CHANGHONG 横向扩展 CHCNMS 系统

技术概述

2023 年 5 月

四川长虹佳华信息产品有限责任公司

目　录

简介

随着时间的推移，传统三层存储模型（文件系统、卷管理器和数据保护）不断演进，以满足小规模存储体系结构的需求，但也带来了极大的复杂性，同时不太适合 PB 级别的系统。CHCNMS 操作系统取代了所有这些模型，提供具有内置可扩展数据保护功能的统一群集文件系统，同时无需进行卷管理。CHCNMS 是横向扩展基础架构的基本构造块，可实现极大规模和巨大效率，用于为所有用于为所有 NS 横向扩展 NAS 存储解决方案提供支持。

CHCNMS 的设计目的是不仅在机器方面，在技术复杂性上，CHCNMS 消除了复杂性，并整合了自我修复和自我管理功能，从而大幅减轻存储管理负担。CHCNMS 还在极深的操作 系统级别纳入并行度，使得几乎每项关键系统服务均跨多个硬件单元分布。随着基础架构的扩展，可使 CHCNMS 几乎扩展到每个维度，从而确保当前正常运转的设备将随着数据集的增长继续发挥作用。

CHCNMS 是完全对称的文件系统，不存在单点故障，利用群集不仅可扩展性能和容量，而且还可实现任意互联的故障切换和远超出 RAID 能力的多级冗余。磁盘子系统的发展趋势是缓慢提高性能，同时快速增加存储密度。CHCNMS 通过扩展冗余量以及故障修复速度，从而对这一现实作出了回应。这样有利于 CHCNMS 增至多 PB 规模，同时提供比小规模的传统存储系统更高的可靠性。

NS 硬件提供 CHCNMS 执行的一体机。硬件组件非同一般，但基于商用硬件，以确保从商用硬件不断改善的成本和效率曲线中受益。CHCNMS 允许随时任意安装硬件或从群集中删除硬件，以及从硬件中提取数据和应用程序。数据可无限期保存，免受硬件更新换代的影响。这就消除了数据迁移和硬件更新的成本和难题。

CHCNMS 非常适合企业环境中基于文件和非结构化的“大数据”应用程序，包括大规模主目录、文件共享、归档、虚拟化和业务分析。为此，CHCNMS 广泛用于当今的各类数据密集型行业，包括能源、金融服务、互联网和托管服务、商业智能、工程设计、制造、媒体和娱乐、生物信息、科学研究和其他高性能计算环境。

CHCNMS 概述

CHCNMS 将三层传统存储体系结构(文件系统、卷管理器和数据保护)整合到一个统一的软件层，创建了在 CHCNMS 支持的存储群集上运行的单个智能分布式文件系统。

图 1：CHCNMS 将文件系统、卷管理器和数据保护整合为单个智能、分布式系统。

这是一种核心创新，直接允许企业在当前的环境中成功利用横向扩展 NAS。它遵照关键的横向扩展原则；智能软件、商用硬件和分布式 体系结构。CHCNMS 不仅是操作系统，还是在群集中驱动和存储数据的底层文件系统。

NS 节点

CHCNMS 专门用于专用平台节点（称为“群集”）。一个群集由多个节点组成，这些节点是架装式企业一体机，其中包含：内存、CPU、 网络、以太网或低延迟 InfiniBand 互联、磁盘控制器和存储介质。因而，分布式群集中的每个节点都具有计算以及存储或容量功能。

借助第 6 代体系结构，创建一个群集需要一个 4U（机架单元）外形规格中 4 个节点的单个机箱，在 CHCNMS 8.2 及更高版本中最多可扩展至 252 个节点。各节点平台需要至少 3 个节点和 3U 机架空间来形成群集。有许多不同类型的节点，所有这些节点均可整合到单 个群集中，其中不同的节点提供不同的容量与吞吐量比或每秒输入/输出操作数 (IOPS)。传统第 6 代机箱和 NS 全闪存 NS F690、NS F660 和 NS F620 独立节点将在同一群集中共存。

每个添加到群集中的节点或机箱都会增加聚合磁盘、缓存、CPU 和网络容量。CHCNMS 充分利

用了每个硬件构造块，使整体能力远超部 件性能之和。RAM 组合为单一连贯缓存，使群集任何部分的 I/O 都能受益于在任何地方缓存的数据。文件系统日志可确保在电源故障期间提供安全的写入。磁盘轴和 CPU 被组合在一起，以便随着群集的增长而增加吞吐量、容量和 IOPS，适合于访问一个文件或多个文件。群集的存储容量可以从数十 TB 到数十 PB 不等。随着存储介质和节点机箱变得更加密集，最大容量将持续增加。

网络

有两种与群集相关的网络类型：后端网络和前端网络。

后端网络

群集中的所有节点间通信均通过专用后端网络（包括 10、40 或 100 Gb 以太网或者低延迟 QDR InfiniBand (IB)）来执行。这种后端网络（配置有冗余交换机以实现高可用性）用作群集的背板。这样，每个节点都可以充当群集中的参与者，并将节点对节点通信隔离到专 用的高速低延迟网络。该后端网络利用互联网协议 (IP) 进行节点对节点通信。

前端网络

客户端使用所有节点上可用的以太网连接（10 GbE、25 GbE、40 GbE 或 100 GbE）连接到群集。由于每个节点都提供自己的以太网端口，因此群集的可用网络带宽量随着性能和容量呈现线性扩展。群集支持与客户网络达成的标准网络通信协议，包括 NFS、SMB、 HTTP、FTP、HDFS 和 S3。此外，CHCNMS 还提供与 IPv4 和 IPv6 环境的完全集成。

完整群集视图

完整群集与硬件、软件和网络组合，如下图所示：

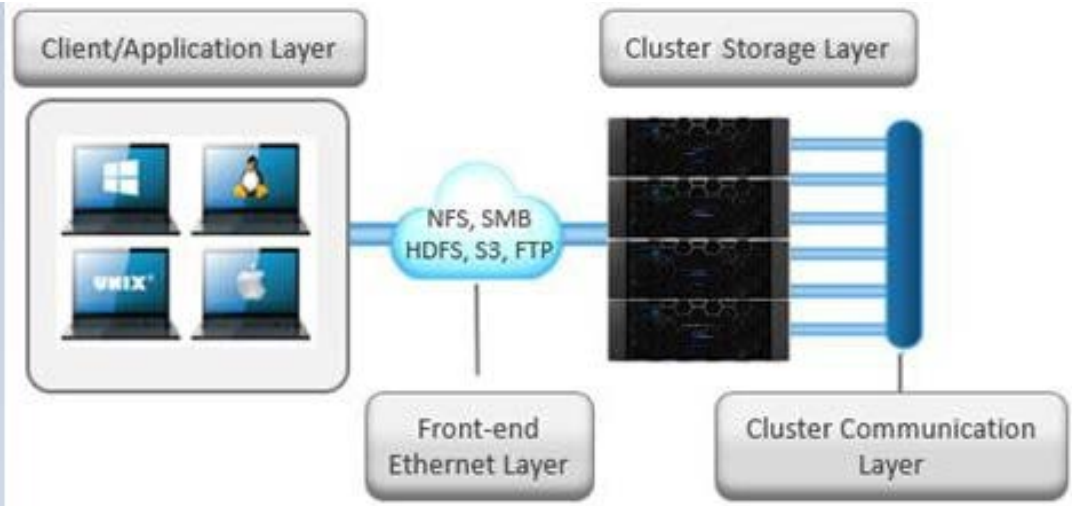

图 2：运行中的所有 CHCNMS 组件

上图描述了完整的体系结构；软件、硬件和网络在您的环境中共同协作，利用服务器提供完全分布式单一文件系统，可随着横向扩展环 境中的工作负载和容量需求或吞吐量需求变化而动态扩展。

CHCNMS SmartConnect 是一个负载均衡器，它在前端以太网层运行，以便在群集中均匀分发客户端连接。SmartConnect 支持 Linux 和 UNIX 客户端的动态 NFS 故障切换与回切以及 Windows 客户端的 SMB3 连续可用性。这可确保在发生节点故障或执行预防性维护时， 所有运行中的读写操作都转由群集中的另一节点处理，以便完成操作而不中断任何用户或应用程序。

故障切换期间，客户被均匀重新分布到群集中的所有剩余节点，确保对性能产生的影响极小。如果某个节点出于任何原因发生问题，包括故障，该节点上的虚拟 IP 地址将无缝迁移到群集中的另一个节点。当离线节点恢复为在线状态时，SmartConnect 将自动重新平衡整 个群集的 NFS 和 SMB3 客户端，以确保更大限度地提高存储和性能利用率。对于定期系统维护和软件更新，本功能允许每节点滚动升级，从而在维护窗口的整个持续时间内提供完全可用性。

CHCNMS 软件概述

操作系统

CHCNMS在基于BSD的UNIX操作系统(OS)基础上进行构建。它本机支持Linux/UNIX和Windows语义，包括硬链接、关闭时删除、原子重命名、ACL和扩展属性。这是一种成熟的经验证操作系统，它将 BSD 用作其基本 OS，同时可充分利用开源社区以实现创新。从 CHCNMS 8.2 开始，底层操作系统版本为FreeBSD11。长虹NS系列集群NAS存储与市场上基于同样操作系统的集群式NAS产品完全兼容，如：EMC、DELLEMC的ISILON系列，或POWERSCALE系列，可以作为其完全兼容的替换产品。

客户端服务

客户端用于与 CHCNMS 交互的前端协议称为客户端服务。为了解 CHCNMS 与客户端 的通信方式，我们将 I/O 子系统分为两部分:上半部分或“发起者”与下半部分或“参与者”。对于具体的 I/O 操作，群集中的每个节点都是参与者。客户端连接的节点是发起者，同时该节点在整个 I/O 操作中起“统领”作用。

群集操作

在群集体系结构中，有的群集作业负责监控群集本身的运行状况和维护，这些作业全部由 CHCNMS 作业引擎进行管理。作业引擎在整个群集中运行，负责划分和处理大型存储管理与保护任务。为此，作业引擎会将任务减少为更小的工作项，然后将整个作业的这些部分分配或映射到每个节点上的多个工作线程。在整个作业执行过程中，将跟踪和报告进度，并在作业完成或终止时显示详细的报告和状态。

作业引擎包括全面的检查点系统，该系统不仅可停止和启动作业，还可暂停和恢复作业。作业引擎框架还包括自适应影响管理系统。 作业引擎通常使用备用或特别保留的容量和资源在群集中将作业作为后台任务执行。这些作业本身可归为三个主要类别：

文件系统维护作业

这些作业执行后台文件系统维护，并且通常需要访问所有节点。这些作业需要在默认配置下运行，并且通常出现在已降级的群集情况 下。示例包括文件系统保护和驱动器重建。

功能支持作业

功能支持作业执行可方便某些扩展存储管理功能的工作，并且通常仅在已配置该功能时运行。示例包括重复数据消除和防病毒扫描。

用户操作作业

这些作业直接由存储管理员运行，以完成某些数据管理目标。示例包括并行树删除和权限维护。下表提供公开的作业引擎作业、这些作业执行的操作及其各自的文件系统访问方法的完整列表：

作业名称	作业描述	访问方法
AutoBalance	平衡群集中的可用空间。	驱动器 + LIN
AutoBalanceLin	平衡群集中的可用空间。	LIN
AVScan	防病毒服务器运行的病毒扫描作业。	树
ChangelistCreate	创建前后两个 SyncIQ 快照之间的更改列表。	更改列表
CloudPoolsLin	按照一个文件池策略的要求将数据归档到外部云提供商。	LIN
CloudPoolsTreewalk	按照一个文件池策略的要求将数据归档到外部云提供商。	树
Collect	回收由于节点或驱动器不可用而无法释放，同时出现各种故障情况的磁 盘空间。	驱动器 + LIN
ComplianceStoreDelete	SmartLock 合规模式垃圾数据收集作业。	树
重复数据消除	消除文件系统中的相同数据块。	树
DedupeAssessment	重复数据消除的优势的预演评估。	树

DomainMark	将路径及其内容与域关联。	树
DomainTag	将路径及其内容与域关联。	树
EsrsMftDownload	由 ESRS 管理的针对许可证文件的文件传输作业。	
FilePolicy	高效的 SmartPools 文件池策略作业。	更改列表
FlexProtect	重新构建和重新保护文件系统，以便从故障情形中恢复。	驱动器 + LIN
FlexProtectLin	重新保护文件系统。	LIN
FSAnalyze	收集与 InsightIQ 一起使用的文件系统分析数据。	更改列表
IndexUpdate	为 FilePolicy 和 FSAnalyze 作业创建并更新一个高效的文件系统索引。	更改列表
IntegrityScan	执行任何文件系统不一致性的在线验证和更正。	LIN
LinCount	扫描并计数文件系统逻辑信息节点 (LIN)。	LIN
MediaScan	扫描驱动器的介质级别错误。	驱动器 + LIN
MultiScan	同时运行 Collect 和 AutoBalance 作业。	LIN
PermissionRepair	更正文件和目录权限。	树
QuotaScan	更新在现有目录路径上创建的域的配额核算。	树
SetProtectPlus	应用默认文件策略。如果在群集中激活了 SmartPools，则此作业将被 禁用。	LIN
ShadowStoreDelete	释放与卷影存储相关的空间。	LIN
ShadowStoreProtect	用更高的请求保护级别保护由一个 LIN 引用的卷影存储。	LIN
ShadowStoreRepair	修复卷影存储。	LIN
SmartPools	在同一群集内的节点层之间运行和移动数据的作业。如果已经许可并配 置，还可以执行 CloudPools 功能。	LIN

SmartPoolsTree	在子树上实施 SmartPools 文件策略。	树
SnapRevert	将整个快照恢复到原点。	LIN
SnapshotDelete	释放与删除的快照相关的磁盘空间。	LIN
TreeDelete	直接从群集本身删除文件系统中的路径。	树
Undedupe	删除文件系统中的相同数据块重复消除功能。	树
升级	将群集升级到一个更高 CHCNMS 版本。	树
WormQueue	扫描 SmartLock LIN 队列	LIN

图 1：CHCNMS 作业引擎作业描述

虽然文件系统维护作业可根据默认设置运行（按计划运行，或者为了响应特定文件系统事件而运行），但是可通过配置任何作业引擎作 业的优先级（与其他作业相关）及其影响策略来管理这些作业。影响策略可以包括一个或许多影响间隔，这些间隔是给定周内的时间块。可对每个影响间隔进行配置，以使用单个预定义影响级别，这些级别规定了特定群集操作使用的群集资源量。可用作业引擎影响级别包括：**已暂停/低/中/高**等几个级别。

此粒度级别允许针对每个作业配置影响间隔和级别，以确保平稳顺畅的群集操作。得出的影响策略指示作业的运行时间及其可占用的资源。

此外，作业引擎作业根据范围（1 至 10）确定优先级；数值越小，优先级越高。这在概念上类似于 UNIX 调度应用工具“nice”。作业引擎允许同时运行最多三个作业。此并发作业执行按照以下标准进行管理：

- 作业优先级
- 排除集，不能同时运行的作业（即 FlexProtect 和 AutoBalance）
- 群集运行状况 — 群集处于降级状态时，大部分作业都无法运行。

文件系统结构

CHCNMS 文件系统基于 UNIX 文件系统(UFS)，因此是非常快速的分布式文件系统。每个群集均创建单独的命名空间和文件系统。这意味着，文件系统跨群集中的所有节点分布，同时可以通过连接到群集中任何节点的客户端访问。这里没有分区，也无需创建卷。CHCNMS 通过共享和文件权限，以及提供目录级配额管理的 SmartQuotas 服务，提供软件中的相同功能，而不是限制访问物理卷级别的可用空间和非授权文件。

