议，因此可以方便地与其他存储系统进行集成，实现了数据的共享和备份等功能。

此外，StarGFS 还具备高可用性和可扩展性等特点。它采用了分布式架构，可以部署在多个节点上，并且支持在线扩容和容灾等功能。这些特点使得 StarGFS 成为了一种理想的分

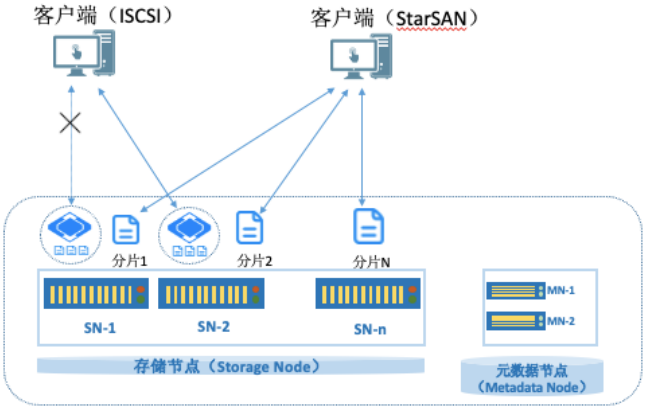
布式存储解决方案，适用于各种大规模的数据中心和云计算环境。



块设备由文件系统中一个或者多个文件组成

由于 StarGFS 是一个全局共享的文件系统，因此多个文件聚合而成的块设备可以在所有的智能存储节点上进行访问。具体而言：

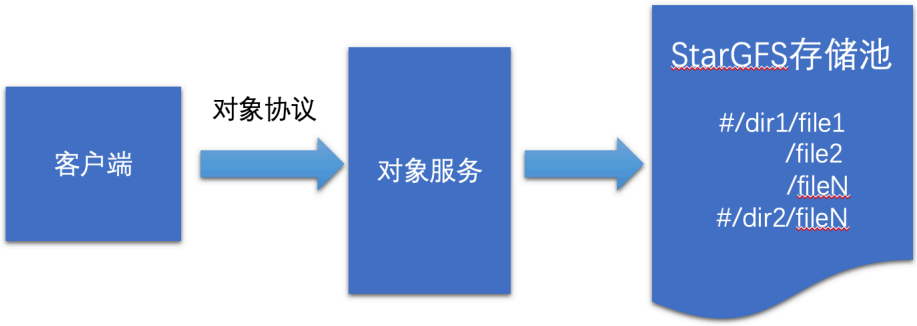
- 对于 ISCSI 客户端的连接访问，一台智能存储节点上的 ISCSI target 将接收到的 ISCSI 读写请求转换为 StarGFS 正常的文件读写请求，并将其分布到所有的智能存储节点中。如果这台智能存储节点发生故障，系统将通过多路径切换到另一台智能存储节点，确保数据的可用性和可靠性。
- 而对于 StarSAN 客户端，它呈现的是裸块设备，客户端对该映射的块设备进行数据读写操作时，会同时对应访问多台智能存储节点上的文件分片，最终实现高度并发的数据读写访问。这种设计方式提高了数据访问的效率和性能，并降低了存储管理的复杂度。



2.12 对象存储接口

StarGFS 基于全局共享的文件系统架构和自身强大的技术支持，能够高效地存储千亿级文件，并且通过优化存储技术，将小文件的性能提高十倍。在此基础上，StarGFS 将文件系

