WHITE PAPER

ChangHong Content Platform

架构白皮书

2023 年

摘要

稳定高效、持久可靠、管理简单且横向扩展的分布式文件存储平台，提供最高级别性能、数据保护能力、海量扩展能力、更高存储效率及全面的兼容性。

目录

一、序言

数字业务被一致认为是企业未来发展的必然方向，而当今各行业的企业都在面临数字化的搅局者(Digital Disruptors)带来的挑战，比如金融行业的数字银行和互联网金融、媒体行业的互联网新媒体等。这是一种长期的趋势，企业要么拥抱数字化的产品、服务和信息并将其作为业务的核心所在，要么就是被数字化的搅局者所打败。IT 在企业的数字化转型战略中处于至关重要的位置，云、大数据、移动和社交等 IT 发展趋势也不仅仅是口号，而是迫在眉睫。

随着 IT 不断发展和数据的积累，企业发现越来越难管理和利用这些数据，特别是分散在不同应用系统的这些数据。而新的技术，互联网或者移动设备等更需要随时随地的访问这些数据。那么如何实现这些数据的集中管控、持久可靠的存储、更好的被保护、随时随地安全访问呢?

Chang Hong Content Platform (CHCP) 是一款完全由四川长虹佳华信息产品有限责任公司自主研发、生产的多用途的分布式文件存储系统，采用软硬件一体化设计，旨在支持大规模非结构化数据资产的高速、持久可靠存储和访问。他可以帮助企业实现私有的、混合的或者公有的云存储服务，而构建成本优于公有云。同时，相较于传统的文件存储等，他可以提供更好的扩展性以支持海量文件长期存放，提供更持久可靠的空间以解决传统备份的挑战，提供更完整的数据描述信息实现跨应用系统的数据检索和调阅，提供更多的数据传输接口更适合互联网和移动访问。

CHCP 的亮点

- 全球最快的共享文件系统
- 支持裸机、容器化、虚拟和云（本地、混合、公有云）环境
- 能够以聚合平台、专用存储设备或云原生形式部署
- 灵活的应用存储访问，包括 POSIX、NFS、SMB、S3 和 NVIDIA® GPUDirect®存储
- 零性能调优，同时支持各种大小的文件，同时支持混合随机和顺序输入/输出模式
- 应用级 4K I/O、高速网络上一致的亚 250 微秒延迟、无限随机 IOPs 性能–根据集群大小线性扩展

• 自动化内置混合存储，通过本地或云对象存储将命名空间从快速闪存扩展到硬盘存储

• 强大的安全功能，包括加密（静态和动态）、身份验证、密钥管理、LDAP

• 完全分布式的数据和元数据，确保存储集群中没有热点

• 大规模横向扩展支撑海量非结构化数据，单个系统可以支持 1000PB 级的存储容量和 1000 亿级的文件数量

• 与云环境全面集成，针对混合云模式或 100%公有云的提供云爆发能力

• 最持久且可靠的数据平台，提供全面的数据保护能力，无需要借助于耗时且耗资的传统备份，即可实现物理错误、逻辑错误(人为和病毒等)、比特错误和灾难等场景的数据保护和修复，可以实现 99.9999999999999% 的数据持久性

二、云时代常见的存储挑战

现代应用有各种不同的性能要求（IOPs、带宽、延迟），另外，由于应用文件格式、访问协议和数据结构的多样性，这些都会导致 IT 复杂性的增加。

存储架构师试图通过对特定应用采用多种存储架构来克服这些限制。全闪存存储区域网络（通常称为全闪存阵列 (AFA)）进行了性能优化，但它们无法满足云规模的要求，也不具备网络连接存储 (NAS) 的简单性和可共享性。此外，AFA 提供了不能跨服务器共享的块存储协议，因此不适合共享存储用例。这些临时替代方案的结果是需要重新构建基础架构，以跟上不断变化的应用需求，而这是一个代价高昂且无休止的循环。

由于可用的选项多种多样，因此，确定哪种存储解决方案最适合特定的环境或应用工作负载是一项棘手的任务。有些解决方案针对性能进行了优化，而有些则针对规模进行了优化。技术计算空间中的工作负载（如人工智能 (AI) 和机器学习 (ML)、基因组研究和金融分析）都会对超大数据集进行大文件顺序访问和小文件随机访问，因此，问题尤其突出。没有一种传统的存储设计能够满足所有这些工作负载模式的需求。临时解决方法一直是使用多个存储系统和复杂的数据管理平台。

三、设计统一的云时代存储解决方案

在设计现代存储解决方案时，一个关键的考量是要考虑到技术的不断演变和改进。真正的软件定义的存储解决方案应该能够适应这些变化，这意味着它必须能够在商用服务器硬件上运行，适应客户环境，并增加像云环境一样的灵活性、可扩展性和按需调配的性能。它还应该能够轻松地部署和扩展，而不会像传统外部存储设备那样存在采购延迟。

传统设计约束带来的局限性促使我们开发出一种全新的文件系统 CHCP，该系统在单个架构中提供了全闪存阵列的性能、横向扩展的 NAS 的简单性和云环境的可扩展性。图 1 列出了 CHCP 文件系统的价值主张。

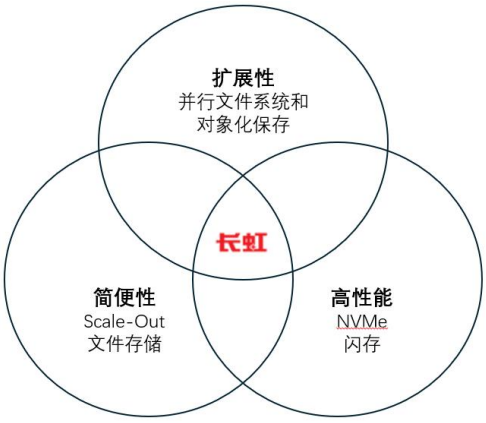


图 1：CHCP 分布式文件系统设计理念

CHCP 文件系统是一种全新的高性能存储解决方案，它采用全新设计，解决了与传统存储系统相关的问题。它在任何基于 AMD 或 Intel x86 的标准服务器硬件上运行，并配有商用 NVMe 固态硬盘 (SSD)，而无需定制的专用硬件。这种理念使您利用技术的进步，而不需要完全升级到下一代架构，包括公有云部署。

四、CHCP – 性能不打折扣的存储

CHCP 通过消除影响应用性能的瓶颈而解决前面提到的常见存储挑战。它非常适合需要低延迟、高性能和云环境扩展性的共享存储的苛刻环境。

用例包括：

- 人工智能 (AI) 和机器学习，包括 AIOps 和 MLOps
- 生命科学，包括基因组学、Cryo-EM、定量药理学 (NONMEM, PsN)
- 金融交易，包括回溯测试、时间序列分析和风险管理
- 工程 DevOps
- 电子设计和自动化 (EDA)
- 媒体渲染和视觉效果 (VFX)
- 高性能计算 (HPC)

通过以新的方式利用现有技术，并借助工程技术的创新成果增强这些技术的能力，CHCP 的软件提供了一个更强大、更简单的解决方案，而在过去，这需要多个不同的存储系统才能实现。这种软件解决方案为所有工作负载（大小文件、读写、随机、顺序和元数据密集）提供了卓越的性能。此外，由于其设计初衷是在商用服务器基础架构上运行，因此不依赖任何专用硬件。

CHCP 是一个完全分布式的并行文件系统，由高性能节点和归档节点组成，高性能节点利用 NVMe 闪存提供最高性能的文件服务，归档节点利用大容量磁盘（HDD）和对象化存储技术提供海量、持久、可靠的文件服务。CHCP 还包括智能数据分层功能，可以根据策略无缝地将数据从高性能节点迁移至归档节点（包括回迁），而不需要特殊的数据迁移软件或复杂的脚本；所有数据都驻留在一个命名空间中，便于访问和管理。直观的图形用户界面允许单个管理员快速轻松地管理上千 PB 的数据，而无需接受任何专门的存储培训。

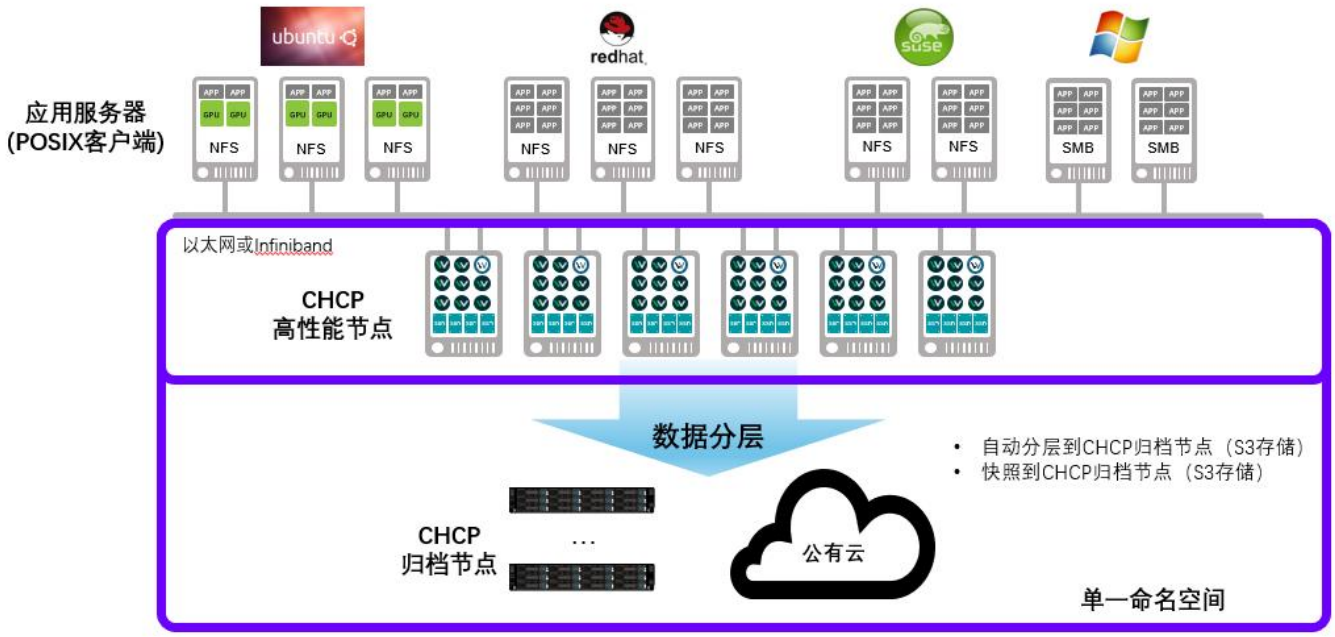


图 2–CHCP 将 NVMe 闪存与合在单个全局命名空间中

如图 2 所示，CHCP 独特的架构与传统存储系统、设备和基于系统管理程序的软件定义存储解决方案截然不同，因为它不仅克服了传统的存储扩展和文件共享限制，还允许通过 POSIX、NFS、SMB、S3 和 GPUDirect 存储并行访问文件。它提供了丰富的企业级特性，包括本地和远程云端快照、克隆、自动分层、云爆发、动态集群重新平衡、私有云多租户、备份、加密、身份验证、密钥管理、用户组、咨询、软硬配额等。

CHCP 的优点：

- 最高性能 – 适合大小文件工作负载
- 容量可扩展 – 最小配置为 30TB，并且可在单个命名空间中扩展到数千 PB
- 强大安全性 – 通过加密和身份验证确保数据免受威胁或恶意人员的攻击
- 混合云 – 可部署到公有云，以提高计算敏捷性或在云端以原生方式运行
- 备份 – 直接将备份推送到私有云或公有云进行长期保存
- 最佳经济性 – 将闪存与磁盘结合在一起，实现最佳的经济性

五、CHCP 架构

CHCP 的并行文件系统旨在提供类似云的体验，无论是在本地运行应用还是计划将应用移动到云端。CHCP 提供了本地和云端之间无缝迁移的能力。

大多数旧有的并行文件系统针对块存储运行文件管理软件，这种分层架构会影响性能。CHCP 是一种分布式并行文件系统，它消除了管理底层存储资源的传统块卷层。这种垂直集成的架构不受其他共享存储解决方案的限制，并提高了扩展性和性能效率（生产力）。