CHCNMS 是带一个命名空间的真正单一文件系统。数据和元数据跨节点进行分条，以提供冗余和可用性。存储已实现对用户和管理员的完全可视化。文件树可有机增长，而无需计划或监管树的增长状况或用户的使用方式。管理员无需特别考虑分层文件到适当磁盘，因为CHCNMS SmartPools将自动处理，且无需中断单个树。无需特别考虑如何复制如此大型的文件树，因为CHCNMS SyncIQ服务会自动将文件树的传输并行到一个或多个备用群集，而不考虑文件树的形状或深度。

该设计应与命名空间聚合进行比较，这是让传统 NAS“看起来”具有单个命名空间的常用技术。通过命名空间聚合，文件仍需在单独卷中管理，但是简单的“粘合”层可通过符号链接将卷中的单个目录“粘合”到“顶级”树。在该模型中，LUN 和卷，以及卷限制仍存 在。文件必须以卷到卷的方式手动移动，以实现负载平衡。管理员必须注意树的布局。分层并不是无缝的，需要大量持续的干预。故障切换需要在卷之间镜像文件，这会降低效率并增加购买成本以及电源和冷却成本。整体来说，使用命名空间聚合时的管理员负担比使用简单的传统 NAS 设备的负担更大。这可防止此类基础架构增长到过大规模。

数据布局

CHCNMS 使用物理指针，扩展元数据，并在信息节点中存储文件和目录元数据。CHCNMS 逻辑信息节点 (LIN) 的大小通常为 512 字节，便 于其能够适应大多数硬盘驱动器进行格式化的本机扇区。

此外，还提供对 8 KB 信息节点的支持，以便支持现在使用 4 KB 扇区进行格式 化的更密集类型的硬盘驱动器。

B 树广泛用于文件系统，可扩展到数十亿个对象，并可对数据或元数据进行近即时查找。CHCNMS 是完全对称的高度分布式文件系统。数据和元数据始终跨多个硬件设备冗余。使用跨群集节点的擦除代码保护数据，这将创建极为高效的群集，允许五个节点或更多节点的群 集具有 80%或更高的原始与可用比例。元数据（通常占系统的不到 1%）被镜像到群集，以获得性能和可用性。由于 CHCNMS 不依赖于 RAID，因此在超出群集默认的文件或目录级别上，冗余量可由管理员选择。在对等体系结构中，元数据访问和锁定任务由所有节点进行 集中平等的管理。该对称性是实现体系结构的精简性和恢复能力的关键。没有单一元数据服务器、锁定管理器或网关节点。

由于 CHCNMS 必须同时从几个设备访问数据块，因此用于数据和元数据的地址方案在物理级别上由（节点，驱动器，偏移）的元组编制索引。例如，如果 12345 是节点 3 磁盘 2 上数据块的块地址，则它应读取（3,2,12345）。群集内的所有元数据将进行多层镜像以保护数据，至少达到关联文件的冗余级别。例如，如果文件的擦除代码保护设置为“+2n”，这表示文件可以承受两次同时发生的故障，然后访问该文 件所需的全部元数据将 3 倍镜像，以便其也可以承受两次故障。文件系统既定允许任何结构使用该群集中任何节点上的数据块。

其他存储系统通过 RAID 和卷管理层发送数据，导致数据布局效率低下，并提供未经优化的数据块访问。CHCNMS 直接控制文件的放置，深入到群集上任何驱动器的扇区级别。这可提供最佳数据放置和 I/O 模式，并避免不必要的读取-修改-写入操作。通过以逐个文件的方式将数据放置到磁盘，CHCNMS 能够灵活控制系统、目录甚至文件级别的分条类型以及存储系统的冗余级别。传统存储系统要求将整个 RAID 卷用于某种性能类型和保护设置。例如，对于某个数据库，一组磁盘可能以 RAID 1+0 保护的方式排列。这使得优化整个存储设备的磁盘轴使用非常困难，同时还导致无法满足业务要求的不灵活设计。CHCNMS 允许在任何时候、完全在线地进行单独调整和灵活更改。

文件写入

CHCNMS 软件在所有节点上均等运行 — 创建了跨每个节点运行的单一文件系统。任何一个节点均无法控制或“主导”群集；所有节点真正对等。

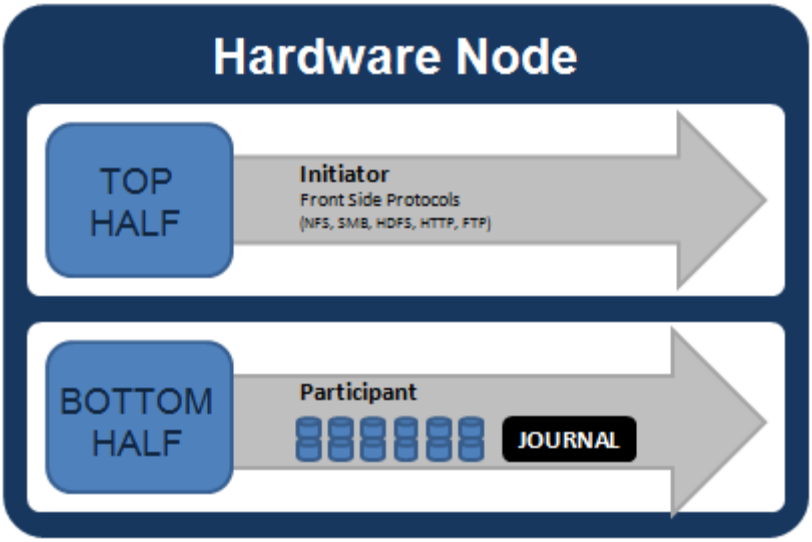

图 6：I/O 涉及的节点组件模型

如果我们能够从高层面查看 I/O 涉及的某一群集每个节点中的所有组件，它们看起来将与以上图 6 类似。我们将堆栈分成“顶”层（称 为启动器）和“底”层（称为参与者）。这种划分方法被用作任何给定读取或写入分析的“逻辑模型”。在物理级别上，节点中的 CPU 和 RAM 缓存同时处理启动器和参与者任务，以便在整个群集中进行 I/O。以上图表中未包含缓存和分布式锁定管理器，以保持简洁。后 文部分将对其进行介绍。

当客户端连接到节点以写入文件时，它将连接到该节点的上半部分或启动器。在写入节点（磁盘）下半部分或参与者之前，文件被分成称作条带的较小逻辑块。使用写入合并器的故障安全缓冲区将用于确保有效写入，同时避免读取-修改-写入操作。每个文件块的大小称为条带单元大小。

CHCNMS 通过软件擦除代码或镜像技术，可跨所有节点对数据进行分条（不仅仅限于跨磁盘），并保护文件、目录和关联元数据。对于数 据，CHCNMS 可以使用（根据管理员的判断）Reed-Solomon 擦除代码系统进行数据保护或镜像（不常用）。当对用户数据应用镜像时，更多用于高事务性能情况。

批量用户数据通常使用擦除代码，因为它提供了极高性能而不牺牲磁盘效率。擦除代码可在具有五个或更多节点的原始磁盘上提供超过 80%的效率，甚至在大型群集上也可实现，同时还能提供四倍级别冗余。任何给定文件的条带宽度是文件要 写入的节点（不是驱动器）数量。这由群集中的节点数量、文件大小和保护设置决定（例如，+2n）。

CHCNMS 使用高级算法确定数据布局，以提供最高效率和性能。当客户端连接节点时，该节点的启动器将在该文件的写入数据布局中发挥“统领”作用。数据、擦除代码（ECC）保护、元数据和信息节点都分布在群集的多个节点上，甚至跨节点内的多个驱动器分布。以下图 7 显示了跨 3 节点群集中所有节点的文件写入。

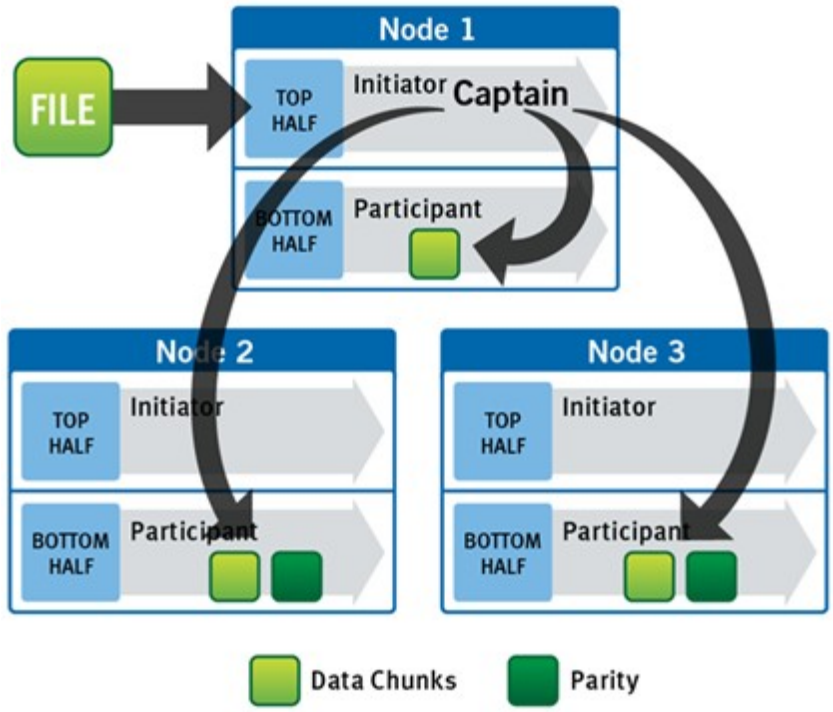

图 7：3 节点群集中的文件写入操作

CHCNMS 使用后端网络自动跨群集中的所有节点进行数据分配和分条，因此无需任何额外处理。随着数据被写入，它将受到指定级别的保护。进行写入时，CHCNMS 将数据拆分为称作保护组的原子单元。将在保护组中构建冗余，这样在每个保护组安全的情况下，整个文件都是安全的。对于受擦除代码保护的文件，保护组由一系列数据块以及这些数据块的一组擦除代码组成；对于镜像文件，保

护组由一组数 据块的所有镜像组成。写入时，CHCNMS 能够动态切换文件使用的保护组类型。这可提供许多其他功能，例如在群集出现会阻止使用所需 擦除代码数量的临时节点故障时，允许系统继续运行而不阻止。镜像可在这些情况下临时使用，以允许继续写入。当恢复节点到群集时，这些镜像的保护组将无缝地自动还原到擦除代码保护状态，而无需管理员干预。

CHCNMS 文件系统块大小是 8KB。小于 8KB 的文件使用完整的 8KB 块。根据数据保护级别的不同，此 8KB 文件最后可能使用超过 8KB 的数据空间。然而，本文稍后部分会对数据保护设置进行详述。CHCNMS 能够以极高性能支持具有数十亿个小文件的文件系统，因为所有磁盘结构均旨在扩展到此类大小，且无论对象总数如何，都提供对任何对象的近即时访问。对于较大文件，CHCNMS 可以利用使用多个连续 8KB 数据块的优势。在这些情况下，可将最多十六个连续块分条到单个节点的磁盘。如果某个文件的大小是 32KB，则使用四个连续的 8KB 块。

对于更大的文件，CHCNMS 可以利用由 16 个连续块组成的条带单元实现连续性能最大化，每个条带单元总共 128KB。写入期间，数据被 分为条带单元，同时将作为一个保护组分布到多个节点。由于数据跨群集分布，因此必要时，擦除代码或镜像将在每个保护组内分布，以确保文件始终受到保护。

CHCNMS AutoBalance 的主要功能之一是重新分配和重新平衡数据，并尽可能提高存储空间的可用性和效率。在大部分情况下，可以增加较大文件的条带宽度，以利用新的可用空间（在添加节点后），并让磁盘分条更加高效。AutoBalance 保持高磁盘效率并自动消除磁盘“热点”。

“统领”节点的上半部分启动器使用经过修改的两阶段提交事务，以安全地将写入分布到整个群集的多个 NVRAM 中，如下图 8 所示。

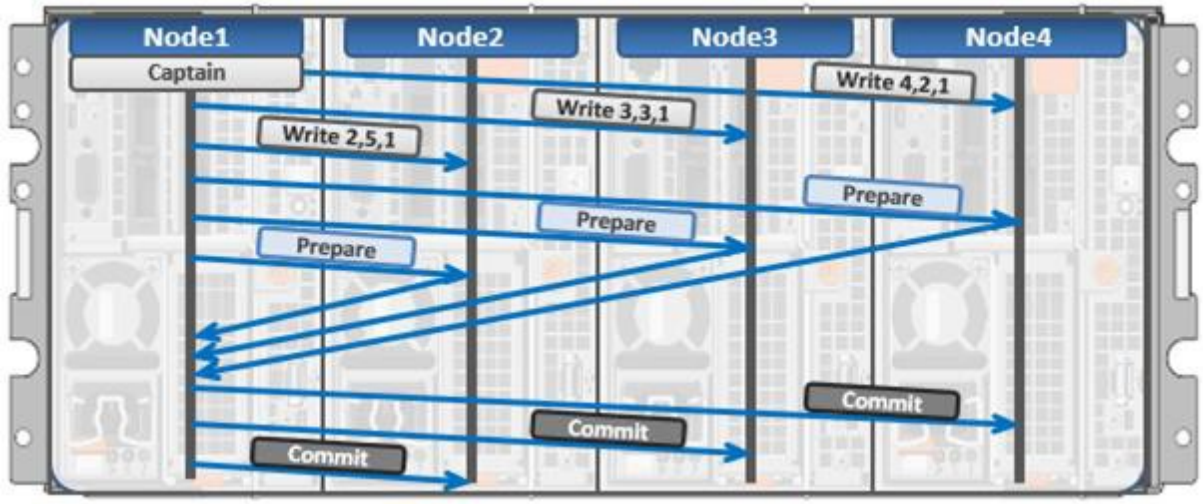

图 8：分布式事务和两阶段提交

两阶段提交涉及每个在特定写入中拥有数据块的节点。该机制依赖于 NVRAM 日志记录所有在存储群集的每个节点发生的事务。并行使 用多个 NVRAM 可实现高吞吐量写入，同时保证数据安全，防止所有形式的故障，包括电源故障。如果节点在事务中发生故障，事务会在不涉及该节点的情况下立即重启。当节点恢复时，节点唯一所需的操作是重放来自 NVRAM 的日志（需要数秒或数分钟），某些情况 下会让 AutoBalance 重新平衡事务涉及的文件。无需昂贵的“fsck”或“disk-check”进程。无需进行持续的重新同步。写入决不会因故障而受阻。这种获得专利的事务系统是 CHCNMS 消除单点故障甚至多点故障的方式之一。

在写入操作中，启动器“统领”或编排数据和元数据布局；创建擦除代码；以及进行锁定管理和权限控制的正常操作。来自 Web 管理或 CLI 界面任何点的管理员均可优化 CHCNMS 作出的布局决策，从而更好地适合工作流。管理员可选择以下逐文件或目录级别的访问模式：

- 并发：优化群集上的当前负载，特点是有许多并发客户端。本设置为混合工作负载提供最佳行为
- 流式：优化单个文件的高速流，例如使用单个客户端进行快速读取
- 随机：通过调整分条并禁用任何预取高速缓存的使用，优化对文件不可预测的访问

CHCNMS 还包括实时自适应预取，为具有可识别访问模式的文件提供最佳读取性能，无需任何管理干预。

CHCNMS 高速缓存

CHCNMS 缓存基础架构设计的前提是将群集中各节点上的缓存聚集成一个可全局访问的内存池。为此，CHCNMS 使用与非一致内存访问（NUMA）类似的高效信息传输系统。这可让所有节点的内存缓存面向群集中的每个节点都可用。远程内存通过内部互连来访问，其延迟性比访问硬盘低得多。