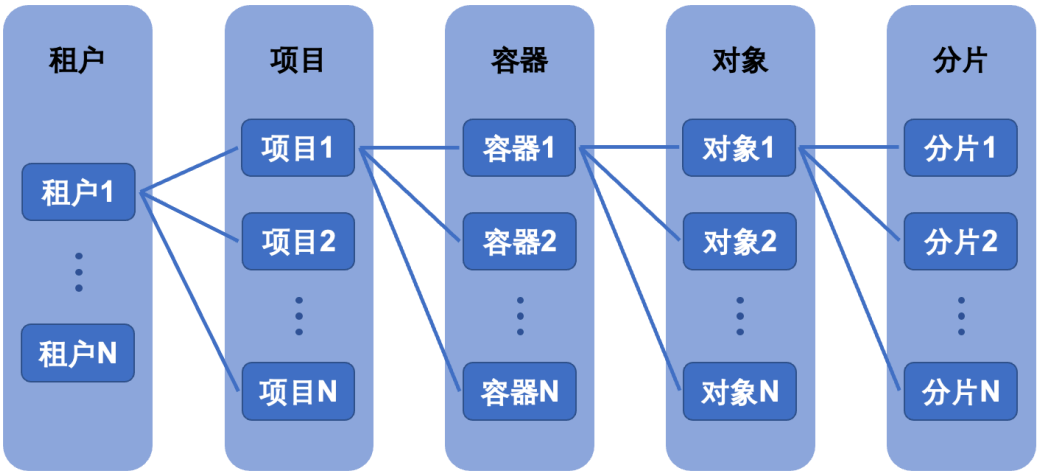
统中的文件以对象协议的方式对外输出 S3 访问接口，并且针对对象业务特点定制负载均衡算法，确保对象服务的整体架构无单点故障。



StarGFS 的对象服务支持多租户模式，有效提高了硬件资源的利用率。为了实现数据资源的绝对隔离和互不干扰，数据资源被单独存储。这种设计方式确保了各个租户之间的数据安全和隐私。

此外，StarGFS 的对象服务还具有无限扩展的特性。每层节点数没有限制，可以根据需要任意扩展。这种可扩展性使得 StarGFS 能够应对大规模的数据存储和访问需求，无论是对于单个租户还是多个租户，都能够提供高效、可靠、灵活的数据存储和管理能力。

综上所述，StarGFS 的对象服务通过多租户模式、数据隔离和无限扩展等特性，成为了一种理想的分布式存储解决方案，适用于各种大规模的数据中心和云计算环境。



2.13 StarOND: StarGFS On Demand

在集群计算环境中，计算节点通常配备有内部闪存硬盘，用于存储操作系统和为应用程序提供本地临时数据存储。然而，对于分布式应用程序而言，使用本地临时数据存储可能不太理想，因为它们需要跨多个计算节点共享数据，这可能导致无法充分利用 SSD 的高带宽和高 IOPS。

为了解决这个问题，StarOND（StarGFS On Demand 的缩写）应运而生。它支持为这些内部硬盘上的计算作业创建一个共享的并行文件系统。StarGFS 客户端挂载在为该作业分配

的节点上，为运行中的计算作业提供了一个快速、共享的全闪存文件系统。这种方式非常灵活，可以作为临时的数据缓冲区，也可用于存储临时数据。值得一提的是，由于 StarOND 只将数据存储在内部硬盘的子目录中，因此内部硬盘仍然可以由应用程序直接本地访问。

我们建议将 StarOND 与全局 StarGFS 文件系统结合使用，但它的使用并不依赖于全局共享集群文件系统是基于 StarGFS 还是其他技术。StarOND 能够独立地为一个计算作业创建一个新的独立挂载点。任何标准工具（如 cp 或 rync）都可以用于将数据传入和传出 StarOND。此外，StarOND 还提供了一个并行复制功能，用于在 StarOND 实例和其他文件系统（例如持久的全局 StarGFS）之间传输数据。

通过这种方式，StarOND 不仅能解决分布式应用程序在临时数据存储方面的挑战，还能提供灵活、高效的存储解决方案，从而进一步提高集群计算环境的整体性能。

2.14 其他高级功能

StarGFS 文件系统的一些附加功能如下所述。

2.14.1 硬件和操作系统兼容性

StarGFS 可以在多种 CPU 体系结构（如 x86、鲲鹏、飞腾、龙芯、申威、兆芯、海光等）上运行，使得客户机和服务器可以在这些不同的平台上进行融合部署。此外，StarGFS 还具有良好的软件版本互操作性，为用户在不同版本之间的切换提供了便利。