下图 4 简要列出了从应用层到物理持久介质层的软件架构。CHCP 核心组件（包括 CHCP 统一命名空间和其他功能，例如虚拟元数据服务器 (MDS)）在 Linux 容器 (LXC) 的用户空间中执行，有效消除了分时和内核特定的其他依赖关系。需要注意的例外是 CHCP 虚拟文件系统 (VFS) 内核驱动程序 - 为应用提供 POSIX 文件系统接口。与使用 FUSE 用户空间驱动程序相比，使用内核驱动程序可以提供更高的性能，并且允许需要完全兼容 POSIX 的应用在共享存储系统上运行。

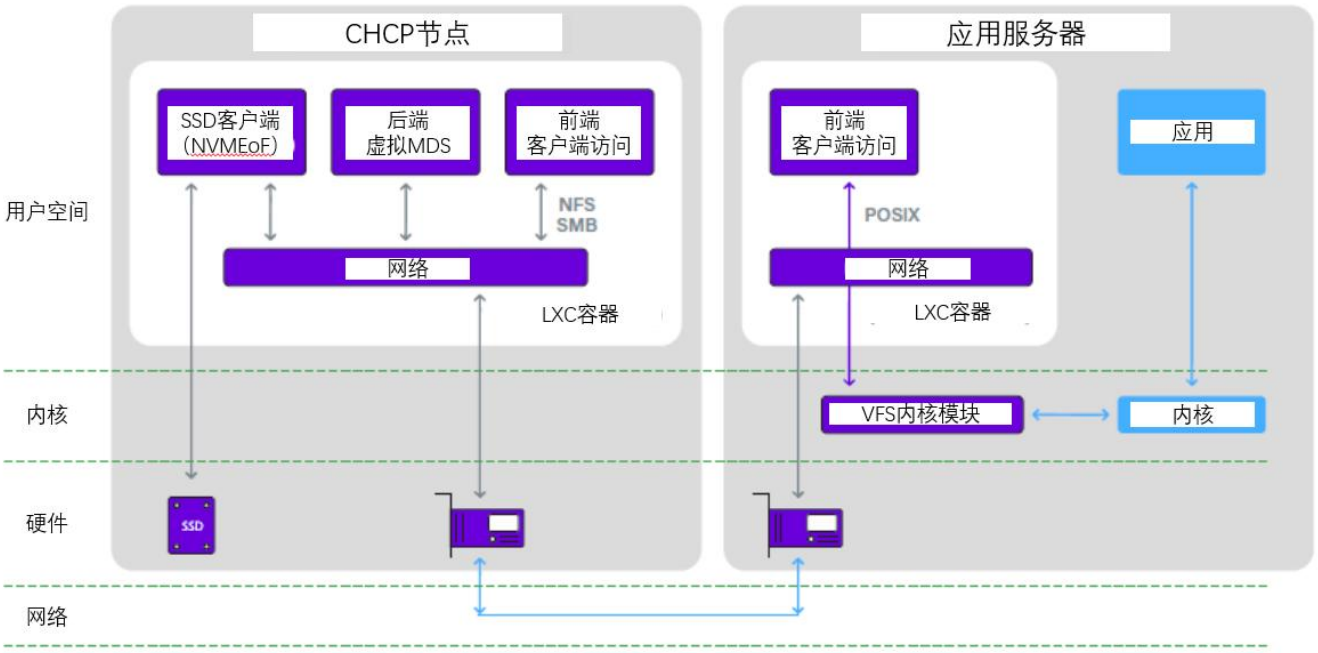


图 3 – CHCP 基于软件的存储架构

CHCP 支持所有主要的 Linux 发行版，并利用虚拟化和低级 Linux 容器技术在用户空间内与原始 Linux 内核一同运行自己的 RTOS（实时操作系统）。CHCP 管理分配给它的资源（CPU 内核、内存区域、网络接口卡和 SSD），以提供进程调度、内存管

理，并控制 I/O 和网络堆栈。由于不依赖 Linux 内核，CHCP 有效地利用了零拷贝架构，而且可以更好地预测延迟。

在 RTOS 中运行的 CHCP 功能（图 3）包括以下软件组件：

- 文件服务（前端）– 管理多协议连接
- 文件系统集群（后端）– 管理数据分布、数据保护和文件系统元数据服务
- SSD 访问代理 – 将 SSD 转换为高效的网络设备
- 管理节点 – 管理事件、CLI、统计和远程管理能力
- 对象连接器 – 读写到对象存储

CHCP 核心软件在 LXC 容器内运行。LXC 容器的优点是与其他服务器进程更好地隔离。CHCP VFS 驱动程序使 CHCP 能够支持完整的 POSIX 语义，对 I/O 采用无锁队列，以实现最佳性能，同时增强互操作性。CHCP POSIX 文件系统与本地 Linux 文件系统（例如 Ext4、XFS 等）具有相同的运行时语义，支持以前由于 POSIX 锁定要求、MMAP 文件、性能限制或其他原因而无法在 NFS 共享存储上运行的应用。与本地文件系统相比，这些应用的性能大大提高。

绕过内核意味着 CHCP 的软件堆栈不仅速度更快、延迟更低，而且可以在不同的裸机、虚拟机、容器化和云实例环境之间移植。

传统基于软件的存储设计通常存在资源消耗问题，因为这些解决方案要么接管整个服务器，要么与应用共享通用资源。这种额外的软件开销会增加延迟，并占用宝贵的 CPU 周期。相比之下，CHCP 只使用 LXC 容器中分配的资源，这意味着共享环境（聚合架构）中少则可以使用一个服务器内核和少量 RAM，多则可使用全部服务器资源（专用设备）。两种情况可使用同一软件堆栈。

5.1 文件系统设计

从最开始，CHCP 的设计初衷就是解决传统横向扩展 NAS 解决方案固有的许多问题。在设计时，关键的考量是构建一个软件定义存储平台，该平台可以在整个组织内或多租户环境中满足不同群体的需求。最常见的横向扩展 NAS 文件系统支持单个文件系统和单个命名空间的结构，利用目录和配额系统来分配资源和管理权限。虽然这种解决方案适用于较小规模的环境，但随着用户和/或目录数量的增加，管理变得越来越复杂。群组的完全隔离需要创建新的文件系统和命名空间，这会形成需要管理的物理存储孤岛。此外，目录扩展也是一个问题，通常需要创建多个目录来保持性能，这导致复杂性进一步提高。因此，CHCP 之所以不同于其他横向扩展 NAS 解决方案，原因在于它采用的是高性能节点和归档节点（S3 存储）二层的架构设计，在全局命名空间中的多个文件系统共享相同物理资源。每个文件系统都有自己的“角色”，可以通过配置而提供自己的快照策略、分层到 CHCP 归档节点（S3 存储）、组织、基于角色的访问控制 (RBAC)、配额等。CHCP 文件系统是一种逻辑结构，与其他解决方案不同，文件系统容量可以即时更改。挂载的客户端可以立即观察文件系统大小的变化，而无需暂停输入/输出。如前文所述，每个文件系统都可以选择分层到对象存储，而且如果是分层文件系统，热 (NVMe) 层和对象 (HDD) 层的比率还可以即时更改。文件系统可以划分为多个组织，由各自的管理员进行管理。单个文件系统可以支持数十亿个目录和数万亿个文件，从而提供一种比 NAS 系统更类似于对象存储的可扩展性模型，而且目录可以在不损失性能的情况下扩展。更多细节可见下文说明。目前，CHCP 在单个全局命名空间中支持多达 1024 个文件系统和 4096 个快照。这是一种人为限制，可以持续扩展。

CHCP 高性能节点扩展能力：

- 多达 6.4 万亿个文件或目录
- 多达 14 EB
- 一个目录中多达 64 亿个文件
- 单个文件多达 4PB

CHCP 归档节点扩展能力：

- 多达 6.4 万亿个文件或目录
- 一个桶中多达 1000 亿个文件
- 单个文件多达 4PB

5.2 支持的协议

CHCP 高性能节点具有适当凭证和权限的客户端可以使用以下协议创建、修改和读取数据：

- POSIX
- NVIDIA® GPUDirect® Storage (GDS)[4]
- NFS（网络文件系统）v3
- SMB（服务器消息块）v2 和 v3
- S3（简单存储服务）

注：CHCP 不支持 HDFS 协议（Hadoop 分布式文件系统）；然而，CHCP 的 POSIX 连接器可以直接挂载到 Hadoop 节点，以提供更高的性能。详情请参阅附件。

采用某种协议写入文件系统的数据可以使用另一种协议读取，因此，数据可以在应用之间完全共享。

CHCP 归档节点具有适当凭证和权限的客户端可以使用以下协议创建、修改和读取数据：

- 传统的 NFS/CIFS/FTP/SFTP 网络文件协议
- 与 OpenStack 兼容的 Swift 对象存储访问协议
- 自有的 REST 服务 API，支持长虹 CHCP 对标准 S3 的功能扩展
- SMTP 协议，可以直接对接邮件系统，实现邮件自动归档
- WebDav 协议

[4] NVIDIA GPUDirect Storage 是 NVIDIA 开发的一种协议，用于提高带宽并减少 NIC 和 GPU 内存之间的延迟。它目前可用于某些基于 NVIDIA GPU 的系统。

5.3 CHCP 存储节点

CHCP 存储节点包括高性能节点和归档节点两种。

CHCP 高性能节点采用软硬件一体化模式，包括两种型号：10224 和 10448。每种型号均配备适当的内存、CPU 处理器、网络和 NVMe 固态硬盘。配置 6 个存储节点才能创建一个能够在两个节点发生故障时继续运行的集群。

技术规格：

每台 CHCP 10224 设备：

- 4 个节点 / 2Rus
- 最大 250TiB

CPU	2x Ice Lake 6326R - 16 core
内存	256GB
硬盘	12 – 24 NVMe SSD
端口	2x 10GBASE-T 2x 25Gbe SFP28
AOCs	1x Intel 2 Port 100Gbe 或 1x Mellanox CX6 200Gbe/IB
读写 IO（M）	4.7
读写吞吐量（GB/s）	66.5

每台 CHCP 10448 设备：

- 8 个节点 / 4RUs
- 最大 3PiB

CPU	Milan 7413P - 24 Core
内存	256GB
硬盘	24 – 48 NVMe SSD

端口	2x 10GBASE-T 2x 25Gbe SFP28
AOCs	1x Intel 2 Port 100Gbe 或 1x Mellanox CX6 200Gbe/IB
读写 IO（M）	11.2
读写吞吐量（GB/s）	133.9

CHCP 归档节点采用软硬件一体化模式，由 G11 访问节点和 S11 存储节点组成。每种型号均配备适当的内存、CPU 处理器、网络、固态盘和大容量 NLSAS 机械盘。

技术规格：

	CHCP 归档节点	
节点型号	CHCP G11	CHCP S11
节点类型	访问节点	存储节点
硬盘	6/12 HDD x 4TB, RAID-6	200 HDD x 10/16/18TB/20TB 纠删码
规模	4 至 80 个节点	1 至 80 个节点
节点总计（容量）	1000PB	
文件数量	1000 亿对象	
硬件	每节点 2RU	每节点 5RU 至 9RU
CPU	2 x 10 核	2 x 8 核
内存	256GB-768GB	128GB
SSD	2 x 1.9TB	6 x 800GB
网络	4 x 10GbE Base-T 4 x 10GbE SFP+	8x SFP+ 10GbE 8x Base-T 10GbE 4x SFP28 25GbE

5.4 用于混合存储的 NVMe 闪存和磁盘层

CHCP 分布式文件存储设计由两个独立的存储层组成，分别是高性能层和归档层：其中高性能层是一个基于 NVMe SSD 的闪存层，为应用提供高性能文件服务，另一个归档层是基于 S3 对象技术的大容量 NLSAS 层，用于管理长期数据湖（如图 3 所示）。这两个层可以在物理上相互分离，但在逻辑上可以作为应用的一个扩展命名空间（Namespace）。CHCP 将命名空间从 NVMe 闪存层扩展到 NLSAS

