宝德 PowerLeader

宝德 PLStor

“芯”系列分布式存储产品技术白皮书

技术白皮书

白皮书版本 V3.2

发布日期：2023/11/9

目 录

1 概述

信息爆炸时代中人们可以获取的数据成指数倍的增长，传统的单机文件系统单纯通过增加硬盘个数来扩展计算机文件系统的存储容量的方式，在容量大小、容量增长速度、数据备份、数据安全等方面的表现都差强人意。为应对不同的存储需求，催生出不同的存储模式：

- 集中式存储：文件的元数据（描述数据的数据，如文件位置、大小等等）、数据信息存放在一起，前端 NFS 挂载后端 SAN、NAS 模式。此类传统存储模式扩展困难，难以支持超大数据量存储（PB 级）。
- 非对称分布式存储：单一元数据服务节点，文件元数据、数据分离存储，例如 lustre、moosefs 等等，独立元数据服务带来的问题：单点故障，可以采用 heart beat 等方式消除单点访问故障，但单点访问的性能瓶颈无法消除。
- 全对称分布式存储：全对称、去中心化的分布式架构，用一致性哈希 consistent hash 算法（DHT 的一种实现）来定位文件在存储节点中的位置，从而取消了元数据服务的角色，系统中只有存储节点的角色，不区分元数据和数据块。但在节点扩容、故障场景下，对一致性哈希算法要求有极高的效率以及均衡性、一致性。

PLStor D 系列采用的是全对称、去中心化的分布式架构，但未采用 DHT 来定位文件存储节点的位置，PLStor D 系列系统内每个节点都能提供元数据服务（MDS）、数据服务（DS）以及外部访问的接口服务（CA），无独立元数据服务节点，消除性能瓶颈，不存在单点故障，在节点扩容、故障场景下都能无缝平滑切换，业务无感知。系统能够给应用服务器提供统一的文件系统空间，满足多台应用服务器之间共享数据的需求。非分布式集群的设备一般使用双控或者多控节点提供服务，每个节点支持特定的业务负载，当容量不够时通过扩展硬盘框的方式增加存储容量。这种方式并不完美：首先业务和节点的绑定，意味着一个业务及其关联的文件系统只在一个节点上工作，容易造成系统

整体的负载不均；其次，这种系统本质上是 Scale-up 的扩容方式，追求单机性能，无法做到系统性能随容量的增加线性增加。

作为 PLStor D 系列系统的软件基础，PLStor D 系列采用全 Active 的 Share Nothing 方式，系统的数据和管理数据（元数据）平均分布在各个节点上，避免了系统资源争用，消除了系统瓶颈；即使出现整节点故障，系统也能够自动识别故障节点，自动重构故障节点涉及的数据和元数据，使故障对业务透明，完全不影响业务连续性。整系统采用全互联全冗余的组网机制，全对称分布式集群设计，实现存储系统节点的全局统一命名空间，从而允许系统中任何节点并发访问整系统的任何文件；并且支持文件内的细粒度的全局锁，提供从多个节点并发访问相同文件的不同区域，实现高并发高性能读写。

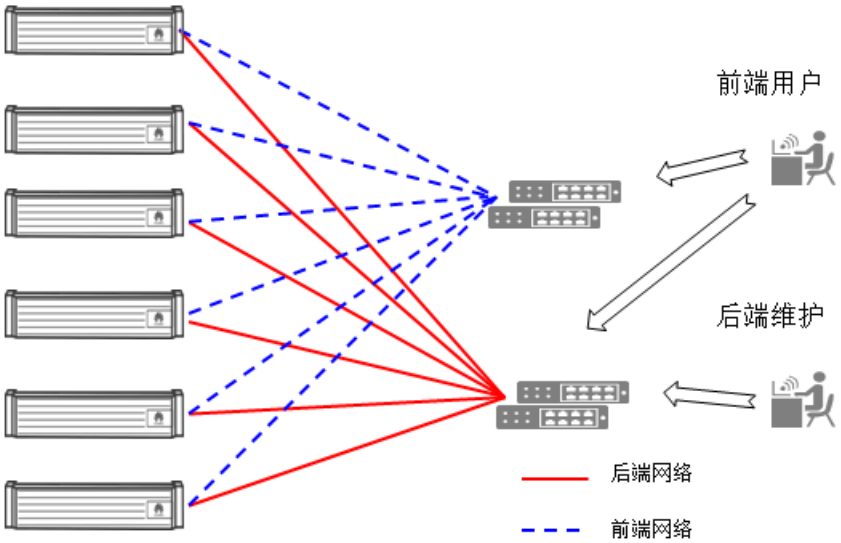
2 硬件、软件与网络

2.1 产品逻辑结构

PLStor D 系列系统采用全对称架构，对相同类型(硬件节点系列参见后文描述)的节点而言，节点内部的软硬件配置完全对等，这样的设计使得用户首次购买或者扩容时，不需要考虑独立元数据服务器或独立网管服务器等问题，只需根据实际需求计算出需要的节点数即可，最小 3 台起配。