为了满足不同用户的需求，StarGFS 提供了高性能的 POSIX 内核客户端，支持 Windows、Linux、Mac 以及麒麟、欧拉、UOS 等多个常用操作系统版本。这些客户端具备快速上线并投入到生产系统的能力，从而为用户提供了高效、可靠的存储服务。

2.14.2 访问控制列表（ACL）

目前，StarGFS 的安全模型是基于 UNIX 文件系统的安全模型，并在此基础上进行了增强，通过 POSIX ACL 实现了更高级别的安全性。此外，StarGFS 还提供了一些额外的特性，例如存储系统自身拥有的权限控制，这使得即使客户端使用 root 权限，也无法对文件进行未授权的操作，从而进一步提高了数据的安全性。

StarGFS 的安全模型不仅符合 UNIX 文件系统的安全标准，还针对分布式存储系统的特点进行了一些扩展和优化。这使得 StarGFS 成为了一个既安全又可靠的分布式存储解决方案，适用于各种需要高度安全性和数据保护的应用场景。

2.14.3 配额–用户和组配额

StarGFS 允许为目录设置一个属性，该属性用于确定该目录及其所有子目录和文件的容量配额。这种配额可以针对目录、用户和用户组进行设置，包括容量、目录数和文件数的限制。这些配额适用于所有的访问接口，并且可以在线进行设定、更改和取消。配额的实时性达到秒级，同时支持配额的嵌套，并具备预警功能。

2.14.4 快照克隆

StarGFS 提供了快照技术，支持文件和卷的快速创建快照和克隆。它采用了 ROW（Redirect On Write）的方式，将新的写入数据重定向到新的位置，同时保留原始数据的快照。这种方式可以显著减少快照的资源消耗，并提高写入性能。

快照功能可以为文件或卷创建无限数量的基于增量的历史时间点拷贝。这些拷贝是只读的，可以随时查看和恢复。当发生数据“软”故障，如软件程序导致的数据损坏、病毒破坏、意外删除等，可以通过回滚到合适的时间点标记来快速恢复数据。这种回滚操作是增量进行的，不会对整个数据集进行全量备份，因此可以大大节省时间和资源。

2.14.5 备份工具

StarGFS 提供了支持复制迁移的工具 StarDM，该工具用以下步骤工作:

首先, StarDM 可以对文件进行快速扫描, 识别自某个特定时间点以来被修改过的文件。它能够高效地扫描文件系统，并生成包含修改文件的路径名列表。

然后，StarDM 启动并发多线程的文件同步推送，将文件从一个文件系统快速、安全地复制到另一个文件系统，无论是通过广域网还是局域网，都可以实现文件的推送和复制。同时，用户可以根据需求选择全量增量、加密、压缩和 QOS 控制等选项。

另外 StarDM 还支持不同存储系统之间的数据快速、高效、安全的复制和迁移。无论是点对点还是多点对多点的复制方式，都可以实现数据的快速恢复，确保用户的业务连续性。StarDM 还具备差异备份和增量备份等机制，并支持多个目标源之间的相互复制和推送。

2.14.6 文件操作日志追溯

StarGFS 的所有系统操作、维护和 IP 信息等都可以被详细记录在后台系统中。这些记录包括但不限于文件的创建、目录的创建、删除和修改等操作。同时，用户可以在线动态定制日志级别, 以及在线定制操作日志的记录内容。对定制内容的修改无需重新挂载或停机维护，从而确保系统的连续性和高可用性。

更重要的是, 这些操作日志可以作为后续文件出现不可控事件时的快速查找和定位事件原因的重要依据。通过这些日志，用户可以追踪到问题的源头，及时找到并解决问题，从而有效地保障了系统的稳定性和可靠性。

2.14.7 访问质量控制 QOS

StarGFS 并行文件系统提供了针对分布式块设备和文件系统的服务质量（QoS）设置功能。通过这些设置，我们可以对块存储的 IOPS、最大写操作和最大读操作进行限制，以确保存储系统资源得到合理分配，并保证核心业务不会受到其他业务的影响。这样，我们能够