对于远程内存访问，CHCNMS 利用冗余、订阅不足的平面以太网网络，实质上是分布式系统总线。远程内存访问的速度虽不如本地内存，但由于 40 Gb 以太网的延迟性较低，因此其速度仍然非常快。CHCNMS 缓存子系统在群集中保持一致。这意味着如果相同的内容存在于多个节点的专用缓存中，那么经缓存的这一数据将在所有实例中保持一致。CHCNMS 利用 MESI 协议来维持缓存一致性。此协议将实施“无效写入”策略，以确保所有数据将在整个共享缓存中保持一致。

CHCNMS 使用多达三级读取缓存，外加一个受 NVRAM 支持的写入缓存或聚合器。这些缓存及其高级别的交互关系如下图所示。

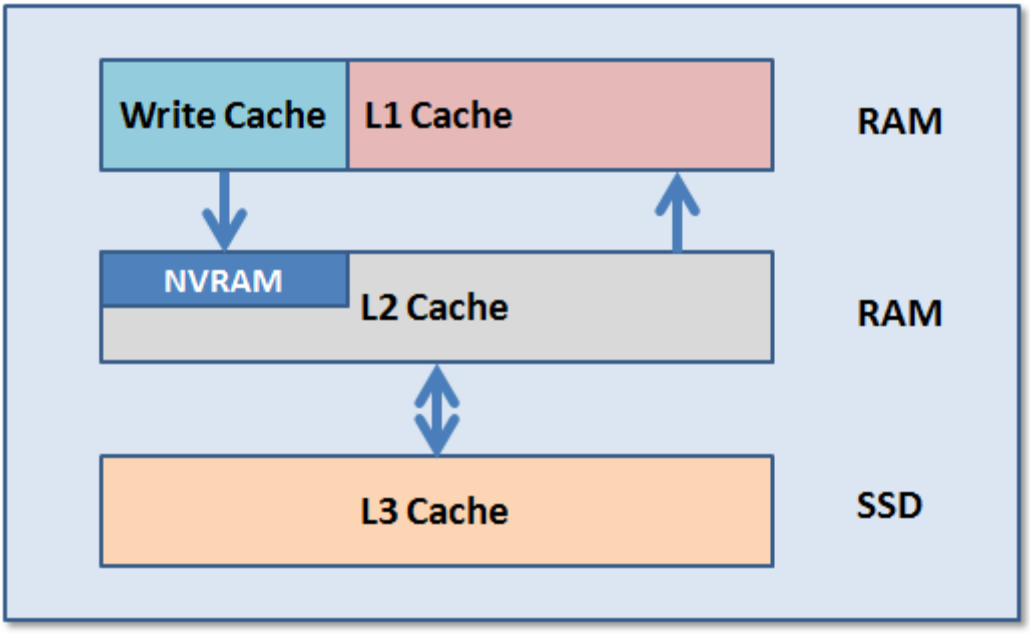

图 9：CHCNMS 高速缓存层次结构

前两类读取缓存（1 级（L1）和 2 级（L2））基于内存（RAM），类似于处理器（CPU）中使用的缓存。这两个高速缓存层存在于所有平台存 储节点中。

名称	类型	永久存储	描述
L1 缓存	RAM	易失性	又称前端高速缓存，将通过前端网络保留经由客户端请求的干净、群集一 致的文件系统数据复制和元数据块
L2 高速缓存	RAM	易失性	后端缓存，包含本地节点上文件系统数据和元数据的干净拷贝
SmartCache/ 写聚合器	NVRAM	非易失性	电池供电的持久性 NVRAM 日志缓存，可将任何挂起的写入缓冲到未提 交到磁盘的前端文件中。
SmartFlash L3 高速缓存	SSD	非易失性	包含从 L2 缓存中回收的文件数据和元数据块，从而有效增加 L2 缓存容量。

CHCNMS 高速缓存一致性

CHCNMS 缓存子系统在群集中保持一致。这意味着如果相同的内容存在于多个节点的专用缓存中，那么经缓存的这一数据将在所有实例中 保持一致。例如，请考虑以下初始状态和系列事件：

1. 节点 1 和节点 5 各有一个位于共享缓存中某一地址的数据拷贝。
2. 节点 5 为响应写入请求而使节点 1 的拷贝无效。
3. 节点 5 随后更新此值。
4. 节点 1 必须从共享缓存重新读取数据，以获取更新后的值。

CHCNMS 利用 MESI 协议来维持缓存一致性。此协议将实施“无效写入”策略，以确保所有数据将在整个共享缓存中保持一致。下图说明了缓存中的数据会出现的各种状态，以及在这些状态间

的转换。图中的不同状态包括：

- M 修改：数据仅存在于本地高速缓存中，且相对于共享高速缓存中的值已经更改。经修改的数据通常被称为脏数据
- E 独占：数据仅存在于本地高速缓存中，但与共享高速存储中的数据相匹配。此类数据通常被称为干净数据
- S 共享：本地高速缓存中的数据也可能位于群集中的其他本地高速缓存中
- I 无效：失去对数据的锁定（独占或共享）

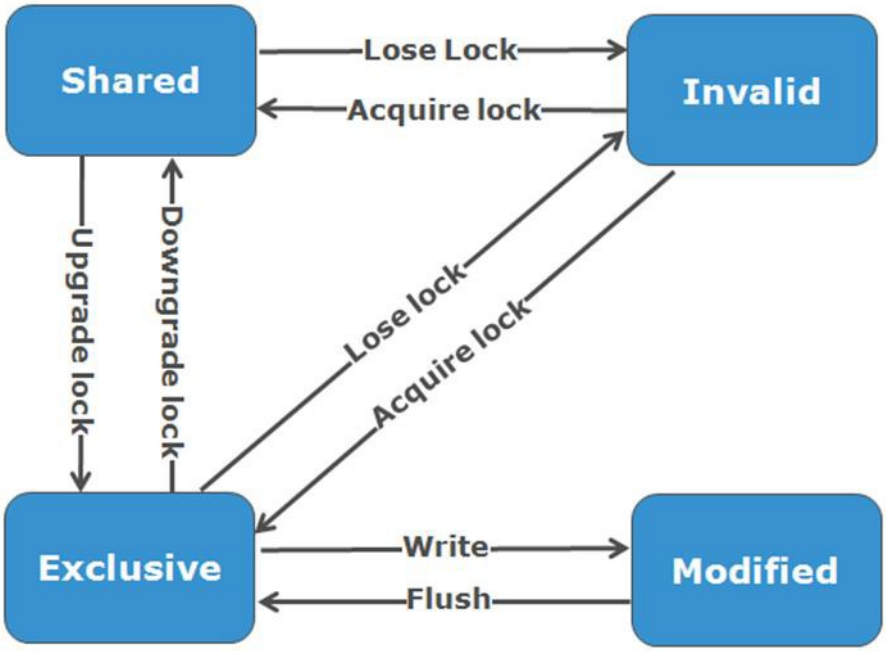

图 10：CHCNMS 高速缓存一致性状态图

一级高速缓存

一级高速缓存（L1）或前端缓存是最靠近客户端或启动器使用的并与此节点连接的协议层（例如 NFS、SMB 等）的内存。L1 缓存的主要目的在于从远程节点预取数据。系统对文件逐一预取数据，这一过程已得以优化，以降低与节点后端网络相关的延迟。由于后端互连延 迟相对较小，因此 L1 缓存的大小和每个请求通常存储的数据量少于 L2 缓存。

L1 缓存中包含从群集中其他节点检索到的数据，因此又称为远程缓存。其在整个群集中保持一致，但仅由其所在的节点使用，无法供其 他节点访问。存储节点上 L1 缓存中的数据在使用后遭到大规模丢弃。L1 缓存使用基于文件的寻址，在此过程中数据将通过一个偏移量存取到文件对象中。

L1 缓存引用与启动器相同的节点上的内存。此类缓存仅可由本地节点访问，而且通常不是数据的主拷贝。这与 CPU 核心上的 L1 缓存类似，可在其他核心写入主内存时变得无效。

如上所述，L1 缓存一致性通过类似于 MESI 的协议使用分布式锁定进行管理。

CHCNMS 还使用一个专用的信息节点缓存，在其中保留最近请求的信息节点。信息节点缓存时常会对性能造成较大影响，这是因为客户端 常常缓存数据，而且很多网络 I/O 活动主要请求文件属性和元数据，而这些都能够从缓存的信息节点中快速返回。

L1 高速缓存在不含任何磁盘驱动器的群集加速器节点中有着不同的应用。相反，整个读取缓存是 L1 缓存，这是因为所有数据都是 从其他存储节点中提取。此外，缓存老化基于“最近最少使用(LRU)”回收策略，与存储节点的 L1 缓存中通常使用的“落后”算法截然相反。由于加速器的 L1 缓存较大，而且其中的数据极有可能被再次请求，因此数据块并不会在使用后立即从缓存中删除。但是，元数据和更新繁重工作负载并不能从中获得太多收益，加速器的缓存仅对与节点直接连接的客户端有益。

二级高速缓存

二级缓存 (L2) 或后端缓存是指存储特定数据块的节点上的本地内存。L2 缓存可从群集中任何节点全局访问，用于通过避免直接从磁盘驱 动器请求寻道降低读取操作延迟。因此，远程节点预取到 L2 缓存中以供使用的数据量远多于 L1 缓存中的数据量。

L2 缓存又称本地缓存，这是因为其中包含从所在节点的磁盘驱动器中检索到的数据，然后将这些数据用于远程节点请求。L2 缓存中的数据将根据“最近最少使用 (LRU)”算法回收。

L2 缓存中的数据由本地节点使用一个偏移量寻址到该节点本地的磁盘驱动器中。由于该节点了

解远程节点请求的数据在磁盘中的位置，因此是检索以远程节点为目标的数据的快捷方法。远程节点通过查找特定文件对象的数据块地址访问 L2 缓存。如上所述，此处不需要进行 MESI 无效操作，缓存将在写入过程中自动更新，并由事务系统和 NVRAM 保持一致。

三级高速缓存

可选的第三层读取缓存称为 SmartFlash 或 3 级缓存（L3），也可在包含固态驱动器（SSD）的节点上配置。SmartFlash (L3) 是一种清除缓存，当 L2 缓存数据块在内存中老化时就会填充到 SmartFlash 中。使用 SSD 而非传统文件系统存储设备进行缓存有诸多优势。例如，将 SSD 预留用于缓存时，整个 SSD 都将得以使用，而且写入也将以高度线性和可预测的方式发生。相比常规文件系统的使用，特别是包含随机写入工作负载的情形，这将大幅优化利用率，同时显著降低磨损，提高耐用性。相比将 SSD 用作存储层，将 SSD 用于缓存还能大为简化 SSD 容量调整，且更不容易产生错误。

下图说明了客户端与 CHCNMS 读取缓存基础架构和写聚合器的交互方式。L1 缓存仍与其所需的任意节点上的 L2 缓存交互，而且 L2 缓存与存储子系统和 L3 缓存交互。L3 缓存存储于节点内的 SSD 上，而且同一节点池中的每个节点均启用 L3 缓存。

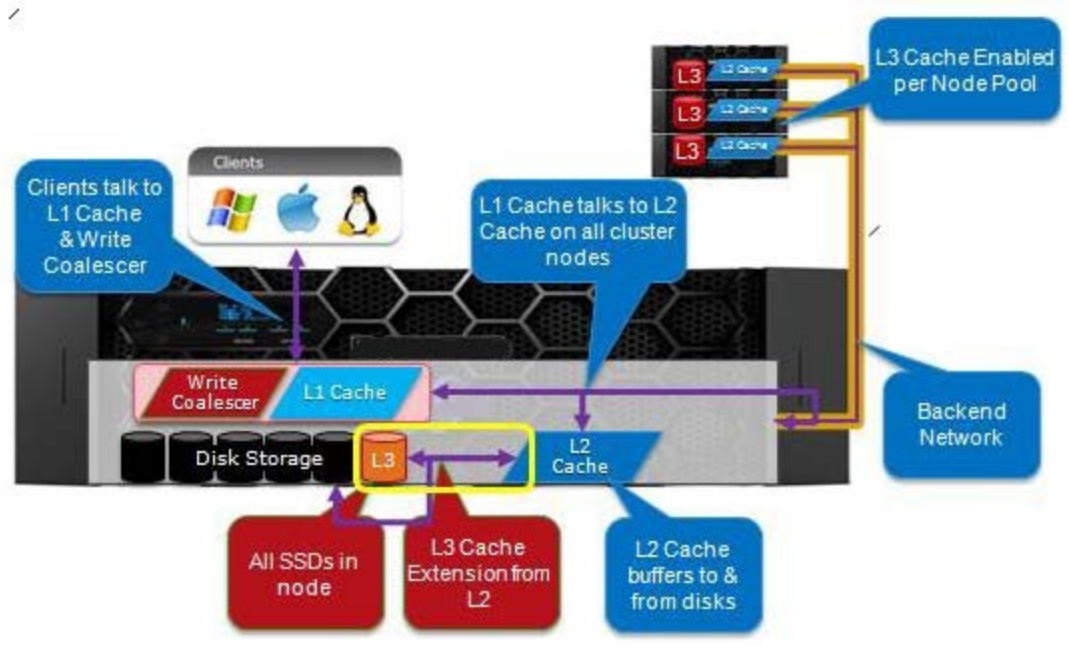

图 11：CHCNMS L1、L2 和 L3 高速缓存体系结构

CHCNMS 规定要跨群集中的多个节点（也可能是节点中的多个驱动器）写入文件，所以所有读请求都涉及读取远程（也可能是本地）数据。收到来自客户端的读取请求时，CHCNMS 将确定所请求的数据是否位于本地缓存中。立即读取驻留在本地缓存中的任何数据。如果请求的数据不在本地缓存中，则从磁盘中读取。对于不在本地节点上的数据，则请求从所在的远程节点读取。在其他的每个节点上，执行另一缓存查找操作。立即返回缓存中的任何数据，不在缓存中的数据将从磁盘中检索。

从本地和远程缓存（还有可能从磁盘）中检索数据后，返回到客户端。 在本地和远程节点上完成读取请求的高级别步骤为：

本地节点（接收请求的节点）：

1. 确定是否部分请求数据位于本地 L1 缓存中。如果是，则返回到客户端。
2. 如果不在本地缓存中，则从远程节点请求数据。

远程节点：

1. 确定请求数据是否位于本地 L2 或 L3 缓存中。如果是，则返回到请求节点。
2. 如果不在本地高速缓存中，则从磁盘读取，然后返回到请求节点。

写入高速缓存可加速将数据写入群集的过程。这可通过分批较小写入请求并以较大数据块发送到磁盘的方式实现，从而消除了大量磁盘写入延迟。当客户端写入群集时，CHCNMS 临时写入数据到启动器节点上基于 NVRAM 的日志缓存，而不是立即写入磁盘。然后，CHCNMS 可在稍后更加方便的时间将这些缓存的写入刷新到磁盘中。此外，这些写入还镜像到参与者节点的 NVRAM 日志，以满足文件的保护要求。因此，在群集拆分或节点意外中断的情况下，未提交的缓存写入将获得完全保护。写缓存的操作如下：

- NFS 客户端为具有+2n 保护的文件向节点 1 发送写入请求。

- 节点 1 接受写入到其 NVRAM 写缓存（快速路径），然后将写入镜像到参与者节点的日志文件进行保护。
- 写入确认立即返回 NFS 客户端，从而避免写入磁盘延迟。
- 随着节点 1 写入高速缓存填满，它将定期刷新，并通过采用相应擦除代码 (ECC) 保护 (+2n) 的两阶段提交过程（上文所述）向磁盘提交写入。
- 清除写缓存和参与者节点日志文件，同时可以接受新写入。