层，提供了一个扩展到 EB 的全局命名空间。下文即将讲到，CHCP 利用基于对象存储技术的归档层来实现数据免备份、灾难恢复到另一个支持 CHCP 的存储集群或文件系统克隆。

5.5 网络

CHCP 系统支持以下网络技术：

- InfiniBand (IB) HDR 和 EDR
- 以太网 – 推荐 100Gbit 以上

客户可用的网络基础架构决定了选择哪种技术，因为 CHCP 无论使用哪种技术都提供了类似的性能。对于网络，CHCP 系统不使用标准的基于内核的 TCP/IP 服务，而是使用以下专有基础架构：

- 使用 DPDK 在用户空间中映射网络设备，并在没有任何上下文切换的情况下以零拷贝访问方式使用网络设备。这种绕过内核堆栈的方式消除了网络操作对内核资源的消耗，并且可以扩展到在多台主机上运行。这适用于后端和客户端主机，使 CHCP 系统能够完全饱和到 200Gbit 的以太网或 InfiniBand 链路。
- 实施基于 UDP 的专有 CHCP 协议，即底层网络可能涉及子网或支持 UDP 的任何其他网络基础架构之间的路由。客户端可以在不同的子网上，前提是它们可以路由到存储节点。

DPDK 的使用可实现高吞吐量和极低延迟的操作。低延迟通过绕过内核并直接从 NIC 发送和接收数据包而实现。由于同一主机中的多个核可以并行工作，消除了任何常见的瓶颈，这样可以实现高吞吐量。

对于不支持 SR-IOV（单根 I/O 虚拟化）和 DPDK 的旧有系统，CHCP 默认采用内核内部处理和 UDP 作为传输协议。这种操作模式通常被称为‘UDP 模式’，通常用于较旧的硬件，如 Mellanox CX3 系列 NIC。

除了兼容旧平台外，UDP 模式还将 CPU 资源提供给其他应用。当出于其他目的需要额外的 CPU 内核时，这可能很有用。

对于支持 RDMA 的环境（在 GPU 加速计算中很常见），CHCP 支持用于 InfiniBand 和以太网的 RDMA，以提供高性能，而无需将内核专用于 CHCP 前端进程。

应用程序客户端通过以太网或 InfiniBand 连接与 CHCP 存储集群连接。CHCP 存储集群支持 10GbE、25GbE、40GbE、50GbE、100GbE、200GbE 以太网，以及 EDR 和 200Gb HDR InfiniBand 网络。为了获得本文列出的最佳性能，CHCP 建议使用至少 100Gbit 的网络链路。

许多企业环境部署了混合网络拓扑，以支持高性能应用客户端（通常在 InfiniBand 上）以及更传统的基于以太网的企业客户端。CHCP 允许 InfiniBand 客户端和以太网客户端访问这些混合网络环境中的同一集群，允许所有应用利用 CHCP 的高性能节点。

5.6 网络高可用性 (HA)

CHCP 支持高可用性 (HA) 网络，以确保在网络接口卡 (NIC) 或网络交换机出现故障时继续运行。HA 在两个接口上执行故障切换和失效自动恢复，以实现可靠性和负载平衡，这可用于以太网和 InfiniBand。要支持 HA，CHCP 系统配置必须没有单点故障。需要多个交换机，而且主机必须与每台交换机连接。客户端的 HA 通过在同一客户端上实施两个网络接口而实现。对于单个双端口 NIC，CHCP 还支持以太网上计算客户端（模式 1 和 4）上的链路聚合控制协议 (LACP)。

CHCP 很容易使单个网络接口卡 (NIC) 的带宽饱和。可以使用多个 NIC 获得更高的吞吐量。使用非 LACP 方法可以设置冗余，使 CHCP 存储能够分别利用两个接口保障 HA 和带宽。

在用 HA 网络时，需要提示系统通过同一交换机在主机之间发送数据，而不是使用交换机互连 (ISL)。CHCP 系统通过网络端口标记来实现这一点，同时也保证了易用性。这样可以减少网络中的总流量。

注：与 RoCE 的实施不同（需要在交换结构中配置基于优先级的流量控制 (PFC)），CHCP 不需要无损网络设置来支持其 NVMe-over-fabrics 的实现，甚至可以在公有云网络中提供如此低延迟的性能。

5.7 协议

CHCP 存储集群支持多种协议以及跨各种协议的数据共享功能，允许不同的应用类型和用户共享单个数据池。与其他并行文件系统不同，CHCP 不需要额外的管理服务器基础架构即可提供这项功能。以下列出了当前支持的所有协议：

- 针对本地文件系统支持的完整 POSIX
- 针对 GPU 加速的 NVIDIA GPUDirect Storage (GDS)
- 针对 Linux 的 NFS
- 针对 Windows 的 SMB
- 针对对象访问的 S3

POSIX

CHCP 系统客户端是安装在应用服务器上的标准、符合 POSIX 协议的文件系统驱动程序，可实现对 CHCP 文件系统的访问。与任何其他文件系统驱动程序一样，CHCP 系统客户端拦截并执行所有文件系统操作。这样，CHCP 系统能够为应用提供本地文件系统语义和性能，同时提供集中管理、可共享且弹性的存储。CHCP 提供了高级能力，例如字节范围锁，并与 Linux 操作系统页面缓存密切集成，这一点稍后将在缓存部分介绍。

CHCP POSIX 客户端提供了最佳的 IOPS、带宽、元数据和延迟性能。

NVIDIA GDS

GPUDirect Storage 是 NVIDIA 开发的一种协议，采用 RDMA 提高服务器 NIC 和 GPU 内存之间的带宽并减少延迟。CHCP 完全支持 GDS，并且已经通过 NVIDIA 验证。

NFS

NFS 协议允许远程系统在没有 CHCP 客户端的情况下从 Linux 客户端访问 CHCP 文件系统。虽然这种实施方式无法提供 CHCP POSIX 客户端的性能，但它为部署和共享来自 CHCP 存储集群的数据提供了一种简单的方法。

CHCP 目前支持 NFS v3。

SMB

SMB 协议允许远程系统从 Windows 或 macOS 客户端连接共享文件服务。该协议支持以可扩展、弹性和分布式的方式实施 SMB，支持多种的 SMB 功能，包括：

- 通过 Active Directory 进行用户身份验证（原生和混合模式）
- POSIX 映射 (uid, gid, rid)
- UNIX 扩展
- SHA 256 签名
- 扩展的标识符空间
- 动态信用评分
- 持久文件打开，以处理连接中断
- 符号链接支持
- 可信域
- 加密
- 访客访问
- 隐藏的共享
- SMB ACL
- 从 Windows 向 POSIX ACL 转换
- 与 SMB 安全相关的共享选项

CHCP 目前支持 SMB v2.x 和 v3.x。

S3

目前，许多基于 Web 的应用都支持 S3 协议，然而，S3 的设计是以牺牲性能为代价实现可扩展性。物联网数据实时分析等应用可以通过高性能 S3 访问而受益。CHCP 针对其高性能文件系统提供了 S3 前端接口，以显著加速 S3 存储 I/O 的运行。特别需要指出的是，CHCP 大大改进了 S3 上小文件的访问性能。

5.8 管理 GUI

CHCP 提供了两种快速简便的 CHCP 文件系统管理方法，可以通过 ChangHong 统一图形用户界面 (Unified - GUI) 或命令行界面 (CLI) 或表述性状态转移 API (REST) 实现。报告、可视化和整体系统管理功能可以采用 REST API、CLI 或直观的 GUI 驱动的管理控制台访问。

用户可通过 ChangHong 统一图形用户界面管理 ChangHong 全系列存储系统（VSP 系列、CHCP 系列）。通过简单的点击式操作快速配置新存储；在全局命名空间中创建和扩展文件系统，建立分层策略、数据保护、加密、身份验证、权限、NFS 和 SMB 配置、只读或读写快照、对象快照和服务质量策略，并监视整体系统健康状况。通过详细事件日志，用户能够查看一段时间内的系统事件和状态，或者通过时间序列图功能深入了解精确时间点的事件详情。

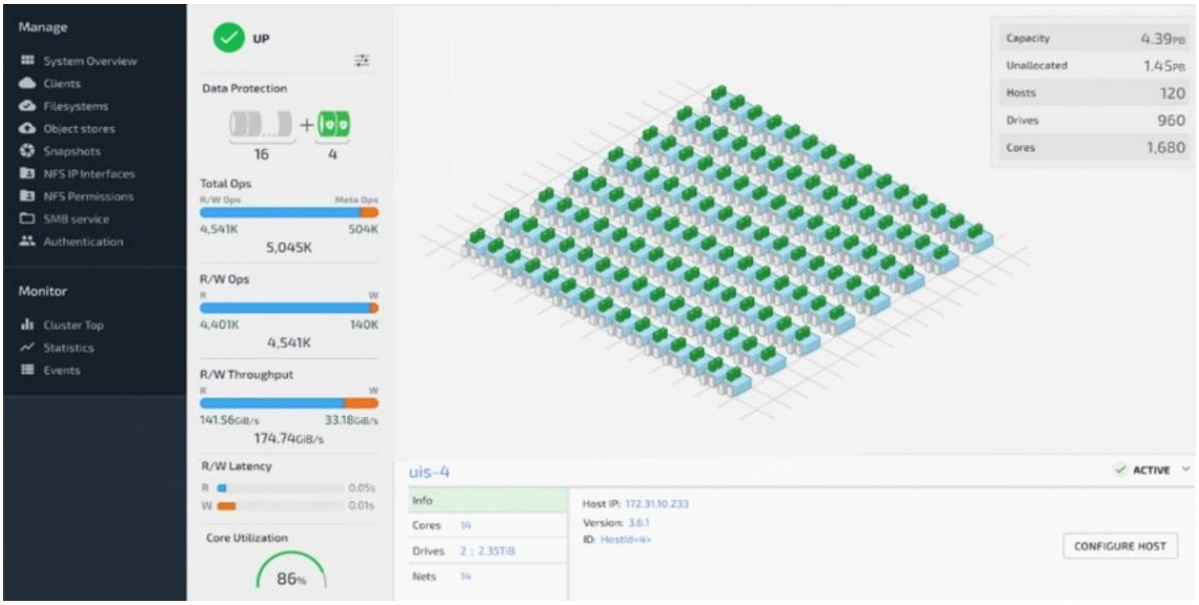


图 4– CHCP 管理软件用户界面

图 5 – 用于事件监控的时序图

系统事件选项卡列出了 CHCP 环境中发生的事件。此窗口中显示的事件也会传输到 CHCP 支持云，这样，后台支持部门可以使用这些事件在必要时积极为您提供协助，或在需要采取行动时主动通知您。CHCP GUI 完全基于 web，无需物理安装和维护任何软件资源，您随时都可以访问最新的管理控制台功能。

5.9 命令行接口 (CLI)

所有 CHCP 系统功能和服务都可以通过 CLI 命令执行。大多数 CHCP 系统命令适用于整个系统，可在所有集群节点上提供相同的结果。某些命令则针对特定节点执行，例如 IP 地址。

5.10 REST API

所有 CHCP 系统功能和服务都可以通过 Web 服务 API 执行。API 遵循 REST 架构的约束，也称为 Restful API。大多数 CHCP RESTful 命令都适用于整个系统，在所有集群节点上提供相同的结果。某些命令则针对特定节点执行。