为用户提供更优质的服务，并确保关键任务能够得到优先处理和保障。

此外，对于文件系统，我们提供了对读写带宽、读写 IOPS、创建文件、删除文件、创建目录、删除目录、查询、readdir 和 rename 等操作的限制功能。这些设置有助于对文件系统的使用进行精细化管理，确保各个应用程序和用户之间的公平性和资源利用的合理性。

通过这些 QoS 设置，我们的存储系统能够在分布式环境下更好地管理和优化资源分配，从而为用户提供更加稳定、高效和可靠的数据存储服务。

2.14.8 数据透明加密存储

StarGFS 提供了目录级文件加密的功能，从而为重要数据提供了强大的加密保护。

StarGFS 采用了独特的数据加密与数据块序列组合策略。在客户端，所有待存储的文件都会被自动切片并封装加密。然后，这些加密后的数据块会通过网络发送到各个存储服务器，以实现密文的保存。只有通过存储系统客户端程序，才能实现文件的组合以及明文的获取。任何其他应用程序，无论是通过网络进行非法截取，还是通过非法获取硬件设备，都无法对数据进行解密和数据块的组合。

这种设计方式确保了数据的机密性和完整性，使得 StarGFS 成为了一个安全、可靠的分布式文件存储解决方案，适用于各种需要高度安全数据存储的场景。

2.14.9 文件名称过滤

StarGFS 支持在客户端一侧对写入的文件名进行过滤或者限制特定类型的文件名，以增强安全性。具体措施包括：

1) 限制只能创建指定几种扩展名的文件。这意味着用户只能创建特定格式的文件，如“.txt”或“.doc”，而不能创建其他扩展名的文件。
2) 禁止指定扩展名的文件被访问或创建。这意味着某些特定扩展名的文件，如“.exe”或“.bat”，将无法被访问或创建，甚至使用“ls”命令也无法看到这些文件。
3) 禁止指定名称的文件被创建。有些病毒可能会在目录下创建自己的文件，为了防止这种情况发生，可以禁止创建特定名称的文件，如“virus.exe”或“trojan.bat”。
4) 设置文件名长度的上限。超过这个长度的文件名将无法被创建，以防止过长的文件名导致系统资源浪费或产生错误。

这些措施可以有效地增强 StarGFS 系统的安全性，防止恶意文件和病毒的入侵，同时也可以规范用户的文件命名行为，提高系统的可维护性和稳定性。

三、 StarGFS 核心优势

3.1 支持单节点、双节点、集群化部署

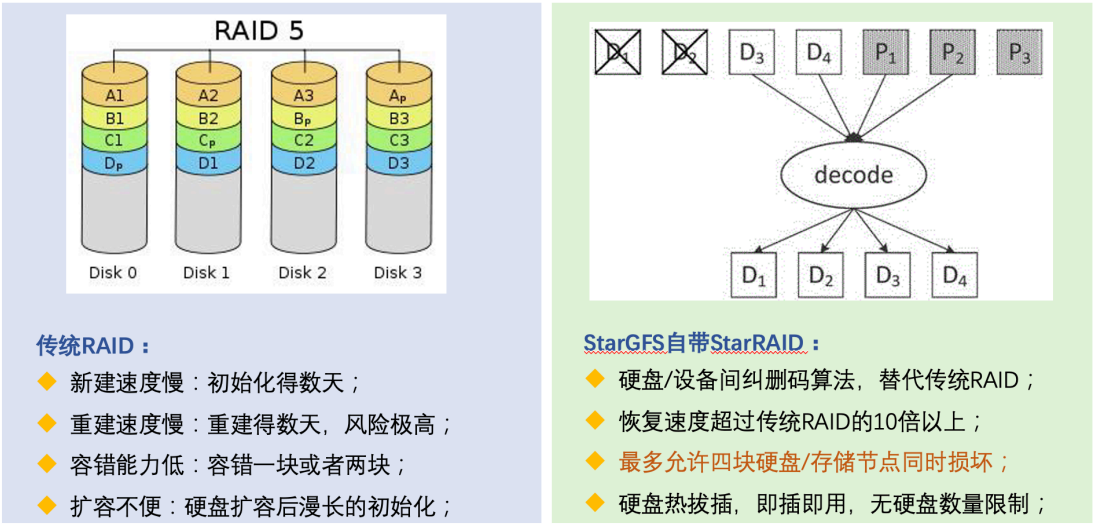
StarGFS 的初始配置仅需一台存储节点（智能存储节点）即可实际上线使用，同时还可以添加两台存储节点以增加系统的冗余性和稳定性。根据业务需求，用户可以按需安装指定数量的硬盘来满足存储需求。在部署了双台存储节点的情况下，StarGFS 支持动态增加硬盘或增加存储节点来扩展存储容量，无需停机。这种设计方式使得容量区间可以从数 TB 扩展到数百 PB，为用户提供了极大的灵活性和可扩展性。