文件读取

在群集中，数据、元数据和信息节点都分布在多个节点上，甚至跨节点内的多个驱动器分布。当读取或写入群集时，连接客户端的节点 将在操作中发挥“统领”作用。

在读取操作中，“统领”节点从群集中的各节点收集所有数据，并以聚合方式提交给请求人。

由于使用经过成本优化的行业标准硬件，因此，群集提供了高缓存与磁盘比（每节点多 GB），可根据需要对读取和写入操作进行动 态分配。此基于 RAM 的缓存在群集中的所有节点上都是统一、连续的，可使某个节点的客户端读取请求从已在另一个节点上处理的 I/O 中受益。这些高速缓存数据块可通过低延迟底板中的任何节点快速访问，从而实现大型的高效 RAM 高速缓存，这将显著提高读取性能。

随着群集变大，缓存的优势将会增加。因此，群集上的磁盘 I/O 量通常明显低于传统平台，从而有利于减少延迟并提供更好的用户体验。

对于标记为并发或流访问模式的文件，CHCNMS 可以基于 SmartRead 组件使用的探试程序，利用预取数据的优势。SmartRead 可以从 L2 缓存创建数据“管道”，预取到“统领”节点上的本地“L1”缓存。这将极大改进所有协议的顺序读取性能，同时意味着直接来自 RAM 的读取在毫秒内。对于高顺序情况，SmartRead 可以非常积极地提取预取，允许以极高的数据率读取或写入单个文件。

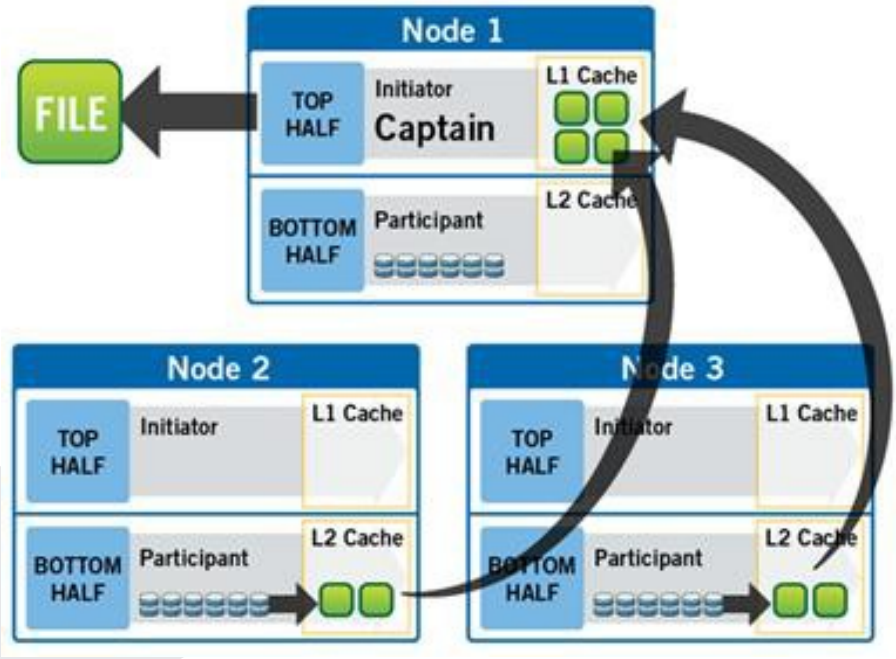

图 12：3 节点群集中的文件读取操作

图 10 说明了 SmartRead 如何读取与 3 节点群集中节点 1 连接的客户端所请求的顺序访问的非高速缓存文件。

1. 节点 1 读取元数据以确定文件数据的所有数据块存在的位置。

2. 节点 1 还检查其 L1 缓存，以查看它是否拥有正在请求的文件数据。

3. 节点 1 构建了读取管道，发送并发请求给具有文件数据块的所有节点，以从磁盘检索该文件数据。

4. 每个节点将文件数据块从磁盘推送到其 L2 高速缓存（或 L3 SmartFlash 高速缓存，如果可用），并将该文件数据传输到节点 1。

5. 节点 1 记录到达 L1 缓存的传入数据，同时向客户端提供文件。同时，预取进程继续进行。

6. 对于高度连续的情形，L1 缓存中的数据可能会选择“落后”，以释放 RAM，从而满足其他 L1 或 L2 缓存需求。

SmartRead 的智能缓存可提供极高的读取性能，并具有高级并发访问权限。重要的是，节点 1 可以从节点 2（在低延迟群集互连上）的 缓存中快速获得文件数据，其速度要比访问本地磁盘更

快。SmartRead 算法控制预取的进度（禁用随机存取情形的预取）以及数据在高 速缓存中的滞留时间，同时优化缓存数据的位置。

锁定与并发

CHCNMS 具有一个完整的分布式锁定管理器，该管理器可以将锁定状态编列到存储群集内所有节点的数据上。锁定管理器具有高可扩展性，允许多个锁定属性，从而支持文件系统锁定和群集一致协议级别锁定，如 SMB 共享模式锁定或 NFS 咨询模式锁定。另外，CHCNMS 还支持委派锁定，如 CIFS oplock 锁定和 NFSv4 委派锁定。

群集内各节点都是一个锁定资源的协调器，一般按照高级散列算法向协调器分配可锁定资源。该算法的设计方式在于，协调器通常在不同于请求的启动器的节点上结束。某个文件请求锁定时，可以是共享锁定（允许多个用户同时共享该锁定，通常用于读取）或独占锁定（允许单个用户随时使用，通常用于写入）。

下图 13 举例说明不同节点的线程如何向协调器请求锁定。

1. 节点 2 被指定为这些资源的协调器。
2. 节点 4 的线程 1 和节点 3 的线程 2 同时向节点 2 请求关于文件的共享锁定。
3. 节点 2 检查请求的文件是否存在独占锁定。
4. 如果不存在独占锁定，节点 2 则向节点 4 线程 1 和节点 3 线程 2 授予请求文件的共享锁定。
5. 节点 3 和节点 4 正在读取请求文件。
6. 节点 3 和节点 4 读取文件时，节点 1 线程 3 请求该文件的独占锁定。
7. 节点 2 检查节点 3 和节点 4 是否可以回收共享锁定。
8. 此时，节点 3 和节点 4 仍在读取，因此，节点 2 请求节点 1 线程 3 等候片刻。
9. 节点 1 的线程 3 受限制，直到节点 2 授予独占锁定为止，此后会完成写入操作。

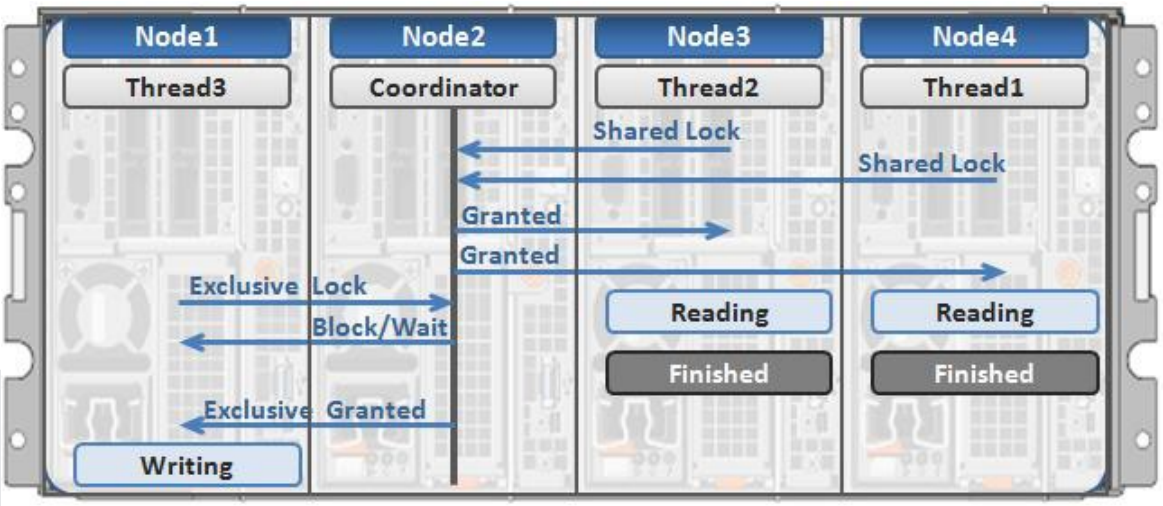

图 13：分布式锁定管理器

多线程 I/O

随着用于服务器虚拟化和企业应用程序支持的大型 NFS 数据存储区不断增加，也带来了对大型文件的高吞吐量和低延迟需求。为此，CHCNMS Multi-writer 支持多个线程向不同文件并发写入。

在上面示例中，并发写入访问大型文件时，会受到独占锁定机制的限制，因为该机制的作用对象是整个文件级别。为避免这种潜在的瓶颈，CHCNMS Multi-writer 先将大文件细分为单独的区域，并向各区域授予独占写入锁定，进而提供了更为精确的写入锁定，因此，它与整个大文件的处理方式截然相反。这样，多个客户端可以同时向同一文件的不同部分写入。

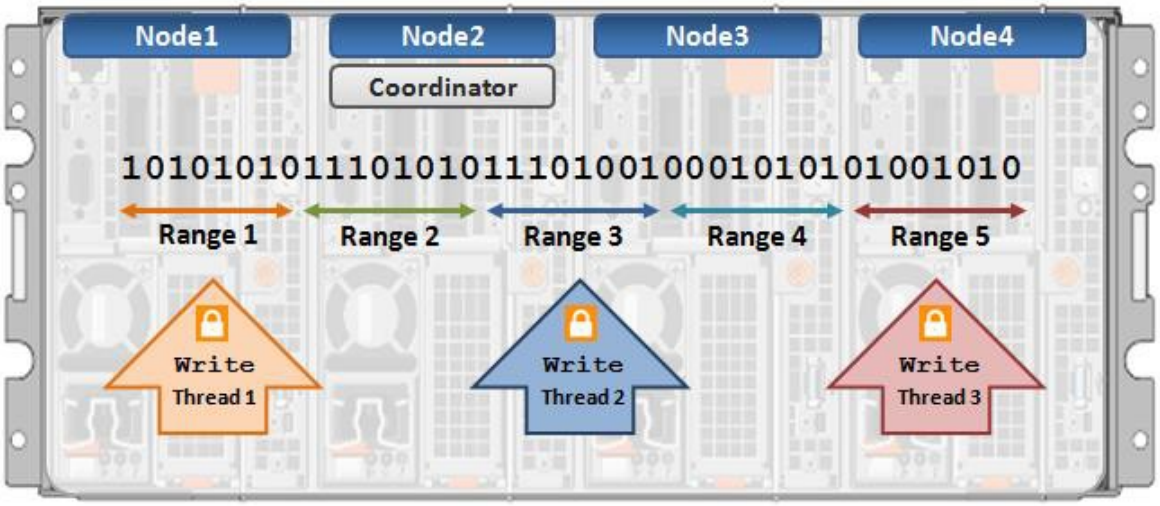

图 14：多线程 I/O 编写器

数据保护

断电

文件系统日志用于存储有关文件系统更改的信息，在系统出现故障或崩溃（如断电）后，该日志旨在促进实现快速的一致性恢复。节点 或群集从断电或其他故障中恢复后，文件系统将重放日志条目。如果没有日志，则文件系统在故障（“fsck”或“chkdsk”操作）发生 后需要单独检查各种潜在更改；在大型文件系统中，此操作可能需要较长时间。

CHCNMS 是一个日志记录的文件系统，在该系统中，每个节点都包含用于保护文件系统的未提交写入的电池供电 NVRAM 卡。NVRAM 卡 的电池充电器可以持续使用很多天，期间无需充电。节点启动后，将检查日志，如果日志记录系统认为有必要，则节点还要有选择地向 磁盘重放事务。

CHCNMS 只有在可以确保尚未记录系统内的所有事务的情况下，才会进行装载。例如，如果未遵循适当的关机程序，且 NVRAM 电池电量 耗尽，则很可能丢失事务；为防止一切潜在问题，节点不会装载文件系统。

硬件故障与仲裁

为了群集能够正常运行并接受数据写入，节点的仲裁应处于活动和响应状态。仲裁定义为简单多数：含节点的群集必须具有一半加 1 个在线节点才允许写入。例如，在七节点的群集中，仲裁必须为四个节点。如果某个节点或一组节点启动并响应，但不属于仲裁的一员，则该节点或该组节点会在只读状态下运行。

如果群集应临时拆分成两个群集，则 CHCNMS 会利用仲裁防止出现“裂脑”状况。按照仲裁规则，体系结构可以确保，无论有多少个节点出现故障或恢复在线，如果进行写入，都会与先前完成的写入保持一致。另外，仲裁还规定需要的节点数，以便转至特定的数据保护 级别。针对基于擦除代码的保护级别+，群集必须包含至少 2+1 个节点。例如，+3n 配置至少应有七个节点；这样一来，在维护运行完 全正常的群集的四节点仲裁过程中，允许同时丢失三个节点。如果群集降至仲裁

之下，文件系统将自动替换为受保护的只读状态并拒绝写入，但仍允许读取访问可用数据。

硬件故障 - 添加/删除节点

组管理协议 (GMP) 系统不仅能随时全面了解群集状态，还能确保持续查看整个群集的所有其他节点状态。如果通过群集互连无法访问单 个或多个节点，则群集会“拆分”或删除该组。所有节点将解析为一致的新群集视图。(可以将其看作是群集被划分为两个单独的节点组，但是请注意，只有一个组可以具有仲裁)。在这种拆分状态下，文件系统中的所有数据都可供访问，对于维护仲裁的一方，可对其修改。利用群集中存储的冗余，对所有存储在“故障”设备中的数据进行重建。

如果该节点再次变为可以访问，则表示发生了“合并”或添加，从而将节点带回群集。(这两组又合并为一组)。无需重建或重新配置，节点也可以重新加入群集。这不同于需要重建驱动器的硬件 RAID 阵列。拆分过程中，如果有些保护组被覆盖并将其转换为更窄的 条带，则 AutoBalance 会对某些文件重新分条以提高效率。

CHCNMS 作业引擎还包含一个名为 Collect 的进程，此进程可以充当一个孤立收集器。如果在写入操作期间出现群集拆分，则为文件分配 的部分数据块可能需在仲裁端进行重新分配。这会“孤立”非仲裁端的已分配数据块。当群集重新合并时，Collect 作业将通过并行的标 记清除扫描来查找这些孤立的数据块，并将其作为群集的可用空间进行回收。

可扩展的重建

CHCNMS 不会依赖硬件 RAID 进行数据分配或故障后数据重建。相反，CHCNMS 会直接管理文件数据的保护，如果出现故障，它会采用并行方式重建数据。通过直接从磁盘上线性读取信息节点数据，CHCNMS 可以在规定时间内确定受故障影响的文件。作业引擎将该组受影响的文件分配到分布在群集节点中的一组工作线程。工作节点并行修复这些文件。这意味着，随着群集规模增大，重建故障数据所需时间 会减少。随着群集规模增大，这在维护其恢复能力方面具有巨大的效率优势。