六、功能详情

6.1 自适应缓存

应用（尤其是那些具有小文件和大量元数据调用的应用）可从本地缓存中大大获益。数据延迟非常低，而且这可以减少共享网络以及后端存储本身的负载。CHCP 文件系统提供了一种独特而先进的缓存能力，称为自适应缓存，允许用户充分利用 Linux 数据缓存（页面缓存）和元数据缓存（目录项缓存）的性能优势，同时确保整个共享存储集群的完全一致性。NFS v3 不支持一致性，因此，利用 Linux 缓存可能会导致读缓存的数据不一致以及写缓存的潜在数据损坏。CHCP 支持利用 Linux 页面缓存（通常是为直连存储 (DAS) 或在块存储上运行的文件服务保留的），使其在共享的网络文件系统上实现，同时保持完整的数据一致性。智能自适应缓存特性将主动通知任何客户端，即文件的独占用户（因此在本地缓存模式下运行）：另一个客户端现在可以访问数据集。一旦设置该标志后，客户端可以继续以本地缓存模式运行，直到文件被其他客户端修改。这时，CHCP 现在将使本地缓存无效，确保两个客户端仅访问数据的最新迭代。这确保了本地缓存在适当时候保持最高性能，并始终确保数据的完全一致性。CHCP 不需要特定的挂载选项即可利用本地页面缓存，因为系统会动态管理缓存，这样，CHCP 环境的配置非常易于管理，不存在导致数据损坏的管理错误的风险。

CHCP 为元数据缓存提供了相同的能力，也称为 Linux 目录项缓存。客户端可以利用目录的本地元数据缓存，从而显著缩短延迟。然而，一旦另一个客户端访问同一目录，CHCP 将确保一个客户端对目录的任何更改都会使访问该目录的所有其他客户端缓存的元数据无效。缓存能力还包括扩展属性和访问控制列表 (ACL)。

尽管某些共享文件存储厂商允许本地缓存，但其他文件系统都不具备 CHCP 的自适应缓存能力。默认情况下，缓存通常处于禁用状态，需要管理员更改挂载选项。这是因为写入一致性通常取决于客户端上某种形式的电池备份保护，以确保写入提交时的数据一致性。CHCP 的缓存立即可用，无需管理员的任何干预，因为它不依赖电池保护。这样，本地运行的软件可以在公有云中无缝部署，而无需更改软件。这一特性非常适合文件解压“Untar”之类的用例，作为本地进程运行的速度比跨共享文件系统运行快得多。

6.2 全局命名空间和扩展

CHCP 将存储系统中的所有数据作为全局命名空间的一部分进行管理，并支持单个混合架构中的两个持久存储层（用于活动数据的 NVMe SSD 的 CHCP 高性能节点和用于数据湖的基于 NLSAS HDD 的 CHCP 归档存储节点）。全局命名空间可添加选项将命名空间扩展到基于对象存储技术的 CHCP 归档节点命名空间，选项可以在安装期间或安装之后随时配置，只需在管理控制台单击几下鼠标即可完成。文件在活动时或者根据预设或用户定义的策略分层到 CHCP 归档节点之前驻留在闪存上。在文件分层到 CHCP 归档

节点后，原始文件将保留在闪存层，直到新数据需要物理空间，因此，在被覆盖之前，原始文件将作为缓存文件。尽管文件数据被转到 CHCP 归档节点中，但文件元数据始终保留在本地闪存层上，因此，所有文件都可供应用使用，无论位于哪个物理位置，即使是存储桶位于公有云中。随着 CHCP 高性能节点中的 NVMe 闪存系统容量即将用尽，使用率达到较高水平，数据被动态推送到归档层，这意味着您永远不必担心闪存层的容量不足。这对于写密集型应用特别有用，因为不需要管理员干预。高性能层和基于 HDD 的归档层可以根据所需的使用容量独立扩展。

全局命名空间可以细分为 1024 个文件系统，而且文件系统容量可以随时动态扩展，无需卸载和挂载文件系统，只需为其分配更多空间即可。通过对命名空间进行分段，可以将存储容量分配给单个用户、项目、客户或任何其他参数，并且可以轻松地集中管理。文件系统中的数据与其他文件系统完全隔离，以防止互相干扰。

6.3 集成数据管理

CHCP 提供了内置的、基于策略的自动数据管理功能，可以根据数据热度在不同存储层之间透明地移动数据。CHCP 支持将数据从基于 NVMe 闪存的高性能层移动到基于 NLSAS 的 CHCP 归档节点或 ChangHong VSP 系列存储。数据的移动在 CHCP 统一管理界面设置，是 CHCP 高性能层的可选扩展功能。例如，为了始终确保最高性能，用户可能希望将某些文件系统仅保留在高性能节点（NVMe SSD）上，而其他文件系统则将数据移动到 CHCP 归档节点或经虚拟化接管的 ChangHong VSP 系列存储系统，以实现最佳成本经济性。

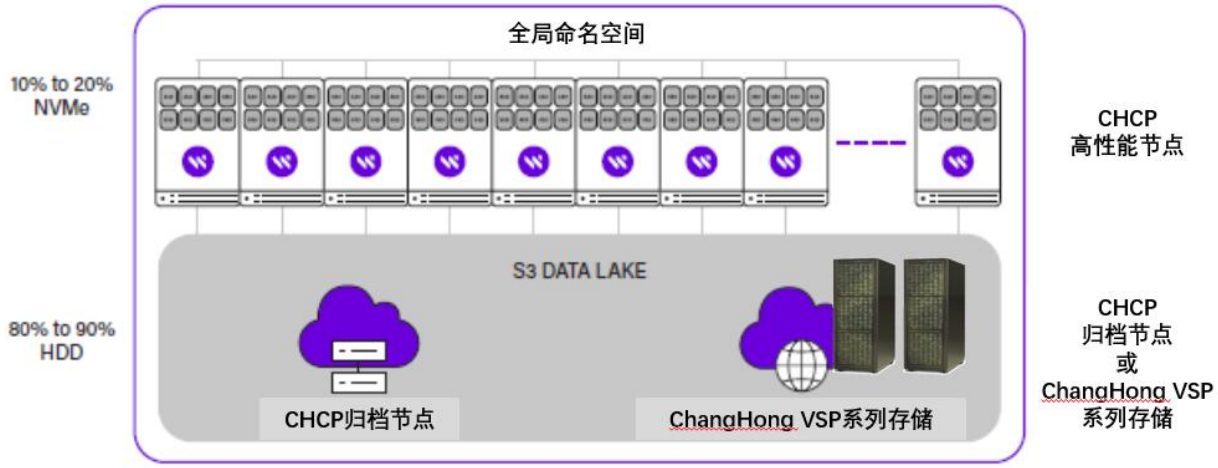


图 6– 分层到 CHCP 归档节点或任何兼容 S3/Swift 的对象存储

元数据始终存储在 CHCP 高性能节点，而且整个文件系统（包括其数据结构）的只读或读写快照可以存储在 CHCP 归档节点或经虚拟化统一纳管的 ChangHong VSP 系列存储系统，以防闪存层出现故障。应用客户端可以看到指定文件系统中的所有文件，无论在哪个位置，因此不需要更改应用即可利用成本经过优化的解决方案。

NVMe SSD 相对于 HDD 上应该存储多少数据并没有硬性规定，但对 CHCP 客户的调查指出，目前闪存上的典型分布约为 20%。我们的经验是测量闪存层，使其能够保存当前工作负载的正常工作数据集，并具有足够的容量用来预处理新工作负载，从而获得最佳性能。

6.4 快照和克隆

CHCP 支持用户自定义的快照，用于进行日常数据保护（包括备份）以及云数据迁移。例如，CHCP 快照可用于在本地备份高性能层上的文件，以及将副本复制到 CHCP 归档层，用于备份或灾难恢复目的。此外，CHCP 快照可用于将文件和数据保存（停放）到成本较低的冷存储中，如公有云和本地对象存储。除了时间点快照外，CHCP 还可以创建完整克隆（可以转换为可写快照的快照），并使用指针指向原始数据。CHCP 快照和克隆即时进行，并且在第一次发生后会有所差异，这样可显著减少数据保护所需的时间和存储容量。此外，整体系统性能不受快照进程或写入克隆时的影响。快照可以从 GUI、CLI 或通过 REST API 调用而创建。CHCP 支持：

- 只读快照
- 读/写快照
- 删除主快照，保留其他所有版本
- 删除任何快照，保留以前和以后的版本
- 将只读快照转换为读/写快照
- 快照到对象（见下节内容）

6.5 WORM 与多版本

CHCP 归档节点支持通过 WORM（一写多读）技术，在保留期间内防止任何人对信息进行修改，满足内控、审计程序对信息原真性的要求。当审计人员需要检索、抽取特定的信息纪录时，能够通过专用的检索工具快速发现所有满足条件的信息。

当更改文件时，新文件被保存，旧文件成为一个新版本被自动保存。实现了数据版本级数据保护，可实现多达 10 个版本的文件版本保存。用户随时可以访问任意某一个版本文件。

6.6 高存储效率

CHCP 的归档节点采用产品化的硬件处理节点，并配置冗余双控制器，磁盘部署采用 20+6 的纠删码技术保护，保证在一个磁盘校验组中任意 6 块盘损坏或故障都不影响数据完整性和数据访问；存储节点内采用冗余控制器与硬盘容量单元分离式设计，控制器任何部件故障都有冗余部件接管且不影响任何底层硬盘容量单元访问，也不会导致节点的硬盘数据重建，纠删码设计见下图：

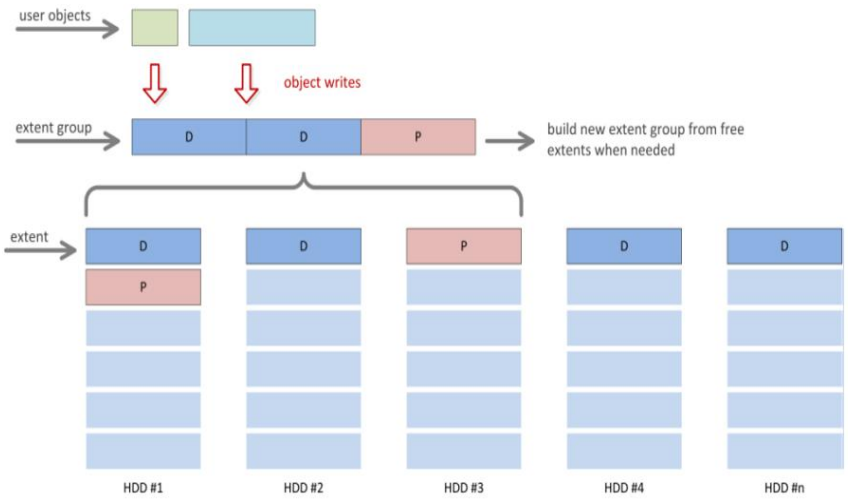


图 7– CHCP 归档节点纠删码（20+6）

存储纠删码设计特点为：

- 避免 RAID 重建的长时间和性能影响；
- 20+6 的 Extent Group 提供 77%的利用率和 15 个 9 的数据可用性；
- 不需要专有 Hot Spare, 所有磁盘的可用空间可用于数据修复；
- 简化部署和管理，且易于扩展。

6.7 重删和压缩

CHCP 归档节点提供的全新数据重复数据删除和压缩服务来最大限度地利用其存储资源，提升存储资源利用率。重复数据删除也称为单一实例存储(SIS)。CHCP 提供的这种新型存储服务远胜于其它同类竞争产品，它可以同时提供 hash 对比和二进制对比，因此能够确认对象是否是重复数据，从而避免了“hash collisions”，避免不同对象却具有相同的加密 hash 密钥的情况。CHCP 的相关机制还能够让用户清楚地看到被删除的重复数据的数量，以及节约下来的总存储容量。

CHCP 归档节点的 Duplication Elimination 是个后台的进程，CHCP 会每天在合适的时间进行，客户可以在控制台进行干预它的启动和停止。该进程在意外停止后，还会持续运行。

6.8 多副本保护

虽然单个存储节点整体发生故障的概率很小，但并不意味着不重要。CHCP 归档节点后端可以挂接多个存储池，而独有的数据多副本技术可以将数据存放到不同的存储池，可根据业务需要设置副本保护级别(DPL)，比如设置 DPL=2 则保存 2 个数据副本，通过多副本即使极端情况单个存储池故障也不会丢失数据且对应用透明。CHCP 可以支持多副本和纠删码同时使用，实现数据的多