随着硬盘数量的增多，StarGFS 系统的安全性和读写性能也会随之提升。StarGFS 系统具有更小化的初始部署能力，这使得并行文件系统可以轻松地从业务发展之初就融入到用户的业务系统中。此外，StarGFS 还支持文件系统持续在线扩展，以支持业务发展所带来的高性能和大容量需求。这种设计方式使得 StarGFS 成为了一个高性能、高可用性、易于扩展的分布式文件存储解决方案，能够满足各种大规模数据存储的需求。

3.2 StarRAID 提高数据安全性和空间利用率

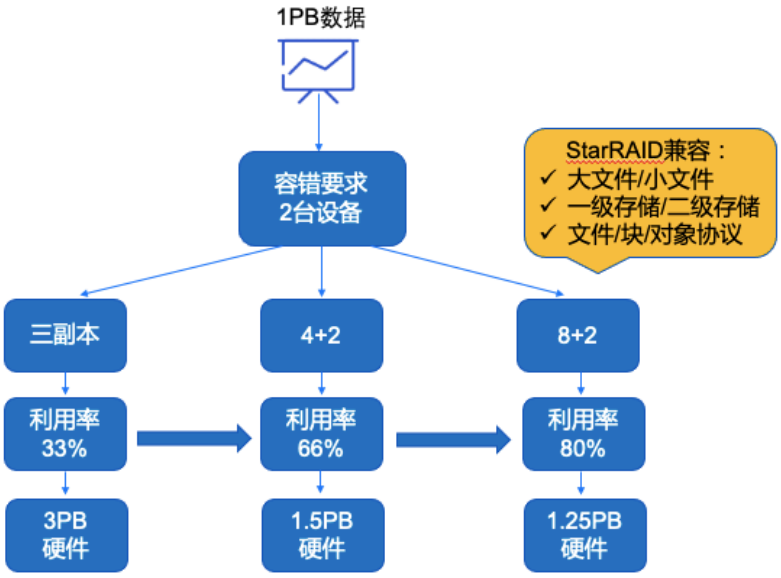
StarGFS 摒弃了传统 RAID 技术以及存储节点之间的数据镜像、HA 等复杂且成本高昂的技术，转而采用副本和 StarRAID 两种数据冗余模式等多种方式来确保数据的可靠性。同时，用户还可以针对单独的目录设置不同的冗余模式，以更好地满足实际业务需求。

在满足性能的前提下，用户可以根据业务需求，灵活地调整性能优先策略或容量优先策略。这种设计方式使得 StarGFS 在提供高性能的同时，也兼顾了数据的安全性和可靠性。此外，StarGFS 还提供了丰富的可配置选项，以便用户根据自身需求进行灵活调整，从而更好地适应各种不同场景下的数据存储需求。



StarGFS 的 StarRAID 冗余机制容错

相比市场上常见的分布式存储系统，它们在上线之初采用了 2 副本甚至 3 副本的方式来保障数据安全，从而付出了昂贵的硬件成本代价。然而，StarGFS 的 StarRAID N+M 冗余功能则提供了更为高效的数据保护方式，可以在文件、块、对象等三种协议基础上进行大规模部署上线，完全具备应用到生产线上的基础能力。相比传统常用的多副本策略，StarGFS 可以在更少的硬件投入情况下，存储相同的数据容量并实现相同的容错能力。这使得 StarGFS 在满足数据安全性的同时，降低了硬件成本，为用户提供了更具竞争力的存储解决方案。