虚拟热备盘

大多数基于 RAID 的传统存储系统都需要调配一个或多个“热备盘”驱动器，才能允许故障驱动器完成独立恢复。热备盘替换 RAID 集 的故障驱动器。出现更多故障之前，如果未更换这些热备盘，系统将面临灾难性数据丢失的风险。CHCNMS 避免使用热备盘驱动器，它只从系统的可用自由空间借取，进而从故障中恢复；此技术称为虚拟热备盘。在这种情况下，群集可以完全自我修复，无需人为干预。管理员可以创建一个虚拟热备盘储备，这样一来，即便用户正在写入，系统也能完成自我修复。

通过擦除代码提供文件级数据保护

群集可以容许同时出现一个或多个组件故障，且不阻止群集提供数据服务。为此，CHCNMS 使用基于擦除代码的保护、Reed-Solomon 纠错（N+M 保护）或镜像系统来保护文件。数据保护应用在文件级别上的软件中，从而让系统能够集中恢复因故障受损的文件，无需检查 并修复整个文件集或卷。CHCNMS 元数据和信息节点始终由镜像而非 Reed-Solomon 编码进行保护，其保护级别至少是其所参考数据的保 护级别。

由于所有数据、元数据以及保护信息分布在群集节点上，因此，群集不需要专用奇偶校验节点或驱动器、专用设备或成套设备来管理元数据。这便确保了节点无法生成单点故障。所有节点在即将执行的任务中平等共享，因此，端对端体系结构中具有完美的对称和负载平衡。

CHCNMS 具有几种不同级别的可配置数据保护设置，您可以随时修改这些设置，而群集或文件系统无需脱机。

对于受擦除代码保护的文件，我们声明每个保护组在 N+M/b 级别上受保护，其中 N>M 且 M>=b。值 N 和 M 分别表示在保护组范围内用于数据和擦除代码的驱动器数量。值 b 与布设该保护组所使用的数据条带数量有关，下面涉及到该值。常见且易懂的情形：当 b=1， 则保护组包含 N 个驱动器的数据；M 个驱动器的冗余，这些值存储在擦除代码中；该保护组应正确布设在一

组节点的一个条带上。这允 许该保护组的 M 个成员同时出现故障，但仍提供 100% 数据可用性。M 擦除代码成员由 N 个数据成员计算得出。下图 13 显示常规 4+2 保护组的情形（N=4、M=2、b=1）。

由于 CHCNMS 跨节点对文件进行分条处理，这表示 N+M 分条文件可以容许同时出现 个节点故障，而不会降低可用性。因此，CHCNMS 在 任何类型故障中都具有恢复能力，不论是驱动器、节点或节点内的组件（比如，卡）。此外，一个节点视为一个单一故障，而无论此节 点内的故障组件数量或类型如何。因此，如果某个节点的五个驱动器发生故障，则仅视为完成 N+M 保护出现的单一故障。

CHCNMS 专门提供不同级别的 M，最多达四种，因而具有四重故障保护。这远远超出了当前常用的 RAID 最高级别，这是对 RAID-6 的双 重故障保护。由于存储的可靠性随着冗余的增加呈几何级数增长，因此，+4n 保护的可靠性比传统硬件 RAID 的可靠性高出几个数量级。 这些额外的保护意味着可以信心十足地添加大容量 SATA 驱动器（如 4 TB 和 6 TB 驱动器）。

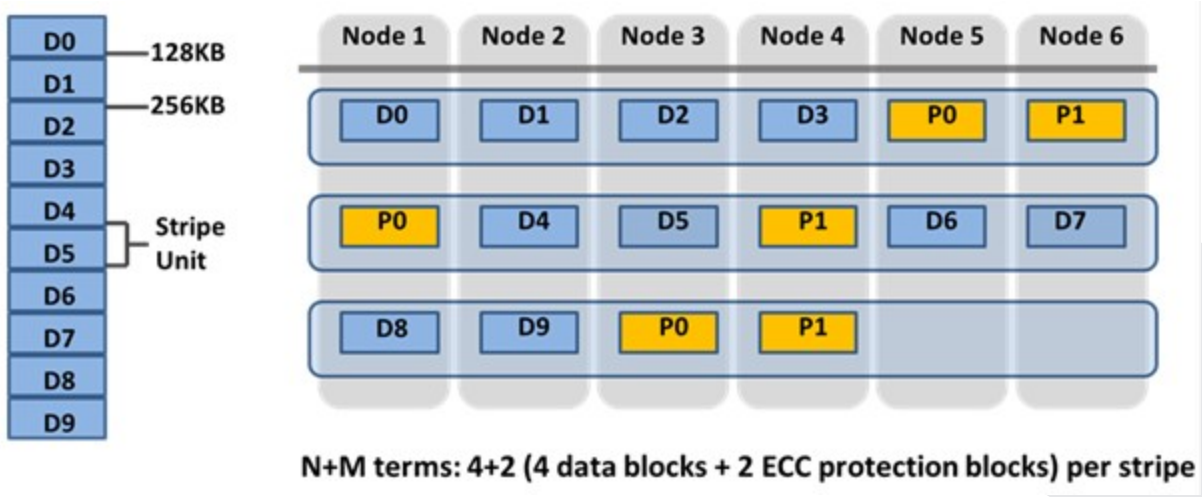

图 15：CHCNMS 冗余 - N+M 擦除代码保护

较小群集可使用 +1n 保护方案进行保护，但这表示可以恢复单个驱动器或节点，而无法恢复两个不同节点中的两个驱动器。驱动器故障 很可能千百倍地超过节点故障。对于具有大型驱动器的群集，即使可接受单节点可恢复性，最好为多个驱动器故障提供保护。

要针对具有双磁盘冗余和单节点冗余的情况提供保护，我们可建立大小为双倍或三倍字宽的保护组。这些双倍或三倍字宽保护组如果布 设在相同的节点集上，则会“隐藏”一两次。由于每个保护组正好包含两个磁盘冗余，此机制将允许群集承受双或三驱动器故障或全部 节点故障，而不会出现任何数据不可用的情况。

对于小型群集而言，最重要的是此分条方法具有高效性，磁盘效率为 M/(N+M)。例如，在具有双重故障保护的五节点群集上，如果我们 使用 N=3、M=2，则将得到 3+2 保护组，其效率为 1－2/5 或 60%。如果使用相同的 5 节点群集，但在 2 个条带上布设各保护组（此时 N 为 8，M=2），我们会得到 1-2/(8+2) 或 80% 的磁盘效率，从而保留了双重驱动器故障保护，但却牺牲了双重节点故障保护。

CHCNMS 支持多种保护方案，其中包括无处不在的 +2d:1n 方案，该方案可防止两个驱动器发生故障或一个节点发生故障。

最佳实践是对特定群集配置使用建议的保护级别。这种建议的保护级别会清楚地在 CHCNMS WebUI 存储池配置页面中标记为 “suggested”，且通常作为默认配置。对于所有当前第 6 代硬件配置，建议的保护级别为“+2d:1n”。

混合保护方案对第 6 代机箱高密度节点配置特别有用，在这些配置中，多个驱动器发生故障的概率远远超过整个节点出现故障的概率。

假设多个设备同时出现故障（机率很小），导致文件“超出相应的保护级别”，CHCNMS 将尽可能重新保护所有文件，并在群集日志上报告受影响文件上的相关错误。

CHCNMS 还提供多种镜像选项（2 倍到 8 倍），可实现指定内容的二至八次镜像。例如，默认元数据在 FEC 以上的一个级别中完成镜像。例如，如果按照 +2n 方案保护某个文件，则其关联元数据对象将被镜像 3 次。

CHCNMS 可以让管理员实时修改保护策略，同时客户端可以连接，进而读取并写入数据。 下

表总结了 CHCNMS 保护级别的完整范围，

保护级别	描述
+1n	可容忍 1 个驱动器或者 1 个节点发生故障
+2d:1n	可容忍 2 个驱动器或者 1 个节点发生故障
+2n	可容忍 2 个驱动器或者 2 个节点发生故障
+3d:1n	可容忍 3 个驱动器或者 1 个节点发生故障
+3d:1n1d	可容忍 3 个驱动器或者 1 个节点和 1 个驱动器发生故障
+3n	可容忍 3 个驱动器或者 3 个节点发生故障
+4d:1n	可容忍 4 个驱动器或者 1 个节点发生故障
+4d:2n	可容忍 4 个驱动器或者 2 个节点发生故障
+4n	可容忍 4 个节点发生故障
2 倍到 8 倍	根据配置在 2 到 8 个节点上镜像

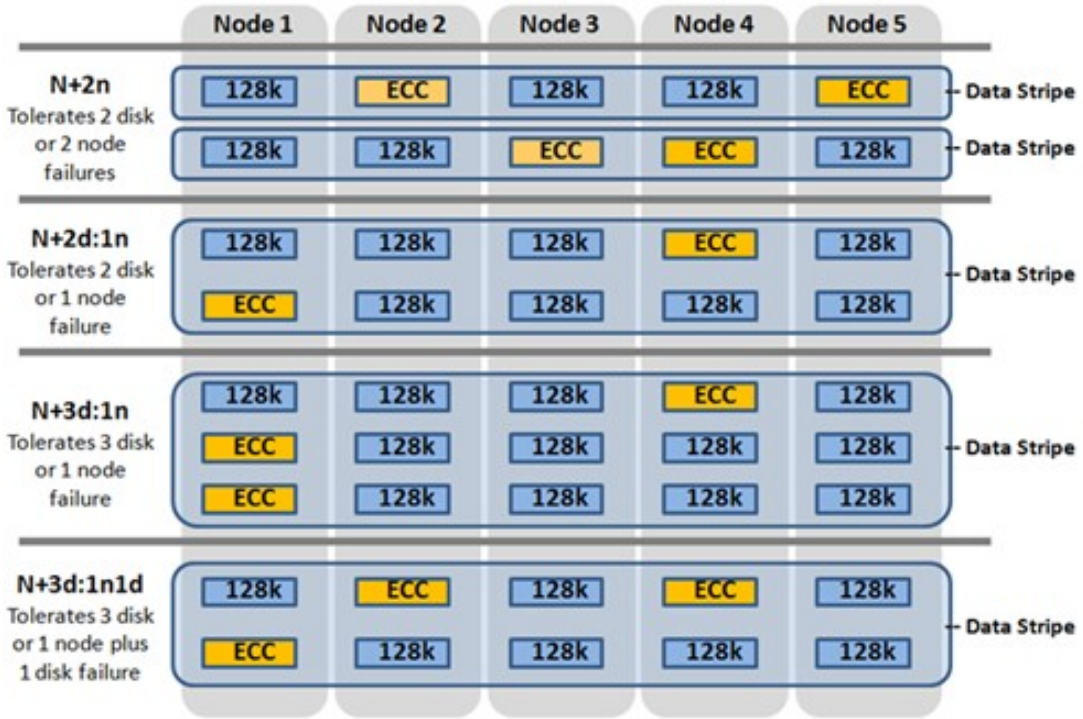

图 16：CHCNMS 混合擦除代码保护方案

CHCNMS 还为新群集安装提供保护不足的警报。如果群集保护不足，则群集事件日志系统(CELOG) 将生成警报，以警告管理员保护 不足，并建议更改此特定群集配置的相应保护级别。

自动分区

CHCNMS 中的数据分层与管理由 SmartPools 框架处理。从数据保护和布局效率的角度来看，SmartPools 可促进细分过程，即将大量高容 量同构节点细分为更多更小的“平均数据丢失时间”(MTTDL) 兼容磁盘池。例如，80 节点 H500 群集通常在 +3d:1n1d 保护级别下运 行。但是，如果将群集分区为四部分，则二十节点磁盘池将允许每个池在 +2d:1n 级别上运行，从而降低了保护开销并提高了空间利用 率，而不会增加任何管理开销。

依照存储管理精简性目标，CHCNMS 将自动计算群集并将其分区为磁盘池或“节点池”，以便针对 MTTDL 和高效空间利用率进行优化。这意味着客户无需决定保护级别（如上文提到的 80 节点群集示例）。

通过自动资源调配，每组兼容的节点硬件将自动分为磁盘池，此类池最多由四十个节点构成，每个节点上最多有六个驱动器。这些节点 池默认受到 +2d:1n 级别保护，并且可通过 SmartPools 文件池策略将多个池整合到逻辑层并进行管理。将某个节点的磁盘细分为多个受 单独保护的池之后，节点对多个磁盘故障的适应能力要比先前更为明显。

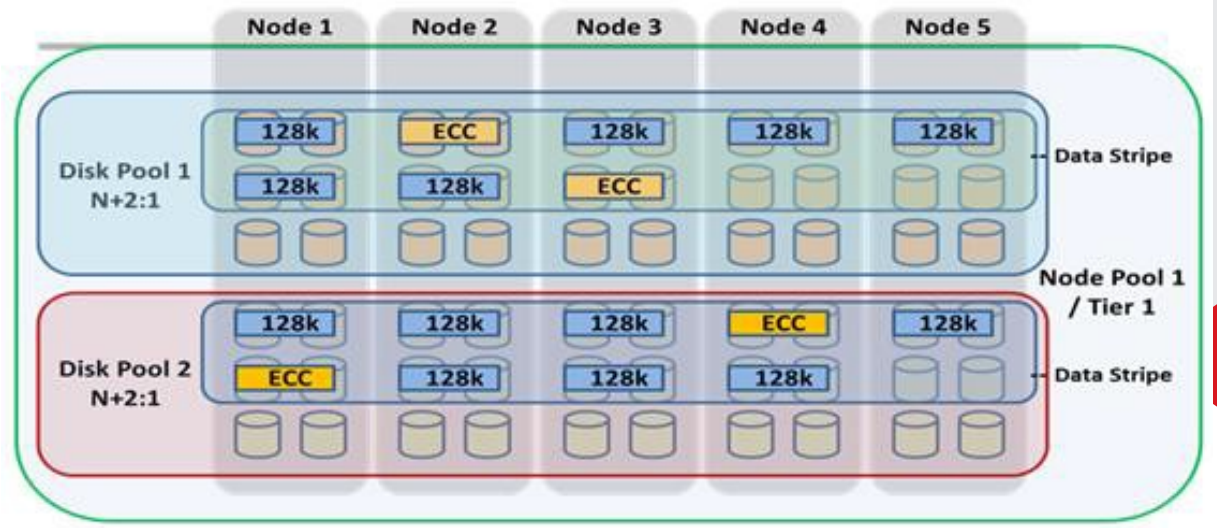

图 17：通过 SmartPools 进行自动分区

NS 第 6 代模块化硬件平台采用高密集的模块化设计，其中四个节点包含在单个 4U 机箱中。这种方式增强了磁盘池、节点池和“邻居”的概念，从而为 CHCNMS 故障域概念增加了另一个弹性级别。每个第 6 代机箱包含四个计算模块（每个节点一个），每个节点五个驱动器容器（或托架）。

图 18.第 6 代平台机箱前视图显示驱动器托架

每个托架是一个托盘，可滑入到机箱前部，并包含三到六个驱动器，具体取决于特定机箱的配置。磁盘池是存储池层次结构中的最小单元。CHCNMS 资源调配的前提是将类似节点的驱动器划分为集合或磁盘池，每个池代表一个单独的故障域。这些磁盘池默认受 +2d:1n 级 别的保护（或者能够承受两个驱动器或一整个节点故障）。