重保护，进一步提高存储空间持久可靠性。

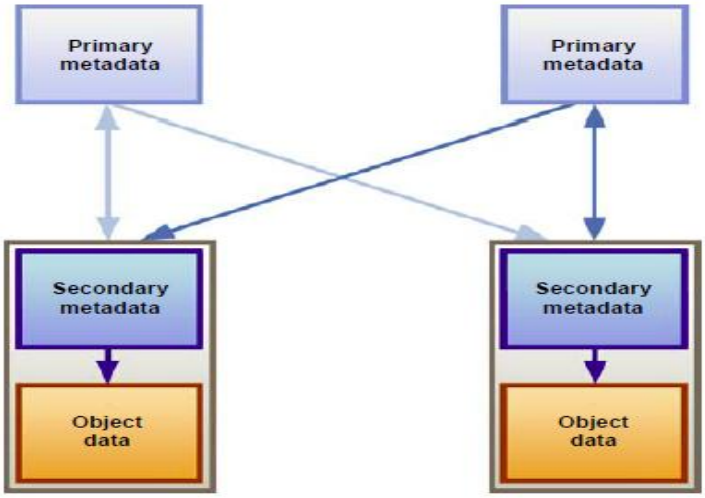


图 8：CHCP 数据跨存储池多副本保存

6.9 数据自愈

CHCP 提供自动检查和自动修复功能，即 CHCP 会定期检查所有对象数据文件，通过检查每一个对象文件的 hash 密钥，查收是否有文件内容被修改或出现错误，如果有文件出现意外损坏或篡改，CHCP 会自动修复这些文件，保证与保存时生成的 hash 密钥一致。

6.10 高性能节点数据保护

对于任何存储系统来说，数据保护都是一项关键功能，这方面的挑战非常巨大。如果没有适当的保护模式，则需要限制文件系统的大小，以适应重构时间窗口，并将数据暴露的风险降至最低。常见的数据保护方案（如 RAID[6]、复制和纠删码）是扩展性、保护、容量和性能之间的权衡。

CHCP 没有数据或元数据局部性的概念，因为所有数据和元数据都均匀分布在存储节点上，这样提高了扩展性、聚合性能和弹性。在高速网络上，数据局部性实际上会产生数据热点和系统扩展性问题，进而导致性能和可靠性问题。通过直接管理 SSD 层上的数据放置布局，CHCP 可以分割数据并将其分布在存储集群中，从而根据用户可配置的存储条带大小达到最佳的放置效果。CHCP 采用先进的算法确定数据布局；分割的数据与底层闪存使用的数据块大小完全匹配，以提高性能并延长 SSD 使用寿命。条带大小可以设置为 4 到 16 之间的任何数值，而奇偶校验可以设置为+2 或+4。图 9 显示了在 4+2 配置中跨 SSD 的数据放置布局。建议的最小集群大小为 6，这允许从 4+2 配置中重构两个完整的虚拟备用。CHCP 集群越大，它支持的条带大小就越大，而存储效率

和写入性能也就越高。

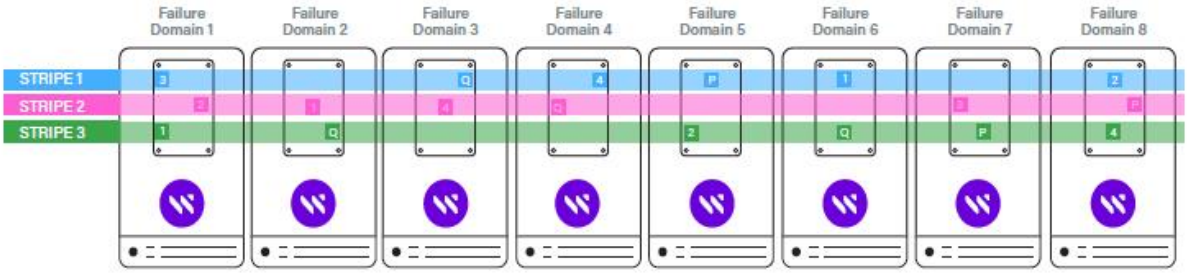


图 9：CHCP 数据分布

6.11 高性能节点数据保护模式

CHCP 管理数据的保护，因此，数据始终安全且可访问：

- 可配置的数据保护等级：从 4+2 到 16+4
- 拥有专利的分布式数据保护模式
- 可配置的故障域
- 端到端数据保护校验和保障数据完整性
- 元数据日志
- 网络冗余:通过两个网络端口支持两个架顶式交换机
- 本地或基于云的快照和克隆
- 快照到对象，用于备份和灾难恢复
- 自动分层到云端

[6] RAID = 独立磁盘冗余阵列

CHCP 使用故障域定义数据保护等级。故障域可以从服务器节点级别开始灵活配置，可提供单个或多个 SSD 级别的配置粒度。根据服务器集群的大小和规模，数据保护级别可灵活设置（集群越大，建议配置的数据条带大小越大），这样能够以最佳方式利用 SSD 容量，提高性能和弹性。对于细粒度保护，数据保护等级在集群级别设置，奇偶校验可以设置为两个或四个，这意味着系统最多可以承受两个或四个节点同时发生故障，而不影响数据可用性。

CHCP 的数据保护模式遵循数据 (N) +奇偶校验（2 到 4）的惯例，而 N+4 数据保护等级是云存储中独有的。与三重复制相比，CHCP 的数据保护模式提供了更好的弹性（三重复制只能保护两次故障），而且不会对存储和吞吐量产生重大影响。N+2 保护等级对于大多数生产环境都已足够，无论是使用聚合集群还是专用设备。对于具有大量（数百个）聚合集群节点的集群，建议使用 N+4 保护等级，因为应用可能会影响服务器可用性。

6.12 业务连续性

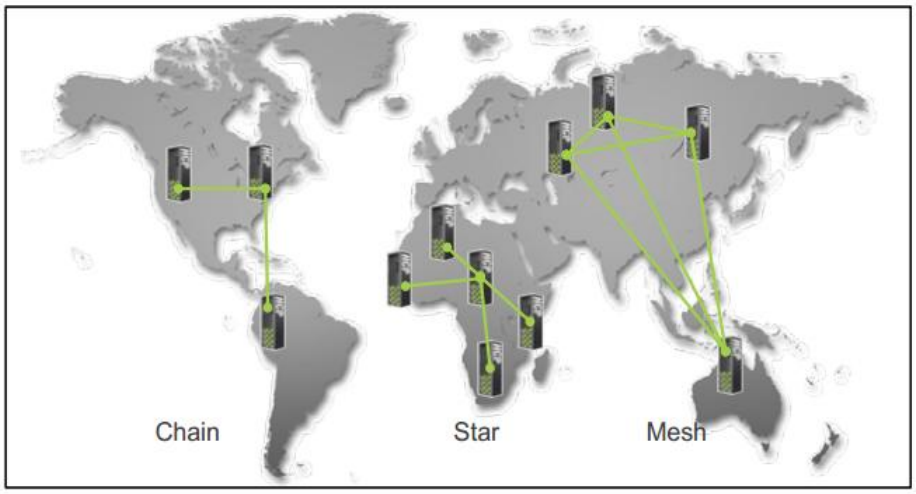
对数据安全的威胁，主要包括硬件故障威胁、逻辑（软件）错误、和灾难影响三大部分。对于硬件故障威胁，主要包括节点失效和磁盘失效两种问题。对于节点失效故障，CHCP 归档节点设计为完整的高可用架构，全部硬件部件均排除单点。G11 处理节点集群满足 N/2-1 的节点冗余性，即集群在 N/2-1 节点故障的情况下，仍能对应用提供 IO 访问并保证数据完整性（注：当处理节点集群后端为集中式存储时，集群可提供 N/2 的节点冗余性，即集群在 1/2 节点故障的情况下，仍能对应用提供 IO 访问并保证数据完整性。）。

S11 存储节点为双控架构，在一个控制器失效情况下仍能对外提供 IO 访问并保证数据完整性。对于磁盘失效故障，CHCP 归档节点支持 20+6 的纠删保护机制，可以满足在最大不超过 6 块盘同时失效的情况下，保证数据不丢失。

对于逻辑错误，CHCP 归档节点提供了纠删码、文件自愈、WORM、多版本、多副本等软件功能，且上述功能可以同时启用。对于容灾，CHCP 提供了 Active-Active 存储复制功能，可以构建多种容灾解决方案。

复制是被设计利用不同地理区域，使用异步复制技术。这个适用于所有的数据、元数据和策略。一旦被启动它将是一个自动的服务。

长虹 CHCP 存储归档节点支持扩展多达 5 个节点集群的跨区域部署，并可实现跨区域多活数据复制和跨区域纠删的不同保护方式，两种方式都支持在任意区域可在线访问多活数据，多活复制基于站点级，实现单套 CHCP 归档节点或单个站点不可用时，数据不会丢失和损坏。站点间建立复制连接，形成链状、星形、全连接三种不同的拓步结构。在复制管理层面，管理粒度可以细化到桶（Bucket）级别。



CHCP 的多活复制为异步复制，因此部署扩展到远程异地范围。在这种方式下，也可以构造类似于 CDN 的访问模式。即各个节点仅复制元数据，当用户和应用访问最近的 CHCP 存储时，按需把远端的数据拉取过来。不仅增加了数据可用性和宕机切换以及数据恢复的流畅性；而且增加了产品性能和生产力，改善合作协同效率。

6.13 数据分布

CHCP 采用了一种专用分布式数据保护编码方案，该方案随着集群规模中节点数量的增加而提高弹性。它提供了纠删码的可扩展性和耐久性，但性能没有降低。与传统的硬件和软件 RAID 及其他数据保护方案不同，随着系统的扩展，CHCP 的重构时间更快，而且更有弹性，因为集群中的每个节点都参与了重构过程。

CHCP 将数据和元数据均匀分布到跨故障域 (FD) 的逻辑节点上。故障域可以是服务器节点、机架，甚至是数据中心。在云环境中，CHCP 可以跨越可用性区域 (AZ)。

与传统硬件和软件数据保护方案不同，CHCP 仅在任何一台服务器（或故障域）中放置一个数据条带，因此，在单个服务器内发生两硬盘故障的情况下，这仍将被视为单一故障。即使节点有多个 SSD，数据条带也始终分布在不同的服务器节点、机架或 AZ 上，具体取决于选择的弹性。CHCP 的弹性可由用户配置，用户可以定义 CHCP 集群内的故障数量，以满足应用工作负载服务等级的要求。当发生故障时，无论定义的域有多大，系统都将故障诊断视为单一故障。除了在 FD 级别分布数据条带外，CHCP 还确保高度随机化的数据放置布局，以提高性能和弹性。随着集群大小的增加，硬件故障的概率成比例提高，但 CHCP 通过以随机方式分布条带克服了这个问题。节点数量越多，随机条带组合的数量就越多，而这导致出现双重故障的概率越低。例如：对于条带大小为 18 (16+2) 和集群大小为 20 的情况，可能的条带组合数为 190，但是随着集群大小增加到 25，可能的条带组合数达到 480,700 个。

6.14 CHCP 重构

CHCP 采用多种创新策略使系统尽快恢复到完全受保护的状态，并准备好处理后续故障。这确保了应用程序不会受到长时间重构过程的影响。

CHCP 在文件级别保护数据，因此只需重构那些在故障服务器或 SSD 上主动存储的数据。这意味着与在数据块层面保护数据的传统 RAID 解决方案或文件服务器相比，重构速度更快。基于 RAID 控制器的系统通常要重构受影响的设备上的所有数据块，包括空块，这样会延长重构和风险暴露时间。CHCP 只需要重构受故障影响的数据。CHCP 分层策略的另一个好处是，已经分层到对象存储的数据永远不会受到节点故障的影响，因为它在对象存储上受到保护。