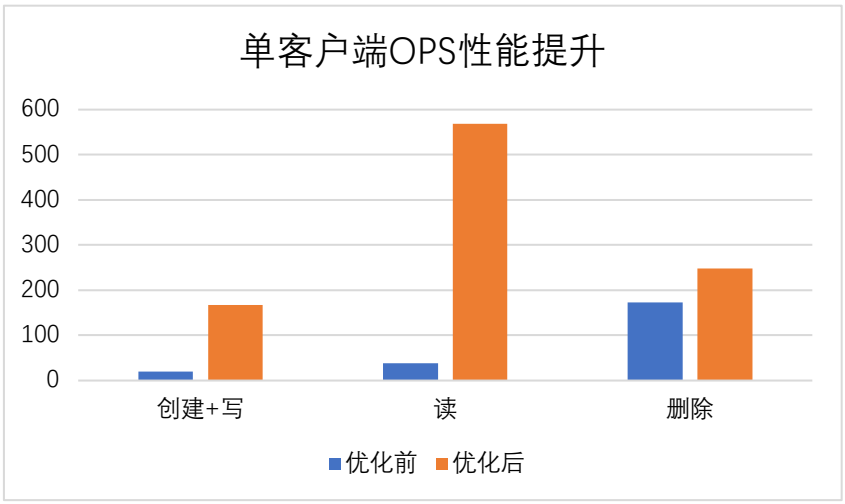
StarRAID 冗余容错机制带来巨大的构建、使用成本优势

3.3 千亿级文件存储能力和聚合存储功能

StarGFS 支持千亿级文件的统一存储和高效率检索。其内部的系统采用了创新的小文件 Container 容器存储技术，该技术能在前端呈现出标准 POSIX 文件系统的名字空间，同时在后端小文件落盘存储时实现智能的聚合成大文件的方式来实际存储。

通过采用该技术，StarGFS 将任何大文件和小文件都视为大文件一样进行存储和检索访问。这完全消除了传统文件系统在文件存储数量上的限制和文件检索效率低下等问题，打破了传统存储的局限。小文件的创建和检索效率提高到传统存储的十倍以上，使得 StarGFS 能够轻松应对任何存在海量小文件的业务场景。

StarGFS 的这种创新技术不仅提高了文件的存储和检索效率，还降低了存储的开销。由于小文件被聚合成大文件进行存储，使得存储空间的使用更加高效，减少了存储资源的浪费。此外，由于小文件的访问速度得到了显著提升，也间接提高了整个系统的性能和效率。