磁盘池分布在每个第 6 代节点的所有 5 个托架上。例如，一个节点（每个托架有 3 个驱动器）具有以下磁盘池配置：

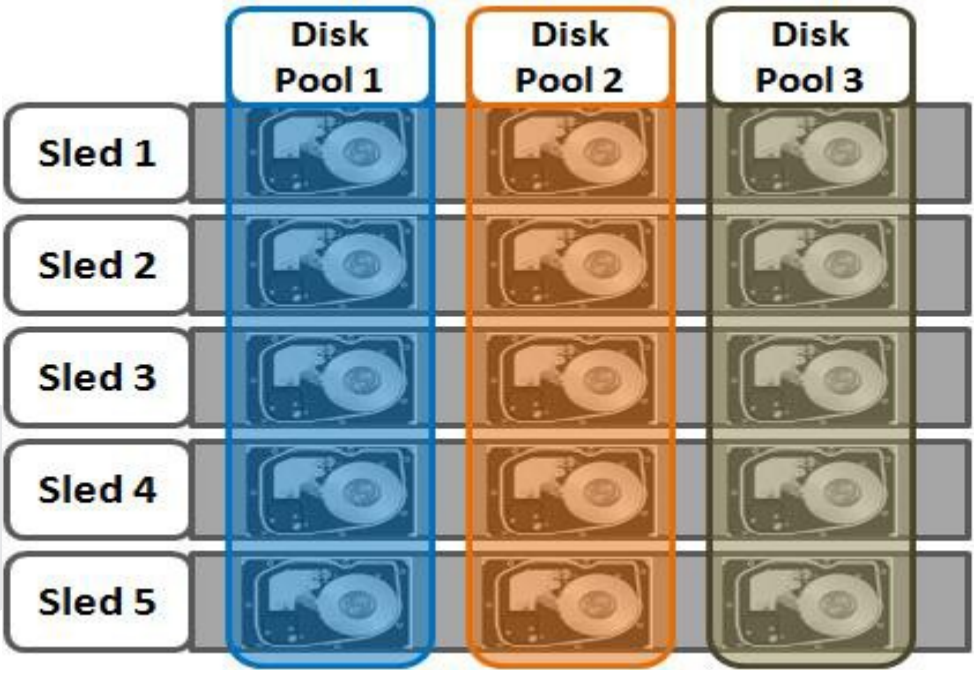

图 19.CHCNMS 磁盘池

节点池是多个磁盘池组，分布在类似存储节点（兼容性机箱）上。如下面的图 20 所示。多组不同节点类型可以在单个异构群集中协同工作。例如：一个用于 I/Ops 密集型应用程序的 F 系列节点的节点池，一个主要用于高并行和串行工作负载的 H 系列节点的节点池，以及一个主要用于近线和/或深度归档工作负载的 A 系列节点的节点池。

这样一来，CHCNMS 就可以呈现一个存储资源池，其中包含多个驱动器介质类型（SSD、高速 SAS、大容量 SATA 等），从而提供一系列不同的性能、保护和容量特性。这种异构存储池反过来可以通过一个统一的管理点来支持多种应用程序和工作负载需求。 它还有利于 融合旧的和新的硬件，轻松实现多代产品的投资保护和无缝硬件更新。

每个节点池仅包含来自相同类型存储节点的磁盘池，一个磁盘池只属于一个节点池。例如，具有 1.6 TB SSD 驱动器的 F 系列节点将位于一个节点池中，而具有 10 TB SATA 驱动器的 A 系列节点将位于另一个池中。如今，对于第 6 代硬件（如 NS H870），每 个节点池至少需要 4 个节点（一个机箱），对于自包含节点（如 NS F690），每个池至少需要 3 个节点。

CHCNMS“邻居”是节点池中的故障域，其目的通常是提高可靠性，并防止由于意外拆卸驱动

器托架而导致数据不可用。对于自包含节点（如 NS F620），CHCNMS 每个节点池的理想大小为 20 个节点，最大大小为 39 个节点。在添加第 40 个节点时，这些节点将拆分为两个邻居，每个邻居 20 个节点。

在第 6 代平台中，邻居的理想大小从 20 个节点变为 10 个节点。这可以防止同时发生节点对日志故障和整个机箱故障。

合作伙伴节点是其日志已镜像的节点。使用第 6 代平台，节点的日志存储在 SSD 上（而不是像以前的平台那样每个节点都将其日志存储在 NVRAM 中），并且每个日志在另一个节点上都有一个镜像副本。包含镜像日志的节点称为合作伙伴节点。从对日志的更改中获得了几个可靠性方面的好处。例如，SSD 比 NVRAM 更持久和可靠，后者需要充电电池才能保留状态。此外，对于镜像日志，必须在两个日志驱动器均“停止工作”后，才能认为日志丢失。因此，除非两个镜像日志驱动器均出现故障，否则两个合作伙伴节点均能正常工作。

通过合作伙伴节点保护，在可能的情况下，节点将放置在不同的邻居中，因此位于不同的故障域。一旦群集达到五个完整的机箱（20 个节点），当进行第一次邻居拆分后，CHCNMS 将合作伙伴节点放置在不同的邻居中时，就可以对合作伙伴节点进行保护：

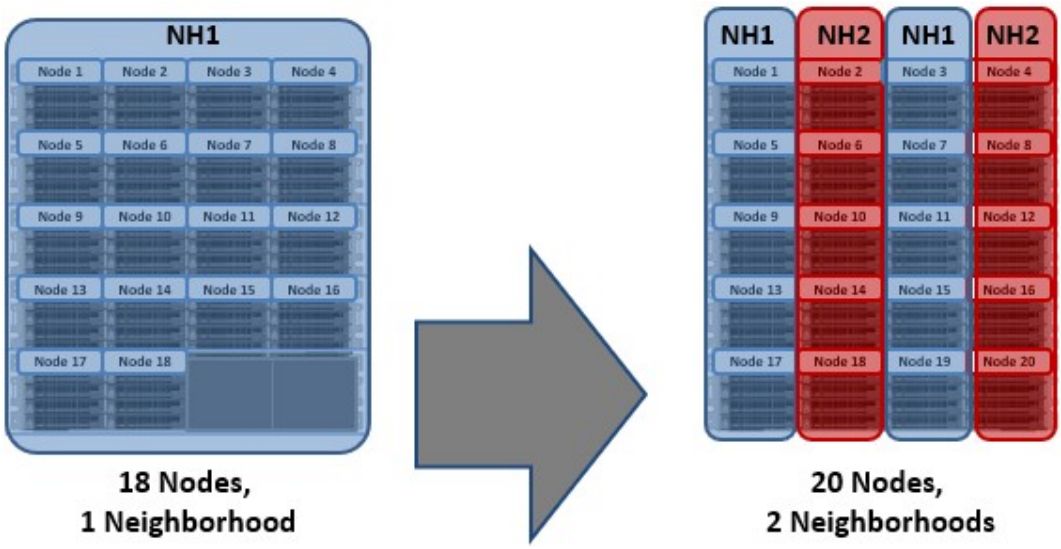

图 20.拆分为 2 个邻居，每个邻居 20 个节点

合作伙伴节点保护提高了可靠性，因为如果两个节点发生故障，它们将位于不同的故障域中，所以它们的故障域只会损失一个节点。 通过机箱保护，在可能的情况下，将一个机箱中的四个节点分别放置在单独的邻居中。可在 40 个节点时实现机箱保护，因为在 40 个节点时的邻居拆分使机箱中的每个节点均可放置在不同的邻居中。因此，当 38 节点的第 6 代群集扩展至 40 个节点时，两个现有的邻居将拆分为 4 个 10 节点邻居： 机箱保护可确保在整个机箱发生故障时，每个故障域只会损失一个节点。

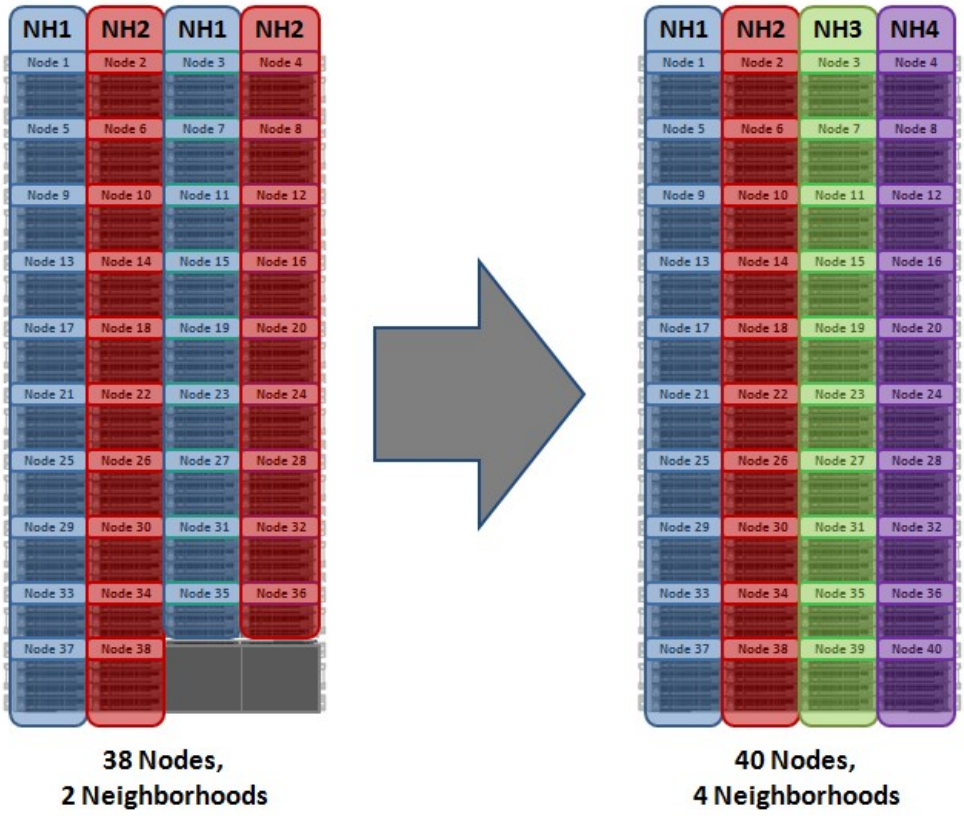

图 21.CHCNMS 邻居 — 四个邻居拆分

包含 4 个邻居的 40 节点或更大群集的默认保护级别为 +2d:1n，每个邻居可以承受一个节点故障。这样可防止群集发生一次第 6 代机箱故障。

总之，第 6 代平台群集的可靠性至少比具有类似容量的上一代群集高一个数量级，这是以下增强功能的直接结果：

- 镜像日志

- 小型邻居
- 镜像启动驱动器

兼容性

根据节点的兼容性，可将某些类似但不完全相同的节点类型调配到现有节点池。CHCNMS 要求节点池必须至少包含 3 个节点。

由于体系结构差异很明显，第 6 代平台、前几代硬件或 NS 节点之间不存在节点兼容性。

CHCNMS 还包含 SSD 兼容性选项，从而允许将具有不同容量 SSD 的节点调配给单个节点池。

在 CHCNMS WebUI SmartPools 兼容性列表中创建和描述 SSD 兼容性，且也将其显示在“Tiers & Node Pools”列表中。

当创建此 SSD 兼容性时，CHCNMS 自动检查要合并的这两个池是否有相同数量的 SSD 层、请求的保护以及 L3 高速缓存设置。如 果这些设置不同，则 CHCNMS WebUI 将提示整合并协调这些设置。

支持的协议

具有充足凭据和权限的客户端可以通过任一种与群集通信的标准支持方法来创建、修改并读取数据：

- NFS (Network File System)
- SMB/CIFS (Server Message Block/Common Internet File System)
- FTP（文件传输协议）
- HTTP (Hypertext Transfer Protocol)
- HDFS（Hadoop 分布式文件系统）
- REST API（表述性状态转移应用程序编程接口）
- S3（对象存储 API）

对于 NFS 协议，CHCNMS 支持 NFSv3 和 NFSv4，以及 CHCNMS 9.3 中的 NFSv4.1。此外，CHCNMS 9.2 及更高版本还包括对 NFSv3overRDMA 的支持。

在 Microsoft Windows 端，支持 SMB 协议，最高版本为 3。作为 SMB3 方言的一部分，CHCNMS 支持以下功能：

- SMB3 多路径
- SMB3 连续可用性与 Witness
- SMB3 加密

可以以每个共享、区域或群集为基础配置 SMB3 加密。只有支持 SMB3 加密的操作系统才可与加密共享配合使用。如果群集配置为允许 非加密连接，则这些操作系统还可与未加密的共享配合使用。只有在群集配置为允许非加密连接时，其他操作系统才可访问未加密的共享。

群集中所有数据的文件系统根为 /ifs（CHCNMS 文件系统）。这通过 SMB 协议显示为“ifs”共享 (\\<cluster_name\ifs)，通过 NFS 协议显 示为“/ifs”导出 (<cluster_name>:/ifs)。

数据在所有协议之间通用，通过一种访问协议完成的文件更改均可通过其他所有协议即时查看。

CHCNMS 通过前端以太网、SmartConnect 以及完整的存储协议和管理工具阵列为 IPv4 和 IPv6 环境提供全面的支持。 此外, CHCNMS CloudPools 还支持以下云提供商的存储 API，从而允许将文件存根到多个存储目标，包括：

- Amazon Web Services S3
- Microsoft Azure
- Google Cloud Service
- 阿里云
- CHCNMS RAN（RESTful 命名空间访问）

无中断操作 - 协议支持

CHCNMS 通过支持 Linux 和 UNIX 客户端的动态 NFSv3 与 NFSv4 故障切换和回切以及 Windows 客户端的 SMB3 连续可用性，对数据可 用性作出贡献。这可确保在发生节点故障或执行预防性维护时，所有运行中的读写操作都转由群集中的另一节点处理，以便完成操作而不中断任何用户或应用程序。

故障切换期间，客户被均匀重新分布到群集中的所有剩余节点，确保对性能产生的影响极小。如果某个节点出于任何原因发生问题，包 括故障，该节点上的虚拟 IP 地址将无缝迁移到群集中的另一个节点。

当离线节点恢复为在线状态时，SmartConnect 将自动重新平衡整个群集的 NFS 和 SMB3 客户端，以确保更大限度地提高存储和性能利 用率。对于定期系统维护和软件更新，本功能允许每节点滚动升级，从而在维护窗口的整个持续时间内提供完全可用性。

文件筛选

可以跨 NFS 和 SMB 客户端使用 CHCNMS 文件过滤功能，以允许或禁止对导出、共享或访问区域执行写入操作。此功能可防止某些类型 的文件扩展被阻止，因为这些文件可能会导致安全问题、生产力中断、吞吐量问题或存储混乱。可通过排除列表（阻止显式文件扩展） 或包含列表（显式允许只写入特定文件类型）进行配置。

重复数据消除 – SmartDedupe

SmartDedupe软件可以在所有NAS节点上支持文件级别数据重删功能，通过减少托管组织数据所需的物理存储量，更大限度提高群集的存 储效率。通过扫描磁盘数据中的相同数据块，然后 消除任何重复项目来实现高效率。此方法通常称为“后处理”或“异步重复数据消除”。

在发现重复数据块之后，SmartDedupe 会将这些数据块的一个副本移到称为“卷影存储”的一组特殊文件中。在此过程中，将从实际文 件中删除重复数据块并替换为指向卷影存储的指针。

利用后处理重复数据消除，新数据将首先存储在存储设备上，然后有一个后续过程分析此数据，以发现公用性。这意味着初始文件写入 或修改性能不受影响，因为在写入路径中无需其他计算。

SmartDedupe 体系结构

CHCNMS SmartDedupe 体系结构包括以下 5 个主要模块：

- 重复数据消除控制路径
- 重复数据消除作业
- 重复数据消除引擎
- 卷影存储
- 重复数据消除基础架构

SmartDedupe 控制路径包括 CHCNMS Web 管理界面 (WebUI)、命令行界面 (CLI) 和 RESTful 平台 API，并负责管理重复数据消除作业的 配置、计划和控制。作业本身是高度分布式后台进程，它管理跨群集中的所有节点的重复数据消除的编排。作业控制包括文件系统扫 描、检测和共享匹配的数据块，这与重复数据消除引擎配合使用。重复数据消除基础架构层是将共享数据块整合到卷影存储中的内核模 块，卷影存储是保存物理数据块和对共享数据块的引用（或指针）的文件系统容器。