此外，任何缓存数据（已分层到对象但仍保留在闪存层上作为缓存的数据）也不需要重构，仅优先重构驻留在闪存层上的数据。

由于所有服务器节点（故障域）共享所有条带，对于发生故障的组件，其条带由其余良好节点共享；因此，所有健康节点都参与恢复过程，包括虚拟（热）备用。这意味着集群越大，重构速度越快，系统就变得越可靠，因为参与重构过程的计算资源越多，条带就越随机。在发生多个故障的情况下，系统会优先从保护水平最低的数据条带开始数据重构。CHCP 寻找故障节点都使用到的数据条带，并首先重构这些数据条带，使系统能够尽快恢复到更高的弹性级别。这种优先重构过程将继续进行，直到系统恢复到完全冗余状态。相反，在复制的系统中，只有镜像服务器参与恢复过程，这会显著影响应用性能。纠删码也面临类似的问题，其中只有一小部分服务器参与恢复。对于 CHCP，恢复率可由用户配置，用于重构的网络流量可以随时更改，因此，管理员可以完全控制，以确定持续应用性能和恢复时间之间的最佳权衡。

6.15 IPv6 支持

CHCP 原生支持 Ipv6。由于 IPv4 最大的问题在于网络地址资源有限，严重制约了互联网的应用和发展。IPv6 的使用，不仅能解决网络地址资源数量的问题，而且也解决了多种接入设备连入互联网的障碍。

6.16 自动数据重新平衡

CHCP 主动监测和管理 CHCP 集群的性能、可用性和容量健康状态。这允许系统计算节点的利用率（性能和容量），从而在集群中自动且透明地重新分布数据，以防止出现热点。

这样做的好处是，随着容量和使用率的变化，CHCP 可以保持良好平衡的集群性能，并提供数据保护。另一个优点是，随着向现有服务器节点添加更多固态硬盘或者使用更多节点扩展集群，CHCP 会自动重新平衡，以提高性能、弹性和容量，而无需花费高昂的停机时间进行数据迁移。这不需要您配置容量相匹配的固态硬盘，您可以在固态硬盘价格下降时利用新技术，并节省资金。

6.17 身份验证和访问控制

CHCP 在用户级和客户端-服务器级提供身份验证服务，以验证用户或客户端是否具有查看和访问数据的安全等级。CHCP 允许不同的挂载验证模式，如只读或读写，并在文件系统级定义。

经过验证的挂载权限在组织级进行定义，并采用加密密钥加密。仅拥有适当密钥的客户端才能访问经过身份验证的挂载点。这种方法极大地限制了对组织内某些子集的访问，并限制了对具有适当加密密钥的客户端的访问，从而提高了安全性。CHCP 支持：

- LDAP（轻量级目录访问协议），一种跨多个不同平台提供目录服务的网络协议。
- Active Directory，Microsoft 的 LDAP 实施协议，这是一种目录服务，可以存储关于网络资源的信息。它主要用于验证想要加入集群的用户和组。
- CHCP 还提供基于角色的访问控制 (RBAC)，为用户和管理员提供不同的权限等级。某些用户可以被授予完全访问权限，而有些用户则具有只读权限。

每个 CHCP 系统用户拥有以下某种定义的角色：

集群管理员：与普通用户相比，该用户拥有额外的权限。其中包括以下能力：

- 创建新用户
- 删除现有用户
- 更改用户密码
- 设置用户角色
- 管理 LDAP 配置
- 管理组织

此外，为避免集群管理员失去对 CHCP 系统集群的访问权限，集群管理员用户有以下限制：

- 集群管理员不能删除自己
- 集群管理员不能将其角色改为普通用户角色

组织管理员：该用户的权限与集群管理员类似，但这些权限局限于组织级。他们可以在组织内执行以下操作：

- 创建新用户
- 删除现有用户
- 更改用户密码
- 设置用户角色
- 管理组织 LDAP 配置

另外，为了避免组织管理员失去对 CHCP 系统集群的访问权限，组织管理员有以下限制：

- 组织管理员不能删除自己
- 组织管理员不能将其角色改为普通用户角色

普通用户：拥有读写权限的用户

只读：拥有只读权限的用户

6.18 动态和静态加密

CHCP 提供了从客户端到对象存储解决方案的完整端到端加密，使其成为最强大的商用加密文件系统。加密是在创建文件系统时在文件系统级设置，因此，某些被视为关键的文件系统可以加密，而其他文件系统则不能。CHCP 的加密解决方案可防止物理介质失窃、SSD 上的低级固件遭黑客攻击和网络上的数据包窃听。文件数据采用 FIPS 140-3 1 级兼容加密密钥 XTS-AES 进行加密（使用 512 位密钥），可提供 256 位的有效安全防护。

CHCP 已经证明，在使用 CHCP 客户端时，加密文件系统对应用性能的影响可以忽略不计。

6.19 密钥轮换和密钥管理

CHCP 支持任何遵循 KMIP（密钥管理互操作性协议）以及 Hashicorp Vault 专有 API 规范的密钥管理系统 (KMS)。集群密钥由 KMS 轮换，文件系统密钥可以通过 KMS 轮换，并用新的 KMS 主密钥重新加密，而文件密钥可以通过复制文件进行轮换。

对象存储上的文件数据也被加密。在将快照上传到对象存储时，除其他文件系统参数外，文件系统密钥也包含在内，并使用特殊的“特定备份”集群密钥进行加密，该密钥可通过 KMS 获得，可用于所有快照到对象的备份和还原。在 CHCP 将快照推送到 AWS 云时，数据受到充分保护，而且只能通过本地 KMS 系统进行身份验证。

七、设备更新

根据 Gartner 的统计信息，平均三分之二的企业 IT 预算用于维持现有的 IT 基础架构的运行，所以加速创新来减少运维成本对保持企业竞争力具有重要作用，CIO 们为此面临着巨大压力。

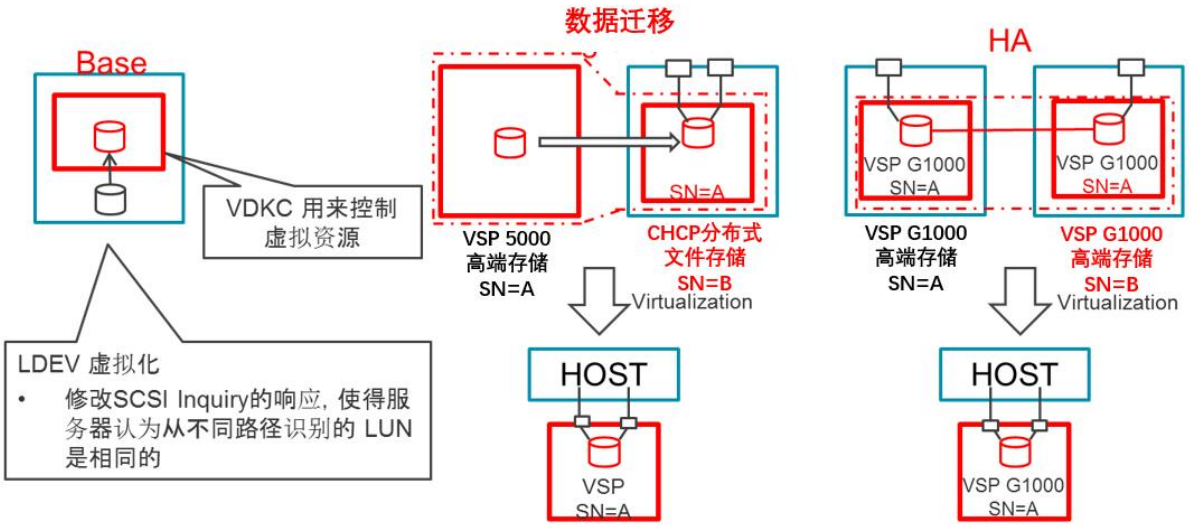
由于数据持续的增长，IT 部门不断地在处理与变更相关的需求，包括存储扩容，存储集中，存储整合，设备更新换代，满足互操作性要求等等。在 IT 设备平均四到五年的生命周期中，数据迁移将会多次发生。

ChangHong 解决之道

借助 ChangHong 无中断数据迁移（NDM），用户将告别停机窗口。无中断数据迁移的技术核心是在 ChangHong 全系列存储系统（VSP 系列、CHCP 系列）中引入了虚拟存储的概念（vDKC）。虚拟存储实际上是物理存储中的一组逻辑资源，虚拟存储的信息包括了虚拟的设备序列号、WWN、SSID，以及数据卷 LDEV 地址信息，由于源存储及目标存储中的虚拟存储使用了完全相同的信息，因此，服务器不会察觉到所使用的资源实际上是分布在不同的存储设备中。

借助虚拟存储的概念，数据源存储设备的 ID 被完整地复制到数据目标存储设备上，而服务器无法察觉存储设备物理身份的变化，这一过程对任何操作系统、虚拟机监控程序，服务器、服务器的路径管理软件，服务器集群软件以及存储网络连接等都是透明的。不仅如此，NDM 相比同类产品对源数据卷的不同类别支持具有局限性问题，比如：数据迁移不支持数据克隆卷、数据快照卷、远程复制卷等等，NDM 可以支持将具有复制关系的数据卷进行完整迁移，迁移后复制关系得到保留，用户无须因为数据迁移而不得不重新建立数据复制关系，而且重新完成数据的初始复制，从而能够满足多样的存储环境，大幅压缩迁移所消耗的时间。

在大型企业中，一台服务器通常连接多台存储设备，NDM 帮助客户实现了迁移过程的无中断和简化，相对业内用于数据迁移的平均人力开销和费用，可大幅节省 90%甚至更多。



如上图所示，主机识别 LUN 是通过控制器 ID 来识别，VDKC 是长虹 VSP 5000 / 长虹 CHCP 上虚拟出来的一个虚拟控制器，它可以将多台存储底层的物理控制器虚拟成一个控制器，这样主机通过虚拟控制器访问后端磁盘资源时始终和一个控制器 ID 交互，无论后台存储如何变化，主机都不会有感知，从而实现了数据迁移、存储底层复制、设备更新、双活等特性。

特性：

- 远端的数据拷贝与本地的数据拷贝或生产数据永远保持一致，远端拷贝永远是本地数据盘的“镜像”
- 目标存储系统总是与源存储系统数据同步，源存储系统与目标存储系统同步进行相同的 I/O 更新，目标存储系统在更新时

总是与源存储系统保持完全一致的顺序，以保证数据的一致性和完整性。当生产中心发生灾难时，不会出现数据丢失。

- 不依赖于主机系统、文件系统、数据库系统，基于存储系统的工作机制，利用存储系统控制器的控制台来启动、监控、控制远程数据备份的操作。节省主机系统的 CPU 资源，提供用户开放的高可用性

八、数据集成及检索

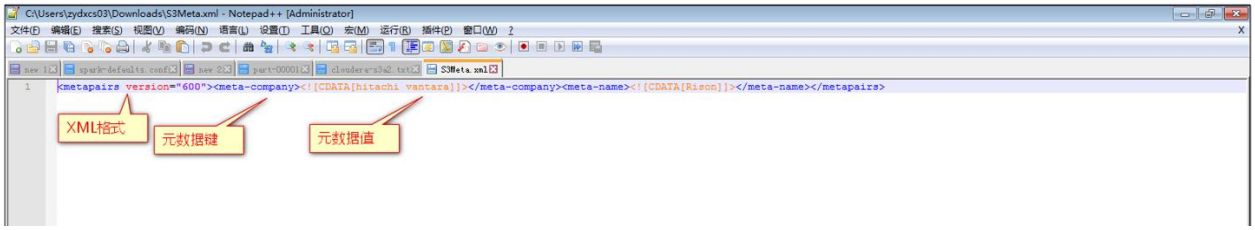
8.1 概述

借助 CHCP，您可以访问元数据和内容搜索工具，以实现更平滑的自动化查询，从而更快获得更准确的结果。通过这些特性，您可以更好地了解存储文件的内容，如何使用内容以及对象如何相互关联。这些知识可以帮您实现更智能的自动化，以及基于最佳元数据架构的大数据分析。

长虹 CHCP 存储包含全面的内置搜索能力，使用户能够搜索桶中的对象，根据元数据分析桶，并且操作对象组，以支持审计和诉讼过程中的电子发现操作。搜索引擎 (Apache Lucene) 在 CHCP 处理节点上执行，并且可以在租户和桶级别启用。CHCP 支持两种搜索方法：