PLStor D 系列系统由交换设备和 PLStor D 系列硬件节点组成，不需要额外的设备。产品结构如图 2-1 所示。

图2-1 PLStor D 系列产品结构



PLStor D 系列针对不同的应用需求提供不同系列的硬件节点类型。不同类型节点可以单独使用，也可以混合部署以达到整体最优，混合部署时每种节点最少配置为 3 台。通过划分不同“节点池”将

不同的硬件统一在单一的文件系统下，满足客户同时多层次的容量和性能级别的需求，并通过分级存储特性适应业务数据在不同层级之间流动。

PLStor D 系列所有的存储节点硬件配置均通过技术手段实现了数据掉电保护，使得数据即使在缓存中也可以得到持久化保护，同时，通过 RDMA 技术，有效减少了网络传输过程中内存拷贝次数。这样在不降低可靠性的前提下，进一步提升整系统响应速度。

PLStor D 系列系统可分为硬件平台和软件系统，硬件平台包括网络设备和物理存储节点。软件系统主要为 PLStor D 系列、管理系统和多种 Info 系列增值特性，PLStor D 系列提供对外统一的 NAS 共享服务，其中 PLStor D 系列基础软件包中包含 NFS、CIFS、FTP、NDMP 多种协议以及客户端负载均衡和性能加速软件。管理系统包含统一的系统资源管理、存储设备管理、网络设备管理（10GE 组网时）和系统统计报表，趋势分析，容量趋势预测，性能对比，诊断分析等功能。

PLStor D 系列的软件如下列表：

表2-1 PLStor D 系列软件列表

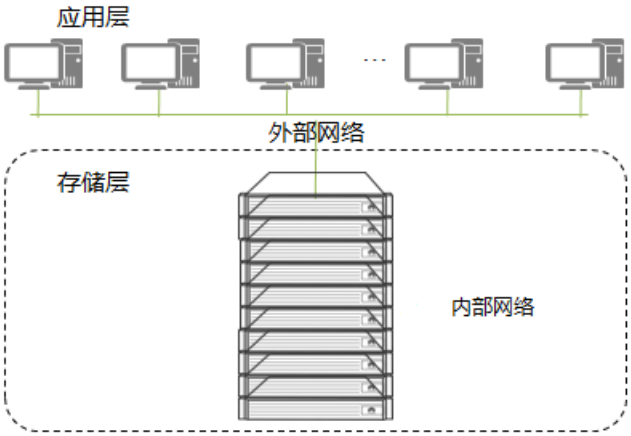
名称		功能
PLStor D 系列		分布式文件系统软件
Device Manager		设备管理软件
NAS 存储增值特性	InfoEqualizer	客户端连接负载均衡
	InfoTurbo	性能加速
	InfoAllocator	配额管理
	InfoTier	自动分级存储
	InfoLocker	WORM
	InfoStamper	快照
	InfoReplicator	远程复制
	InfoScanner	防病毒
	InfoRevive	视频图像修复
	InfoMigrator	文件迁移
	InfoContainer	虚拟机灌装特性
	InfoStreamDS	视频监控宝德自强流媒体软件

2.2 组网概述

PLStor D 系列前后端网络物理隔离，业务网络与管理网络分别使用不同的网络平面，组网结构如下图 2-3 所示，包括：

- 前端业务网络用于 PLStor D 系列与用户网络对接。
- 后端存储网络用于 PLStor D 系列内部节点间互联。

图2-2 PLStor D 系列组网结构



在 PLStor D 系列系统中，集群后端网络可以支持 10GE、25GE、IB 连接，前端网络支持 GE、10GE、25GE、IB 连接，以满足不同场景用户的组网需求。无论哪种组网，PLStor D 系列系统的所有节点网络都是冗余的，任何单一网口故障或者单一交换机故障均不影响系统使用。

PLStor D 系列系统前端和后端可以分别使用不同的物理网卡以达到从网络上相互隔离的目的，并且前端网络可以根据用户现有网络状况选择 GE、10GE、25GE、IB 连接。通过 PLStor D 系列存储设备所提供的 IPMI 网口可以访问设备的管理界面。