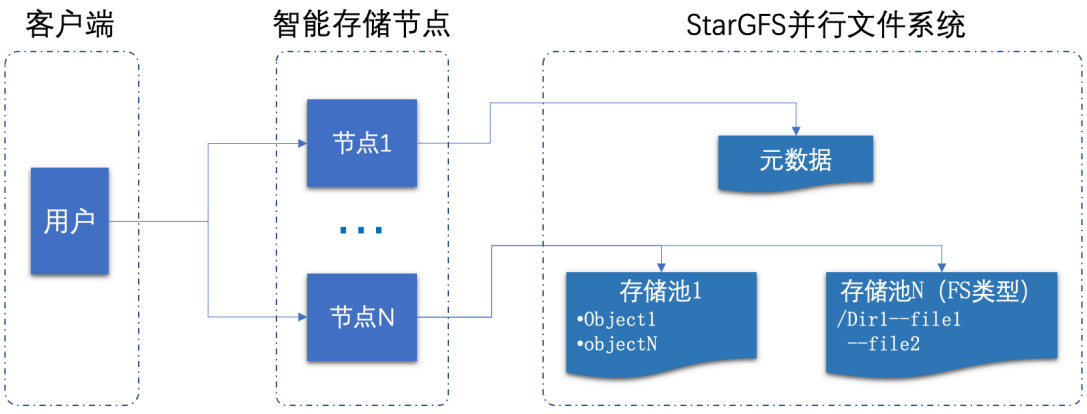


小文件性能优化效率变化

3.4 文件系统和对象存储融合互通共享

StarGFS 不仅支持提供 S3 对象存储接口能力，还具备文件系统和对象存储之间的数据互通功能。这意味着用户可以同时使用 StarGFS 的 POSIX、S3、NFS、CIFS 和 HDFS 存储协议，并按需访问同一份数据。这种设计方式使得企业可以轻松地构建数据湖存储平台，以满足各种不同的业务需求。

通过 StarGFS 的支持，用户可以灵活地按需配置和使用不同的存储协议，从而更好地适应各种不同的应用场景。无论是需要高性能计算、大数据分析还是需要高可用性的存储系统，StarGFS 都可以提供相应的解决方案，帮助企业实现高效、可靠的数据存储和管理。



3.5 超高的读写性能

StarGFS 是一个支持 RDMA 通信技术的分布式文件系统。它可以在客户端到存储服务器之间、存储服务器之间的通信等全过程支持 RDMA 方式进行通信。这种通信技术可以极大

地降低网络延迟，并充分发挥出超高速网络的极限性能。

据我们之前的性能测试数据来看，StarGFS 能够在硬盘性能足够的情况下，轻松地将网络极限性能发挥到接近 100%。例如，**单台存储节点如配置了 400G 网络则能够提供达到 40GB/s 的读写带宽；单客户端如配置了 400G 网络则能够获得并发读写 40GB/s 的读写带宽；单客户端单码流能够支持超过 10GB/s 的性能**等。这些性能指标充分满足了 AI 人工智能时代所需要超高速读写和小文件访问检索性能需求。

StarGFS 的 RDMA 通信技术可以提供高效的数据传输和低延迟的网络连接，使得大规模分布式存储系统可以更好地满足高性能计算、大数据分析和人工智能等应用的需求。

四、 关于星实

星实信息-海量文件数据存储专家

星实信息是一家专注于为客户提供高效、可靠的海量文件数据存储和管理解决方案的公司。我们的核心研发团队拥有多年的数据存储研发和产品化经验，曾在国际知名的存储公司从事相关工作。凭借着深厚的技术背景和专业知识，我们成功地开发出了高性能的 StarGFS 文件系统。

StarGFS 文件系统具有出色的弹性伸缩能力，支持百亿级海量小文件存储，并具有极高的并发读写性能。我们的秒级快照克隆技术使得数据备份和恢复更加高效。此外，StarGFS 还具备一系列企业级的功能特点，以满足不同客户的需求。**基于 StarGFS 文件系统，我们还推出了一系列专注于文件存储产品的解决方案，如并行文件系统和 StarNAS 存储系统。**这些产品覆盖了从办公文件存储、企业信息化的文件存储到教育科研计算的海量科研文件数据存储等应用场景，容量范围从 TB 级扩大到 EB 级。

我们始终坚持自主技术创新，以核心产品为依托，深入研究和探索行业应用。我们以技术创新为导向，致力于为客户提供优质的产品和服务，为客户创造价值，为社会贡献才智。我们将不断进取，务实笃行，为各行各业提供卓越的文件数据存储和管理解决方案。

目前，我们已经成功地为众多大型企业、科研院所、金融行业、电信运营商、广播电视台、影视动漫制作公司以及智慧城市项目等提供了优质的服务。我们相信，通过我们的不懈努力和技术创新，我们将能够继续为客户提供卓越的产品和服务，为客户创造更大的价值。

五、 版权申明

Copyright © 2023 上海星实信息技术有限公司 版权所有，保留一切权利。
非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部， 并不得以任何形式传播。本文档中的信息可能变动，恕不另行通知。

六、 联系星实

上海星实信息技术有限公司

TEL：021-31200121
Mail: Marketing@xinfostor.com
Web: www.stargfs.com
Add:上海市浦东新区金皖路 199 号 A 座 8F