1) 基于 Web 的用户界面（称为搜索控制台）提供了一个交互式界面，用于使用“AND” 和 “OR”逻辑创建和执行搜索查询。带有下拉输入字段的模板会提示用户输入各种选择条件，例如在特定日期之前存储或大于指定大小的对象。 可点击的查询结果显示在屏幕上。在搜索控制台中，搜索用户可以打开对象，对对象执行批量操作（保留、释放、删除、清除、权限删除和清除、更改所有者、设置 ACL），并以标准文件格式导出搜索结果，用作其他应用的输入信息。

2) 元数据查询 API 使 REST 客户端能够以编程方式搜索 CHCP。与搜索控制台一样，对查询的响应结果是符合查询条件的对象的元数据，这些元数据采用 XML 或 JSON 格式。

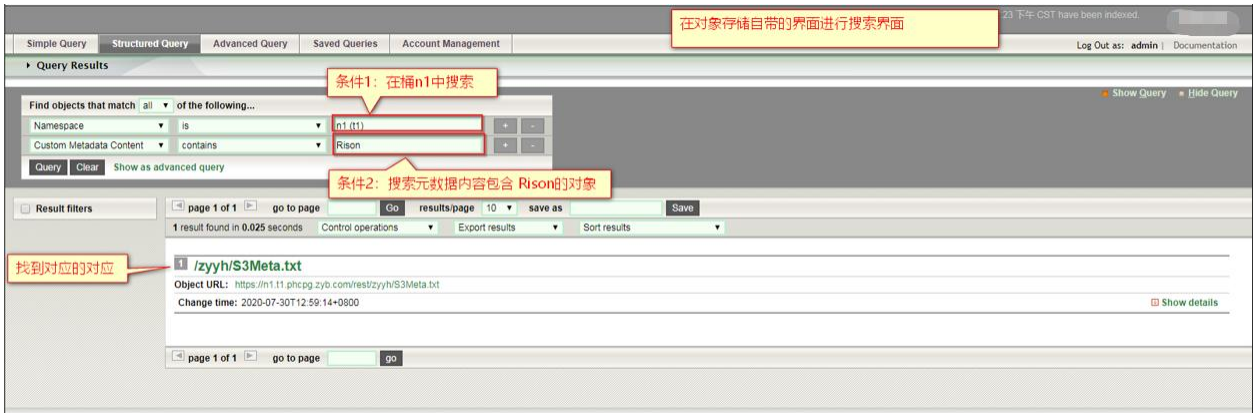


图：XML 格式元数据注入

在任何情况下，支持的查询类型有两种：

基于对象的查询根据其元数据定位存储库中当前存在的对象，包括系统元数据、定制元数据和 ACL，以及对象位置（桶或目录）。可以在基于对象的查询中指定多个可靠的元数据标准。要支持此类查询，必须对对象编制索引。

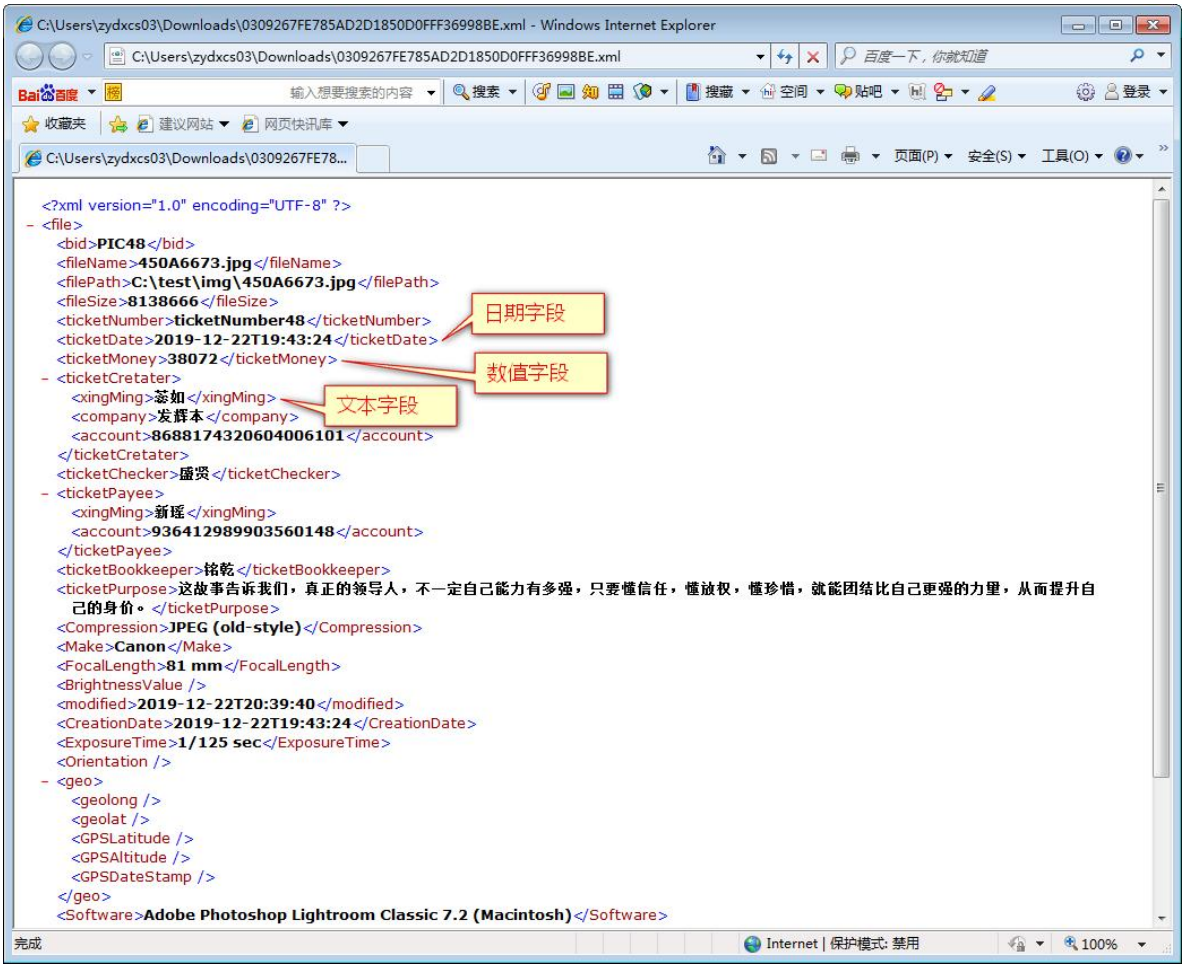
基于操作的查询是基于时间检索对象事务。它根据在指定时间段内对对象执行的操作来搜索对象。另外，它还检索对象创建、删除和清除（用户发起的操作）以及处置和修剪（系统发起的操作）的记录。基于操作的查询不仅返回当前存储库中的对象，还返回已删除、已处置、已清除或已删除的对象。



图：CHCP 内嵌搜索引擎界面

8.2 客户化元数据注释

每个 CHCP 对象最多支持 10 个自由格式的 XML 元数据注释，总容量为 1GB。这样，单独的团队可自由地独立工作和搜索元数据。分析团队可以添加其应用特定的注释，这与计费应用不同。与简单键值对相比，XML 注释具有显著的优点，因为搜索引擎可以使用 XML 返回更相关的结果。



图：自定义元数据 schema

表：元数据注释举例

本 XML 记录样例呈现了单个注释	键值对
<Record> <Dr>John Smith</Dr> <Patient>Jonn Smith</Patient> <Address>St John Smith Square</Address> </Record>	Dr=John smith Patient= Jonn Smith Address=St John Smith Square</Address>

现在设想一下这样的搜索：您需要一个与“John Smith” 医生相关的对象。XML 记录允许您将搜索结果精确定位到这个字段，而键值对会生成更大的搜索命中集。随着对象数量增长到数百万甚至数十亿个，键值搜索很快就会变得速度缓慢，而且任务艰巨。

8.3 CHCP 元数据检索接口 API 实例

CHCP 元数据检索接口 API 实例

基于操作的查询请求的主体是由 XML 或 format.xml JSON 请求体基于操作的查询。

1）基于操作的 XML 请求主体查询

包含一个 query request 出入，除非请求所有可以获得信息的操作入口,所有其他条目都是可选的。请求主体有如下格式。每个条目等级级别可以指定任何顺序：

<queryRequest>

<operation>

<count>number-of-results</count>

<lastResult>

<urlName>object-url</urlName>

<changeTimeMilliseconds>change-time-in-milliseconds.index

</changeTimeMilliseconds>

<version>version-id</version>

</lastResult>

<objectProperties>comma-separated-list-of-properties

</objectProperties>

<systemMetadata>

<changeTime>

<start>start-time-in-milliseconds</start>

<end>end-time-in-milliseconds</end>

</changeTime>

<directories>

<directory>directory-path</directory>

...

</directories>

<indexable>(true|false)</indexable>

<namespaces>

<namespace>namespace-name.tenant-name</namespace>

...

</namespaces>

<replicationCollision>(true|false)</replicationCollision>

<transactions>

<transaction>operation-type</transaction>

...

</transactions>

</systemMetadata>

<verbose>(true|false)</verbose>

</operation>

</queryRequest>

2）基于操作的查询 JSON 请求

JSON 请求体的操作为基础的查询必须包含一个未命名的顶级条目，除非请求所有可用信息。所有其他条目都是可选的。JSON 请求体具有以下格式。每个条目层次级别可以在任何顺序：

{

"operation": {

"count":"number-of-results",

"lastResult": {

"urlName":"object-url",

"changeTimeMilliseconds":"change-time-in-milliseconds.index",

"version":version-id

},

"objectProperties":"comma-separated-list-of-properties",

"systemMetadata": {

"changeTime": {

"start":start-time-in-milliseconds,

"end":end-time-in-milliseconds

},

"directories": {

"directory":["directory-path",...]

},

"indexable":"(true|false)",

"namespaces": {

"namespace":["namespace-name.tenant-name",...]

},

"replicationCollision":"(true|false)",

"transactions": {

"transaction":["operation-type",...]

```
}
},
"verbose":"(true|false)"
}
}
```

3）客户化元数据查询示例

这里是一个元数据查询请求实例，该请求将检索元数据的所有对象：

•首先是是由租户拥有的桶

•同时有自定义元数据，它包含一个名为部门的元素查询 JSON 格式使用一个 XML 请求和请求的结果。此外，在结果集对象的基本信息，这请求返回每个对象的内容和保留设置结果集。该请求还指定了结果集中的对象被列为以改变时间为基础的反向时间顺序。

备注：查询文件是 XML 格式的 accounting.xml

```
<queryRequest>
<object>
<query>customMetadataContent:
"department.Accounting.department"
</query>
<objectProperties>shred,retention</objectProperties>
<sort>changeTimeMilliseconds+desc</sort>
</object>
</queryRequest>
```

【命令行方式】

利用 Curl 工具提交命令

```
curl -k -H "Authorization: CHCP bXl1c2Vy:3f3c6784e97531774380db177774ac8d"
-H "Content-Type: application/xml" -H "Accept: application/json"
-d @Accounting.xml "https://europe.CHCP.example.com/query?prettyprint"
```

【Python 方式】

```
import pycurl
import os
curl = pycurl.Curl()
# Set the URL, command, and headers
curl.setopt(pycurl.URL, "https://europe.CHCP.example.com/" +
"query?prettyprint")
curl.setopt(pycurl.SSL_VERIFYPEER, 0)
curl.setopt(pycurl.SSL_VERIFYHOST, 0)
curl.setopt(pycurl.POST, 1)
```

```
curl.setopt(pycurl.HTTPHEADER,
["Authorization: CHCP bXl1c2Vy:3f3c6784e97531774380db177774ac8d",
"Content-Type: application/xml", "Accept: application/json"])
# Set the request body from an XML file
filehandle = open("Accounting.xml", 'rb')
curl.setopt(pycurl.UPLOAD, 1)
curl.setopt(pycurl.CUSTOMREQUEST, "POST")
curl.setopt(pycurl.INFILESIZE,
os.path.getsize("Accounting.xml"))
curl.setopt(pycurl.READFUNCTION, filehandle.read)
curl.perform()
print curl.getinfo(pycurl.RESPONSE_CODE)
curl.close()
filehandle.close()
```