PLStor D 系列节点支持的组网类型包括：

- 前 2*10GE 后 2*10GE 组网
- 前 2*25GE 后 2*25GE 组网
- 前 2*10GE 后 2*25GE 组网
- 前 2*GE 后 2*10GE 组网
- 前 4*GE 后 2*10GE 组网
- 前 2*IB 后 2*IB 组网
- 前 2*10GE 后 2*IB 组网

- 前 2*25GE 后 2*IB 组网

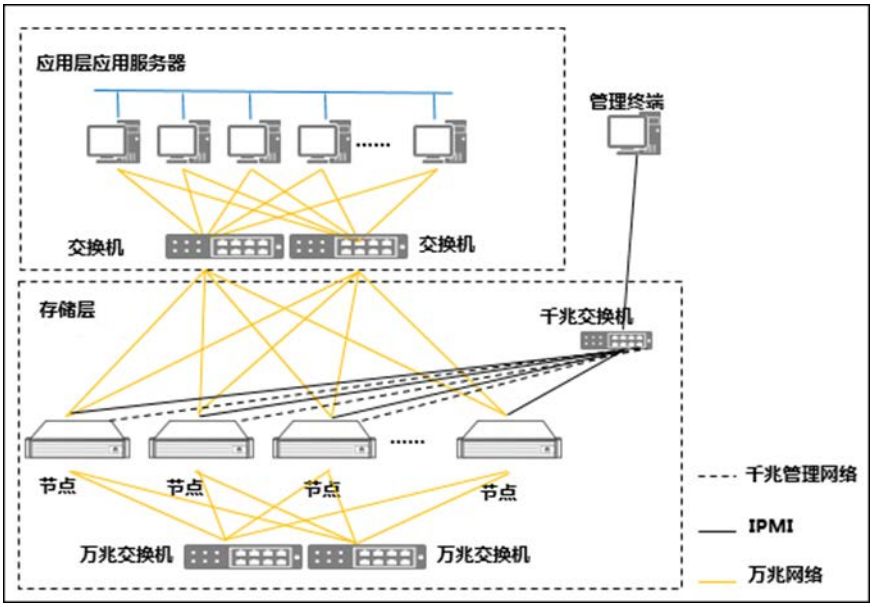
图2-3 组网部署

- 系统支持多种类型节点混合部署，混合部署时相同类型相同配置的节点最少配置为 3 台；
- 系统仅部署 NAS 存储时，最少需要部署 3 台节点；

2.2.1 Ethernet 组网描述（前后端网络均采用 Ethernet 组网）

前后端均采用 Ethernet 交换机组网的典型配置方案如图 2-5 所示。

图2-4 前端和后端网络均采用 Ethernet 交换机组网方案示意



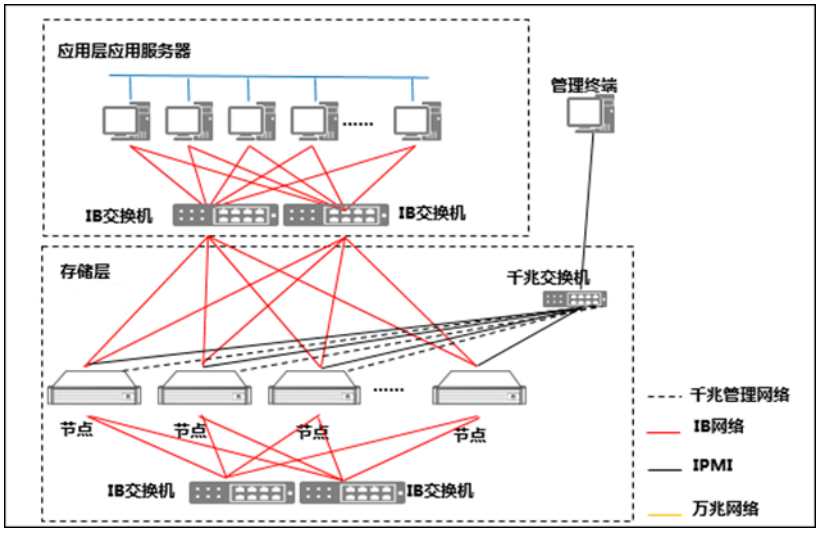
组网说明：

- 当 PLStor D 系列使用 Ethernet 组网时，前端网络对接用户 Ethernet 交换网，后端网络使用内部 Ethernet 交换机。前后端交换机冗余配置。
- GE 交换机通过网线连接管理网口和 IPMI 网口，仅用于管理维护。