卷影存储区

CHCNMS 卷影存储是允许以可共享方式存储数据的文件系统容器。因此，CHCNMS 上的文件可以包含物理数据和指向卷影存储中共享数据 块的指针或引用。

卷影存储类似于常规文件，但一般不包含通常与常规文件信息节点关联的所有元数据。特别是，明确不保留基于时间的属性（创建时间、修改时间等）。每个卷影存储区最多可包含 256 个数据块，每个数据块能够由 32,000 个文件引用。如果超过了这个 32K 引用限 制，将创建新的卷影存储。此外，卷影存储不引用其他卷影存储。不允许卷影存储的快照，因为卷影存储没有硬链接。

除了重复数据消除之外，卷影存储还用于 CHCNMS 文件克隆和小型文件存储效率 (SFSE)。

小型文件存储效率

卷影存储区的另一个主要使用者是 CHCNMS 小型文件存储效率。这项功能通过减少存储通常包含归档数据集的小型文件（如在医疗保健 PACS 工作流中找到的对象）所需的物理存储量，更大限度提高群集的空间利用率。

通过扫描磁盘上的数据以发现受完整拷贝镜像保护的小型文件并将其打包到卷影存储区中，实现高效率。然后，这些卷影存储区采用奇 偶校验保护而不是镜像，通常提供 80%甚至更高的存储效率。

小型文件存储效率可牺牲较小的读取延迟性能，以提高存储利用率。归档文件明显保持可写状态，但是当包含卷影引用的容器文件被删除、截断或覆盖时，它可能会在卷影存储区中留下未引用的数据块。这些数据块稍后将被释放，并可能导致可降低存储效率的漏洞。

实际的效率损失取决于卷影存储使用的保护级别布局。较小的保护组大小更容易受到影响，容器文件也是如此，因为容器中的所有数据块最多有一个引用文件，而且压缩的大小（文件大小）也很小。

提供了碎片整理程序来减少文件片段，从而减少了覆盖和删除等操作。这款卷影存储区碎片整理程序已集成到 ShadowStoreDelete 作业 中。碎片整理流程的运作方式为：将每个容器化文件划分为多个逻辑区块（每个约 32 MB），并评估每个区块的片段。

如果包含片段的区块的存储效率低于目标值，则通过将数据转移到另一个位置来处理该区块。默认目标效率是卷影存储区使用的保护级 别所提供的最大存储效率的 90%。在存储效率下降到低于此阈值之前，较大的保护组大小可以容忍较高的碎片级别。

线内数据减少

NS F690、NS F681、NS F660 和 NS F620 全闪存节点、NS H875/NS H870 和 NS H855 混合机箱以及 NS A130/NS A1300 归档平台提供 CHCNMS 线内数据缩减功能。

CHCNMS 体系结构包含以下主要组件：

- 数据减少平台

- 压缩引擎和区块映射
- 零数据块删除阶段
- 重复数据消除内存中索引和卷影存储基础架构
- 数据减少警报和报告框架
- 数据减少控制路径 线内数据减少写入路径包括三个主要阶段：
- 零数据块删除
- 线内重复数据消除
- 线内压缩

如果在群集上启用了线内压缩和重复数据消除，则先执行零数据块删除，接下来执行重复数据消除，然后进行压缩。此顺序使每个阶段 都可以减少每个后续阶段的工作范围。

图 24：线内数据缩减工作流

NS F681 节点包括硬件压缩分载功能，NS F681 机箱中的每个节点包含一个 Mellanox Innova-2 Flex 适配器。这意味着，Mellanox 适配器以极小的延迟透明地执行压缩和解压缩，从而无需消耗节点昂贵的 CPU 和内存资源。

CHCNMS 硬件压缩引擎使用 zlib，以及 NS F690、NS F681、NS F660、NS F620、NS H870/NS H875、NS H855 和 NS A130/NS A1300 节点的 igzip 软件实施。软件压缩还可用作发生压缩硬件故障时的后备方案，在混合群集中，可用于没有硬件压缩功能的非 NS F681 节点，以及用作压缩硬件发生故障时的后备方案。CHCNMS 使用的压缩区块大小为 128 KB，每个区块包含 16 个 8 KB 数据块。这是一种最佳方法，因为它的大小与 CHCNMS 用于其数据保护条带单元的大小相同，从而通过避免额外区块打包的开销，提供简易性和效率。

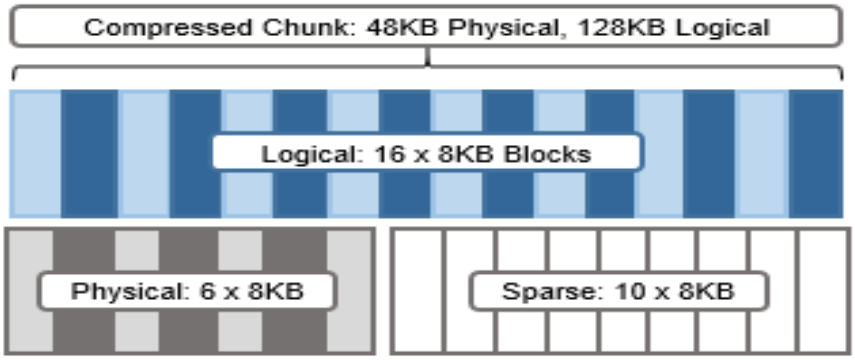

图 25：压缩区块和 CHCNMS 透明覆盖:

请仔细观察上图。压缩后，此区块的大小从 16 个缩小到 6 个 8 KB 数据块。这表示，此区块现在的大小实际为 48 KB。CHCNMS 为物理 属性提供透明的逻辑覆盖。这个覆盖描述了备份数据是否经过压缩，以及区块中的哪些块是物理块或稀疏块，这样文件系统使用者就不 会受到压缩的影响。因此，无论实际物理大小如何，压缩区块的大小在逻辑上都表示为 128 KB。

要进行压缩，必须至少节省 8 KB（一个数据块）的效率，否则该区块或文件将经历传递，并保持其原始的未压缩状态。例如，一个 16 KB 的文件节省 8 KB（一个数据块），将会得到压缩。文件压缩后，就会受到 FEC 保护。

压缩区块绝不会跨越节点池。这样便无需解压缩或重新压缩数据来更改保护级别、执行恢复的写入操作或以其他方式转移保护组边界。

动态扩展/按需扩展

性能和容量

与需要额外性能或容量时必须“纵向扩展”的传统存储系统相较而言，CHCNMS 可以让存储系统“横向扩展”，这不仅可实现性能的线性 提升，还可无缝地将现有文件系统或卷增加到 PB 级容量。

向群集添加容量和性能功能比起其他存储系统来容易得多，存储管理员只需执行简单的三步即可：在机架中添加另一个节点；将该节点连接到后端网络；指示群集添加其他节点。新节点将提供额外的

容量和性能，因为每个节点都包含 CPU、内存、缓存、网络、NVRAM 以及 I/O 控制路径。

CHCNMS 的 AutoBalance 功能将自动、连贯地跨后端网络移动数据，从而让驻留在群集上的现有数据移动到这个新存储节点中。该自动重 新平衡不但确保了新节点无法成为新数据的热点，而且可确保现有数据可以享有更强大存储系统的优势。CHCNMS 的 AutoBalance 功能也 对最终用户完全透明，并且可调节，以尽量减小对高性能工作负载的影响。只有此功能允许 CHCNMS 进行透明的动态扩展，范围为数 TB 到数 PB 之间，期间不会延长管理员的管理时间，也不会增加存储系统的复杂性。

大型存储系统应当具备不同工作流的所需性能，不论它们是顺序工作流、并发工作流还是随机工作流。应用程序之间以及单个应用程序 内存在不同的工作流。CHCNMS 通过智能软件同时满足所有这些需求。更重要的是，利用 CHCNMS，吞吐量和 IOPS 可随单个系统中节点 数的增加线性扩展。由于平衡数据分发、自动重新平衡和分布式处理，CHCNMS 可随着系统扩展利用增设的 CPU、网络端口以及内存。

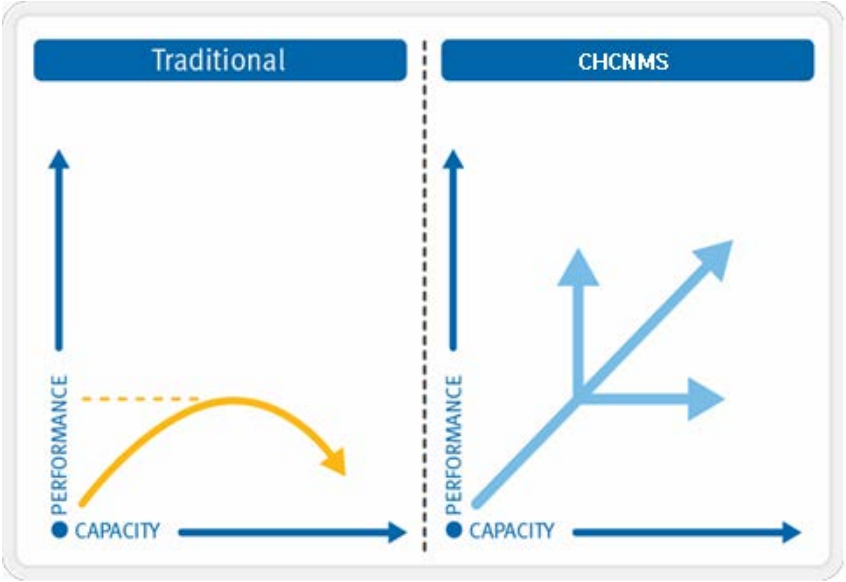

图 26：CHCNMS 线性可扩展性

界面

管理员可以使用多个界面来管理其环境中的存储群集：

- Web 管理用户界面（“WebUI”）
- 通过 SSH 网络访问或 RS232 串行连接的命令行界面
- 节点上自带的 LCD 面板，提供简单的添加/删除功能
- RESTful 平台 API，提供用于群集配置与管理的编程控制和自动化

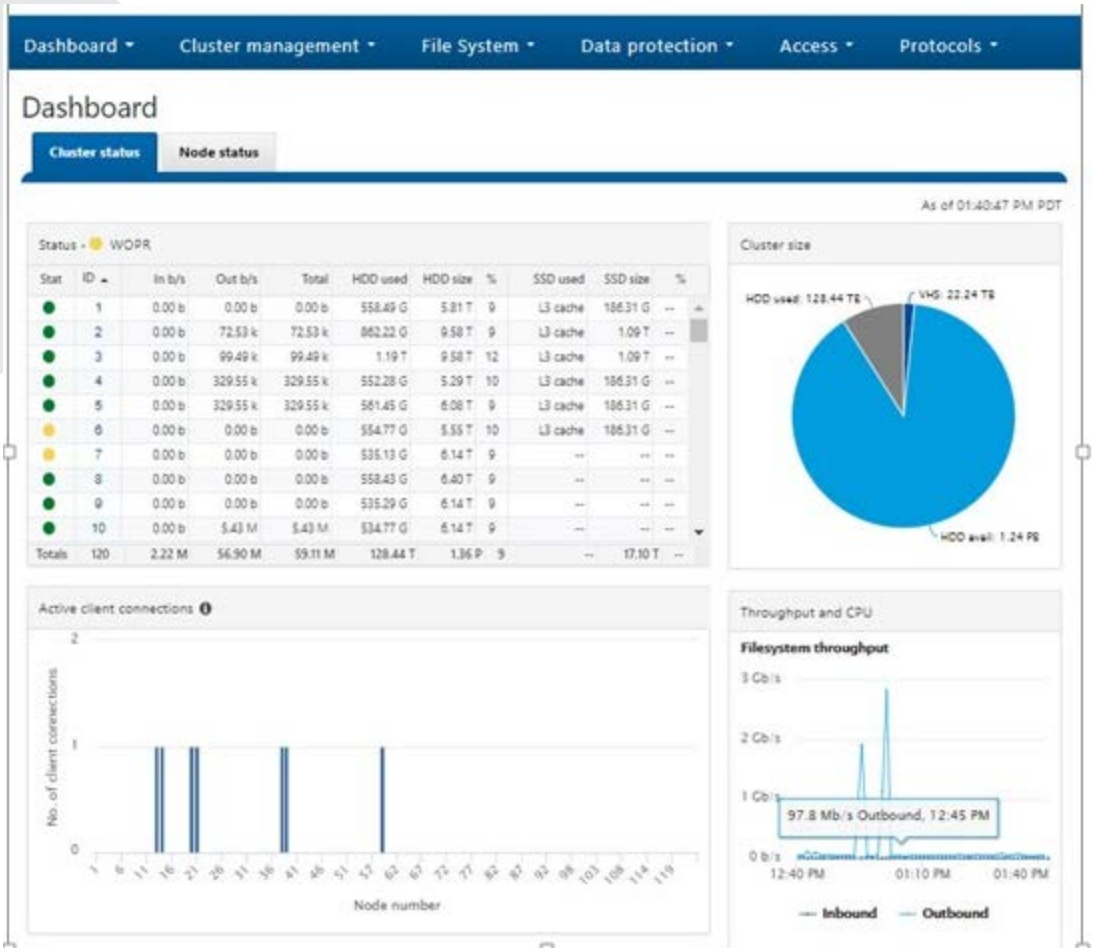

图 27：CHCNMS Web 用户界面

身份认证与访问控制

在允许访问、修改文件之前，身份认证服务通过验证用户凭据来提供安全保障。CHCNMS 支持四种认证用户的方法：

- Active Directory (AD)
- LDAP（轻量级目录访问协议）
- NIS (Network Information Service)

- 本地用户与组

CHCNMS 支持多种身份认证类型的使用。但是，在群集上启用多种身份认证方法之前，建议您充分了解各种身份认证类型之间的交互。有关如何正确配置多个身份认证模式的详细信息，请参阅产品文档。

Active Directory

Active Directory（Microsoft 在 LDAP 的实现）是一项可存储网络资源信息的目录服务。虽然 Active Directory 具有多项功能，但连接群 集和域前仍要进行用户和组的身份认证。您可以通过 Web 管理界面或命令行界面来配置并管理某个群集的 Active Directory 设置；但我们建议您尽量使用 Web 管理。群集内各个节点共享一个 Active Directory 主机帐号，这样一来可以轻松管理并控制。

LDAP

轻量级目录访问协议 (LDAP) 是一种定义、查询和修改服务与资源的联网协议。LDAP 具有开放性使用目录服务的主要优势，还能应用 于多个平台。群集存储系统可以使用 LDAP 对用户和组进行身份验证，以便授予它们对群集的访问权限。

NIS

Sun Microsystems 设计的 Network Information Service (NIS) 是一种 CHCNMS 可以使用的目录服务协议，用于在访问群集时对用户和组 进行身份验证。NIS 有时称作黄页 (YP)，它不同于 NIS+，不受 CHCNMS 支持。

本地用户

CHCNMS 支持本地用户和组身份验证。通过 WebUI 界面，您可以在群集上直接创建本地用户和组帐户。未使用目录服务（Active Directory、LDAP 或 NIS）时，或者特定用户或应用程序需要访问群集时，本地身份认证可能会非常有用。