【请求报文】

```
POST /query?prettyprint HTTP/1.1
Host: europe.CHCP.example.com
Authorization: CHCP bXl1c2Vy:3f3c6784e97531774380db177774ac8d
Content-Type: application/xml
Accept: application/json
Content-Length: 192
```

【返回结果】

```
HTTP/1.1 200 OK
Server: CHCP V7.0.0.16
Transfer-Encoding: chunked
JSON response body
To limit the example size, the JSON below shows only one object in the
result set.
{"queryResult:
{"query":
{"expression":"customMetadataContent:
"department.Accounting.department""},
"resultSet":[
{"version":84689494804123,
"operation":"CREATED",
"urlName":"https://finance.europe.CHCP.example.com/rest/presentations/
```

Q1_2012.ppt",
"changeTimeMilliseconds":"1334244924615.00",
"retention":0,
"shred":false},
.
.
.
],
"status":{
"message":"",
"results":12,
"code":"COMPLETE"}
}
}

Custom metadata file for the Q1_2012.ppt object

<?xml version="1.0">
<presentation>
<presentedBy>Lee Green</presentedBy>
<department>Accounting</department>
<slides>23</slides>
<date>04-01-2012</date>
</presentation>

总之，通过客户化数据查询，可以在跨系统的对象化存储数据中，迅速找到用户希望的信息并快速应用。

8.4 CHCP 数据集成及分析功能 - CHCI

ChangHong Content Intelligence（CHCI）软件作为 CHCP 分布式文件存储的增值软件功能，可帮助企业将多模态数据转换为宝贵业务信息。Content Intelligence 将这些数据汇总在一起，创建了一个集中的信息中心，帮助您的员工快速探索、发现并查找可行的业务洞察。

ChangHong Content Intelligence 可对 ChangHong 全系列存储系统和第三方应用程序（ISV）的数据进行自动提取、分类、丰富和归类，无论这些数据位于本地还是云端，也无论是否在异构数据间（内部和外部）。这种方法显著缩短了搜索所需数据或者创建已存在的数据的时间。此外，Content Intelligence 解决方案：

连接、转换、丰富和处理

CHCI 可对 ChangHong 全系列存储系统和第三方应用程序（ISV）的数据进行自动提取、分类、丰富和归类，无论这些数据位于本地还是云端，也无论是否在异构数据间（内部和外部）。这种方法显著缩短了搜索所需数据或者创建已存在的数据的时间。此外，CHCI 解决方案：

- 根据自动分类和归类进行数据探索指 导。
- 通过统一不同地点和数据类型的数据 接入而立即洞察您的数据。
- 通过将数据转换为宝贵的业务信息而 获得重要洞察。
- 通过精细接入控制和安全系统集成而 管理对敏感数据的接入。
- 采用个性化和自助式特性及用户体验，在正确的数据为正确的人员提供正确的数据。

创新、智能的数据理解方式

CHCI 是一种数据建议解决方案，秉承了开放核心架构的丰富特性、伸缩能力和可扩展性。

集成、封装和扩展开源保证了 CHCI 拥有成熟且广泛采用的代码库，而且该代码库已经在多个行业、垂直市场和多个应用上使用。因此，我们的解决方案对极为活跃的开发人员社区很有吸引力。它们可以利用该产品的可扩展性，将其与您的其他服务集成，或者根据用户需求而创建新的解决方案或用户体验。

CHCI 充分利用数据的配置，并通过查询服务提供索引，进而从企业数据中产生建议。这些建议基于用户的请求。建议可能来自解决方案的第一个适用的用例 – 搜索，或通过用户对 CHCI 和现有应用集成或者用于创建客户应用的体验而得出。

如何运行

如下图所示，CHCI 能够连接并索引 ChangHong Content Platform（CHCP）、ChangHong VSP 系列存储系统和云数据库中驻留的数据。一旦连接了数据库，CHCI 就可以支持多种处理工作流，并允许在数据处理过程中接入 20 多个数据分析、提取、转换和丰富阶段，以供根据具体条件而使用。在数据处理完成后，结果可以而存储在多个具有接入控制措施的地点，以确保仅授权用户能够接入特定数据。

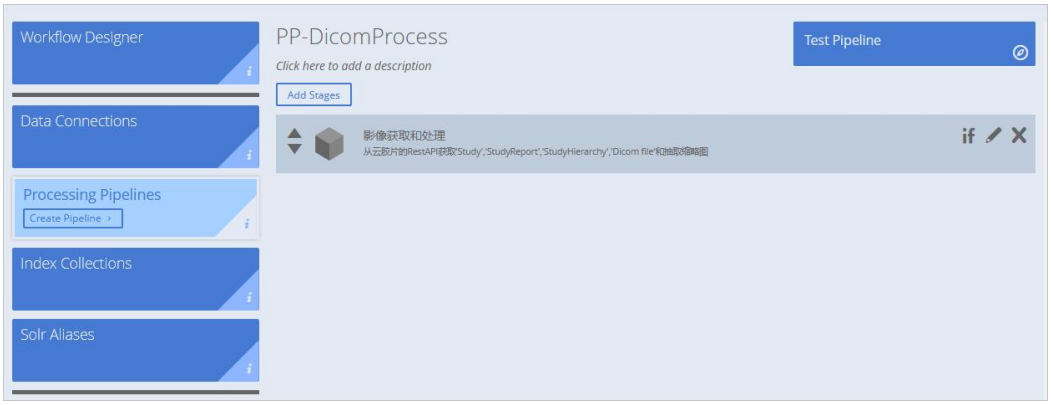
图：CHCI 软件架构

ChangHong Content Intelligence 提供了专门的接口，用于通过以下方式执行必要的解决方案配置、内容分析和转换、个性化结果访

问等功能：

以 PACS 系统为例：

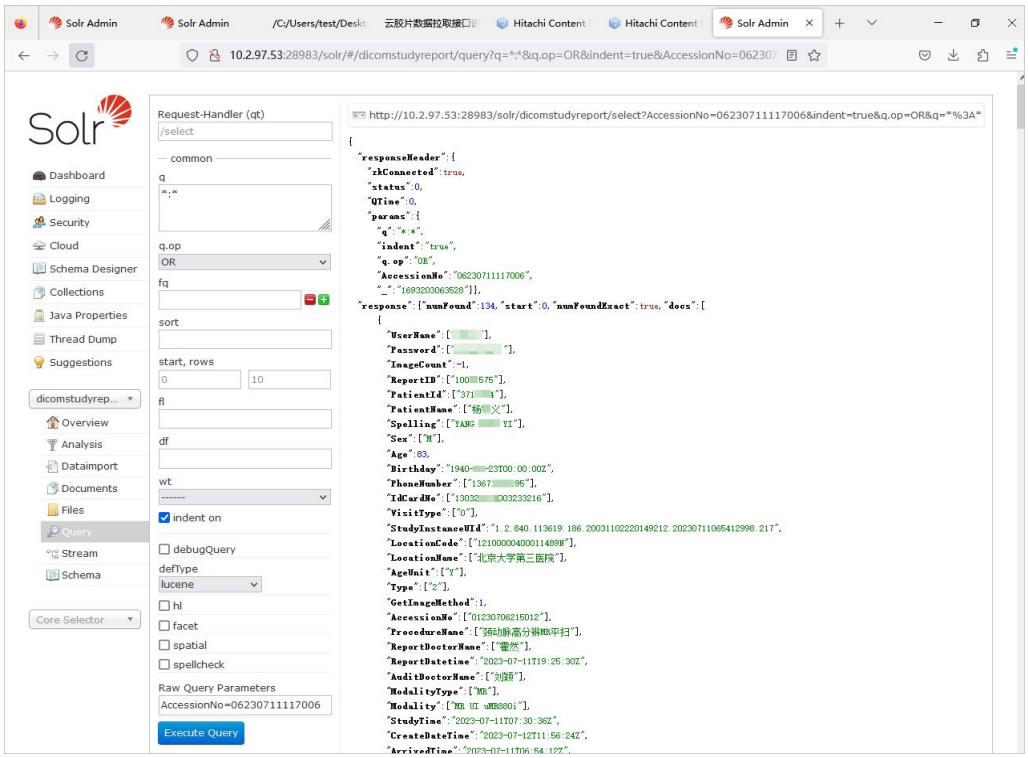
1 – 创建影像文件处理工作流



2 - 数据采集

3 – 建立索引库

4 – 检索查询


Solr Admin
Solr Admin
/C:/Users/test/Deskt
云胶片数据拉取接口
Hitachi Content
Hitachi Content
Solr Admin
10.2.97.53:28983/solr/#/dicomstudyreport/query?q=*:*&q.op=OR&indent=true&AccessionNo=062307
Solr
Dashboard
Logging
Security
Cloud
Schema Designer
Collections
Java Properties
Thread Dump
Suggestions
dicomstudyrep...
Overview
Analysis
Dataimport
Documents
Files
Query
Stream
Schema
Core Selector
Request-Handler (qt)
/select
common
q
:
q.op
OR
fq
sort
start, rows
0
10
fl
df
wt
indent on
debugQuery
defType
lucene
hl
facet
spatial
spellcheck
Raw Query Parameters
AccessionNo=06230711117006
Execute Query
http://10.2.97.53:28983/solr/dicomstudyreport/select?AccessionNo=06230711117006&indent=true&q.op=OR&q=*%3A*
{
"responseHeader":{
"zkConnected":true,
"status":0,
"QTime":0,
"params":{
"q":"*:*",
"indent":"true",
"q.op":"OR",
"AccessionNo":"06230711117006",
"_":"1693203063528"}},
"response":{"numFound":134,"start":0,"numFoundExact":true,"docs":[
{
"UserName":[
"Password":[
"ImageCount":-1,
"ReportID":["100575"],
"PatientId":["371
"PatientName":["杨
"Spelling":["YANG
"Sex":["M"],
"Age":83,
"Birthday":"1940-
"PhoneNumber":["1367
"IdCardNo":["13032
"VisitType":["0"],
"StudyInstanceUId":"1.2.840.113619.186.20031102220149212.20230711065412998.217",
"LocationCode":["121000004000114898"],
"LocationName":["北京大学第三医院"],
"AgeUnit":["Y"],
"Type":["2"],
"GetImageMethod":1,
"AccessionNo":["012307062150012"],
"ProcedureName":["颈动脉高分辨MR平扫"],
"ReportDoctorName":["霍然"],
"ReportDatetime":"2023-07-11T19:25:30Z",
"AuditDoctorName":["刘颖"],
"ModalityType":["MR"],
"Modality":["MR UI uMR880i"],
"StudyTime":"2023-07-11T07:30:36Z",
"CreateDateTime":"2023-07-12T11:56:24Z",

九、总结

CHCP 是遵循 POSIX 规范的高性能集群式并行文件系统，它是从头开始构建的，可在基于 NVMe 的存储上以原生方式运行。该文件系统通过高性能网络（以太网或 InfiniBand）充分利用网络链路，以获得最佳性能。对于需要对多个客户端提供高 I/O 和高并发性的性能密集型应用，它是一种理想解决方案。CHCP 在生命科学、金融分析、基于 GPU 的 DL 和 AI 应用、EDA、HPC 以及其他依赖并行文件系统的大带宽和 I/O 密集型应用中广泛部署。与传统解决方案相比，CHCP 降低了存储成本和复杂性，而且需要更少的硬件资源。它还完全支持传统协议，如 NFS 和 SMB，并提供了一组丰富的企业级特性。

CHCP 通过提供一个快速、高效、有弹性的分布式并行文件系统而解决常见的 IT 存储问题。该文件系统是云原生的，并提供所有闪存阵列所具备的性能、简单的文件存储和云的扩展性。CHCP 的易用性和类似云的体验包括快速资源调配，以缩短部署新工作负载的时间，另外还包括弹性扩展、弹性、高性能和成本效益等。