访问分区

访问分区可用于逻辑分区群集访问权限，并向独立装置分配资源，从而提供一种共享的租户或多租户环境。为方便操作，访问分区共关 联三个核心外部访问组件：

- 群集网络配置
- 文件协议访问
- 验证

同样地，SmartConnect 分区与一组 SMB 共享、NFS 导出、HDFS 机架以及一个或多个身份验证提供程序进行关联，以实现每个分区的 访问控制。这样可提供集中管理的单个文件系统的优势，该系统可以为多个租户调配资源和提供保护。如果中央 IT 部门要向多个单独的 业务部门提供服务，访问分区对企业环境尤为有用。再比如，服务器整合计划期间；合并多个加入不受信任的独立 Active Directory 林的 Windows 文件服务器时。

利用访问分区时，内置系统访问分区默认包括了每个受支持的身份认证提供程序的实例、所有可用的 SMB 共享及所有可用的 NFS 导出。这些身份认证提供程序可能包括 Microsoft Active Directory、LDAP、NIS 以及本地用户或组数据库的多个实例。

基于角色的管理

基于角色的管理是一种基于群集管理角色的访问控制系统 (RBAC)，可以将 root 用户和管理员用户的权力分割成更精细的权限，并允许 将这些权限分配给特定角色。然后，可以向其他非特权用户授予这些角色。例如，向数据中心操作人员授予整个群集的只读权限，允许 全方位监视访问，但不进行任何配置更改。CHCNMS 提供一组内置角色，包括审计、系统与安全管理员，以及为每个访问区域或跨群集创 建自定义角色的功能。基于角色的管理与 CHCNMS 命令行界面、WebUI 和平台 API 集成。

SyncIQ 数据复制概述

利用 CHCNMS，您可通过 SyncIQ 软件模块在 NS 群集之间复制数据。在两个 NS 群集之间复制数据之前，您必须同时在这二者上激活 SyncIQ 许可证。

您可复制目录级别的数据，同时视情况从复制内容中排除特定文件及子目录。SyncIQ 会创建并引用快照来复制源目录的一致时间点映像。元数据（如访问控制列表 (ACL)和备用数据流 (ADS)）随数据一起复制。

SyncIQ 支持特定类型的文件、特定大小的文件复制等；数据复制能够灵活设定时间表和策略；支持一对一、一对多、多对多复制方式；支持单向\双向复制。

SyncIQ 使您可以在其他 NS 群集上维护数据的一致副本以及控制数据复制的频率。例如，您可以将 SyncIQ 配置为在每天晚上 10 点将数据从主群集备份到辅助群集一次。根据数据集的大小，第一次复制操作可能需要相当长的时间。但是此之后，复制操作会更快地完成。

SyncIQ 还提供自动故障切换和回切功能，因此如果主群集变得不可用，您可在辅助 NS 群集上继续执行操作。

复制策略和作业

数据复制将根据复制策略和复制作业进行协调。复制策略指定要复制的数据、数据复制的目标位置及数据复制频率。复制作业是在 NS 群集之间复制数据的操作。SyncIQ 将根据复制策略生成复制作业。

一个复制策略指定两个群集：源和目标。复制策略所在的群集是源群集。数据将复制到的群集是目标群集。复制策略启动时，SyncIQ 将为此策略生成一个复制作业。复制作业运行时，源群集上目录树中的文件将复制到目标群集上的目录树；这些目录树称为源目录和目标目录。

复制策略创建的第一个复制作业完成后，目标目录及其中的所有文件将设置为只读状态，并只能由同一复制策略中的其他复制作业修改。我们建议您不要在群集上创建超过 1,000 个策略。

您可以创建两种类型的复制策略：同步策略和拷贝策略。同步策略用于维护目标群集上源目录的精确复制副本。如果从源目录中删除一个文件或子目录，则策略再次运行时，该文件或目录将从目标群集中删除。

您可使用同步策略在源和目标群集之间执行数据故障切换和回切。源群集变得不可用时，您可以故障切换目标群集上的数据并且将这些数据提供给客户端。源群集再次变得可用时，您可将数据回切至源群集。

拷贝策略维护源群集上存储的文件的最新版本。然而，在源群集上删除的文件不会从目标群集上删除。拷贝策略不支持回切。拷贝策略最常用于归档目的。

您可使用拷贝策略从源群集删除文件，而不会丢失目标群集上的这些文件。在源群集上删除文件可提高源群集的性能，同时在目标群集上维护这些已删除的文件。例如，如果源群集用于生产，而目标群集仅用于归档，这样会很有帮助。

为复制策略创建作业后，SyncIQ 必须等到作业完成后才能为该策略创建另一个作业。群集上适合时候都可存在任意数量的复制作业；但是，在源群集上最多同时只能运行 50 个复制作业。如果群集上的复制作业数超过 50，将运行前 50 个作业，而其他作业排队等待运行。

目标群集可并发支持的复制作业数量没有限制。但是，由于更多复制作业需要更多的群集资源，因此随着添加更多并发作业，复制会降低速度。

运行复制作业时，系统将在源和目标群集上生成工作进程。源群集上的工作进程用于读取和发送数据，而目标群集上的工作进程用于接收和写入数据。系统为每个复制作业的每个节点生成的工作进程数不超过 8。例如，在 5 节点群集中，系统将为复制作业创建的工作进程数不能超过 40 个。

在一个复制作业中，您可复制任意数量的文件和目录。通过限制数据同步可占用的群集资源和网络带宽，您可以防止大型复制作业造成系统负载过重。由于群集中的每个节点均可收发数据，因此群集越大，数据复制速度越快。

自动复制策略

您随时都可以手动启动复制策略，但也可将复制策略配置为根据源目录修改或计划自动启动。您可将复制策略配置为根据计划运行，以便于您控制复制执行时间。您还可以配置策略来复制以目录快照形式捕获的数据。您还可将复制策略配置为在 SyncIQ 检测到源目录修改时启动，以便 SyncIQ 在目标群集上维护最新的数据版本。

在以下条件下，计划策略会很有帮助：

- 您想要在用户活动最少的情况下复制数据
- 您可以准确预测数据修改时间

如果某个策略配置为按计划运行，您可以将策略配置为自上次运行作业以来未更改源目录内容时不运行。但是，如果更改了源目录的父目录或源目录的同级目录，然后拍摄父目录的快照，SyncIQ 将创建作业的策略，即使未更改源目录也是如此。此外，如果通过 InsightIQ 的文件系统分析(FSA)功能监视群集，FSA 作业将创建/ifs 的快照，这最有可能会导致只要运行 FSA 作业就启动复制作业。

复制目录快照中包含的数据在以下情况下会很有用：

- 您想要按计划复制数据，并且您已在通过快照计划生成源目录的快照
- 您想要保持源和目标群集上的快照相同
- 您想要将现有快照复制到目标群集

要执行此操作，必须启用目标群集上的归档快照。仅当创建策略时，才能启用此设置。如果配置某个策略来复制快照，您可以将 SyncIQ 配置为仅复制与指定命名模式匹配的快照。

完整和差异复制

如果复制策略遇到无法修复的问题（例如，目标群集上的关联断开），您可能需要重置复制策略。如果重置复制策略，SyncIQ 将在下次运行此策略时执行完全复制或差异复制。您可指定 SyncIQ 执行的复制类型。

在完全复制过程中，无论目标群集上存在什么数据，SyncIQ 都会传输源群集中的所有数据。完全复制会消耗大量网络带宽，可能需要很长时间才能完成。但是，完全复制占用的 CPU 资源比差异复制更少。

在差异复制过程中，SyncIQ 首先会检查文件是否存在于目标群集上，然后仅传输目标群集上不存在的数据。差异复制占用的网络带宽比完全复制更少；但是，前者要占用更多的 CPU 资源。只要有足够的 CPU 资源可供复制作业使用，差异复制要比完全复制快很多。

复制报告

复制作业完成后，SyncIQ 将生成一份包含详细作业信息的复制报告，其中包括作业的运行时长、传输的数据量以及发生的错误。

如果复制报告被中断，SyncIQ 可能创建一份关于截至目前的作业进度的子报告。如果随后重新启动此作业，SyncIQ 将创建一份关于截至作业完成或再次中断的作业进度的子报告。每次作业中断，SyncIQ 都会创建一份子报告，直到作业成功完成。如果为某作业创建了多份子报告，SyncIQ 会将子报告中的信息组合为一份报告。

SyncIQ 会定期删除复制报告。您可以指定 SyncIQ 保留的最大复制报告数以及 SyncIQ 保留复制报告的时长。如果群集上超出最大复制报告数，则每当创建新报告时，SyncIQ 都会删除最旧的报告。

CHCNMS 审计

CHCNMS 提供的功能可在群集上审核系统配置以及 NFS、SMB 和 HDFS 协议活动。这样，组织就可以满足其可能绑定的各种数据管理和 法规遵从性要求。

所有审计数据均受保护且存储在群集文件系统中，并按审计主题进行组织。在这里，可以通过 Common Event Enabler (CEE) 框架将审计数据导出到第三方应用程序，如 Varonis DatAdvantage 和 Symantec Data Insight。可以为每个访问分区启用 CHCNMS 协议审 计，从而允许在整个群集中进行精细控制。

群集可以采用并行负载均衡配置为每个节点中最多 5 台 CEE 服务器写入审计事件。这使得 CHCNMS 能够提供端到端企业级审计解决方案。

软件升级

升级到 CHCNMS 的最新版之后，即可使用所有新功能、修复及功能。群集可以使用两种方法进行升级：同步或滚动升级

同步升级

同步升级将安装新操作系统，并且可同时重新启动群集内的所有节点。在同步升级过程中，当重新启动节点时，该升级需要 2 分钟以内 的临时服务中断。

滚动升级

滚动升级单独完成升级，并依次重新启动群集内的各个节点。滚动升级过程中，群集保持联机状态，并向客户端继续提供数据服务，期 间无服务中断。在 CHCNMS 8.0 之前，滚动升级只能在 CHCNMS 代码版本系列中完成，无法在 CHCNMS 主代码版本修订版之间完成。从 CHCNMS 8.0 开始，每个新版本都将从以前的版本进行滚动升级。

无中断升级

无中断升级（NDU）允许群集管理员升级存储操作系统，同时他们的终端用户可继续访问数据而不会出现错误或中断。更新群集上的操作 系统只是一个简单的滚动升级问题。在此过程中，每次将一个节点升级到新代码，并且附加到该节点上的活动 NFS 和 SMB3 客户端将 自动迁移到群集中的其他节点。此外还允许部分升级，从而可以升级群集节点的子集。在升级过程中，可能还会增加节点的子集。升级 可以暂停和恢复，使客户能够跨多个较小的维护时段进行升级。此外，CHCNMS8.2.2及更高版本还提供并行升级，群集可以一次升级整

个邻居或容错域，从而大幅缩短大型群集升级的持续时间。CHCNMS9.2 及更高版本结合了操作系统和固件升级，允许它们同时进行，从 而大大减少了升级的影响和持续时间。9.2 及更高版本还包括基于排出的升级，在所有 SMB 客户端都与节点断开连接之前，节点将无法 重新启动或重新启动协议服务。

回滚功能

CHCNMS 支持升级回滚，从而能够将具有未提交升级的群集恢复到以前的 CHCNMS 版本。

自动固件更新

在无中断固件更新过程中，CHCNMS 支持的群集支持新驱动器和更换驱动器的自动驱动器固件更新。固件更新通过驱动器支持包进行交付，此类包简化了整个群集中现有驱动器和新驱动器的管理。这样可确保驱动器固件是最新固件，并降低因已知驱动器问题而发生故障的可能性。因此，自动驱动器固件更新是 CHCNMS 高可用性和无中断运营战略的重要组成部分。驱动器和节点固件可以在滚动升级中应用，也可以通过完全群集重新启动来应用。

在 CHCNMS 8.2 之前，节点固件更新必须一次安装一个节点，这项操作非常耗时，尤其是在大型群集中。现在，可以提供要同时更新的节 点列表来跨群集编排节点固件更新。升级助手工具可用于选择能同时更新的所需节点组合，以及不应一起更新的显式节点列表（例如，节点对中的节点）。

执行升级

在升级过程中，CHCNMS 会自动运行安装前验证检查。这可验证当前安装的 CHCNMS 配置是否与 CHCNMS 的升级版本相兼容。如果发现不 支持的配置，则升级会立即停止，同时显示故障排除说明。在开始升级之前，主动运行预安装升级检查有助于避免因不兼容配置引起的任何中断。

CHCNMS 数据保护和管理软件

CHCNMS 提供全面的数据保护与管理软件产品组合，以满足您的需要：

软件模块	功能	描述
CloudIQ	群集运行状况监控	实施智能和预测性分析来主动监控群集的运行状况。
InsightIQ	性能管理	利用创新的性能监控和报告工具，更大限度地提升群集的性能
DataIQ	数据分析和管理	无论数据驻留在何处（在文件和对象存储中、在本地或在云中），都能在几秒内找到、访问和管理数据。通过单一控制台，可跨异构存储系统获得全面视图，从而有效地细分在孤岛中捕获的数据。
SmartPools	资源管理	实施高效、自动化分层存储策略，优化存储性能和成本。
SmartQuotas	数据管理	分别在群集、目录、子目录、用户以及组等级别上分配并管理配额，这些配额可以 无缝将存储分区并精简调配到易于管理的片段中。
SmartConnect	数据访问	启动客户端连接负载平衡和存储节点上客户端连接的动态 NFS 故障切换和回切，从而优化群集资源的使用
SnapshotIQ	数据保护	通过安全的近即时快照，高效可靠地保护数据，并且几乎不产生任何性能开销。通过近即时的按需快照修复功能，加快关键数据的恢复。使用 CHCNMS 可写快照创建只 读快照的空间和时间高效且可修改的副本。
SyncIQ	数据复制	复制大型关键任务数据集并将其异步分发到多个站点的多个共享存储系统，从而实现可靠的灾难恢复功能。一键式故障切换和回切的简易性可提高关键任务数据的可 用性。

SmartLock	数据保留	通过我们基于软件的一写多读 (WORM) 方法，防止关键数据意外、过早或被恶意更 改或删除，同时满足严格合规性要求和管理要求，如 SEC 17a-4 要求。
SmartDedupe	重复数据消除	扫描群集中的相同数据块，然后消除重复项以减少所需的物理存储量，从而更大限 度地提高存储效率。
CloudPools	云分层	通过 CloudPools，您可以定义应将群集上的哪些数据归档到云存储。云提供商包括 Microsoft Azure、Google Cloud、Amazon S3、和原生 CHCNMS。

结论

借助 CHCNMS 操作系统提供支持的长虹 NS 横向扩展 NAS 解决方案，组织可以在单个文件系统、单个卷中通过单点管理从数 TB 扩展到数 PB。CHCNMS 提供高性能和/或高吞吐量，而且不会增加管理复杂性。

新一代数据中心的构建必须要能够实现可持续的扩展性。这些数据中心将借助自动化力量，利用硬件的商品化，确保网络结构层的充分运用，并向致力于满足不断变化需求的组织提供更高的灵活性。CHCNMS 是新一代文件系统，专为满足这些挑战而设计。CHCNMS 具有：

- 完全分布式单一文件系统
- 高性能、完全对称的群集
- 群集内所有节点上的文件分条
- 自动化软件以消除复杂性
- 动态内容平衡
- 灵活数据保护
- 高可用性
- 基于 Web 的管理和命令行管理

CHCNMS 非常适用于企业数据湖环境中基于文件的非结构化“大数据”应用程序（包括大型主目录、文件共享、归档、虚拟化以及业务分 析）。此外，它还适用于各种不同的数据密集型高性能计算环境，包括能源勘探、金融服务、Internet 与托管服务、商业智能、工程设 计、制造业、媒体与娱乐、生物信息学以及科学研究。

四川长虹佳华信息产品有限责任公司
地址：北京市丰台区南四环西路188号18区26号楼 北京长虹科技大厦九层 电话:(8610) 58292588
网址：www.changhongit.com