2.2.2 IB 组网描述（前后端网络均采用 IB 组网）

前后端均采用 IB 交换机组网的典型配置方案如图 2-6 所示。

图2-5 前端和后端网络均采用 IB 交换机组网方案示意



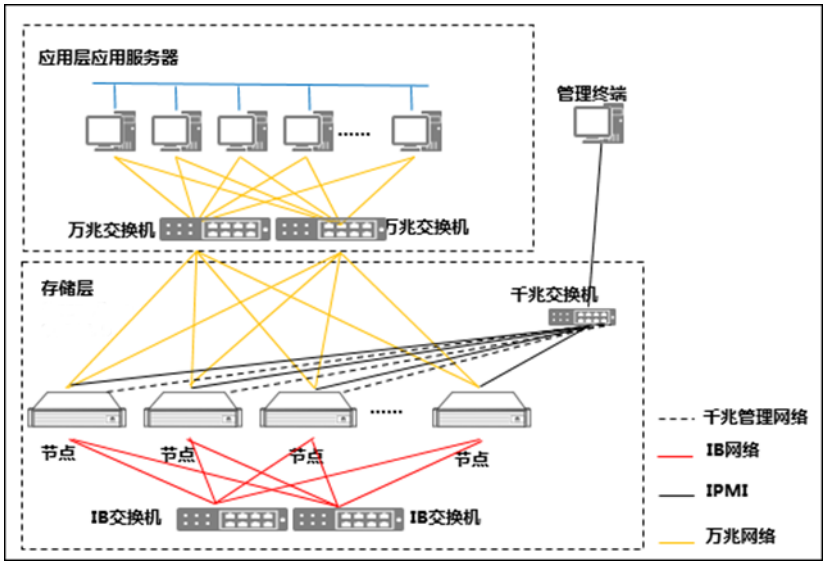
组网说明：

- 当 PLStor D 系列使用全 IB 组网时，前端网络对接用户 IB 交换网后端网络使用内部 IB 交换机。前后端交换机冗余配置。
- GE 交换机通过网线连接管理网口和 IPMI 网口，仅用于管理维护。

2.2.3 前端 Ethernet 后端 IB 组网描述（前端网络采用 Ethernet，后端网络采用 IB 组网）

前端网络采用 Ethernet 交换机，后端网络采用 IB 交换机组网的典型配置方案如图 2-7 所示。

图2-6 前端网络采用 Ethernet 交换机，后端网络采用 IB 交换机组网方案示意



组网说明：

- 前端网络使用 Ethernet 交换机，后端网络使用 IB 交换机。前后端交换机冗余配置。
- GE 交换机通过网线连接管理网口和 IPMI 网口，仅用于管理维护。

2.3 系统运行环境

PLStor D 系列通过 NFS 共享、CIFS 共享等方式为用户提供文件服务，对最终用户来说，PLStor D 系列就是一个文件服务器，用户通过该服务器存取文件。用户所在的环境可能是比较复杂的，在提供 NAS 服务时，需要配合如 AD 域、NIS 域、LDAP 等环境，PLStor D 系列均能够支持。PLStor D 系列提供对以上环境的支持，用户只需要进行相应的配置，即可将 PLStor D 系列系统在现有的域环境中运行起来供应用主机访问。

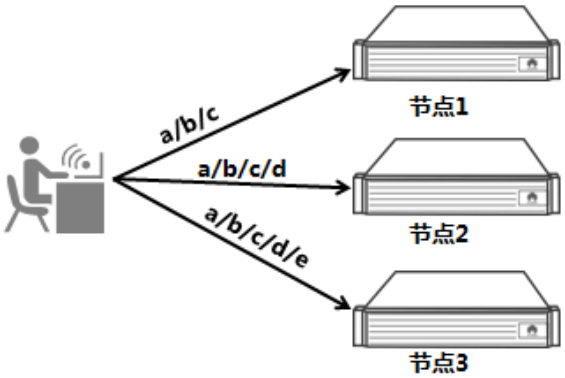
3 分布式文件系统架构

3.1 分布式文件系统架构概述

PLStor D 系列是 PLStor D 系列系统中的核心部件，将系统中所有节点的硬盘整合成一个统一的资源池，对外提供统一命名空间。同时对用户数据提供跨节点、跨机架、不同级别的数据冗余保护，可以兼顾高硬盘利用率和高可用的需求，避免了传统存储的烟囱式弊端。

单一的文件系统上，PLStor D 系列提供目录级的业务控制能力，可以基于目录配置数据的保护级别，配额控制，快照等特性，从而在单一文件系统上满足客户多层次差异化的需求。

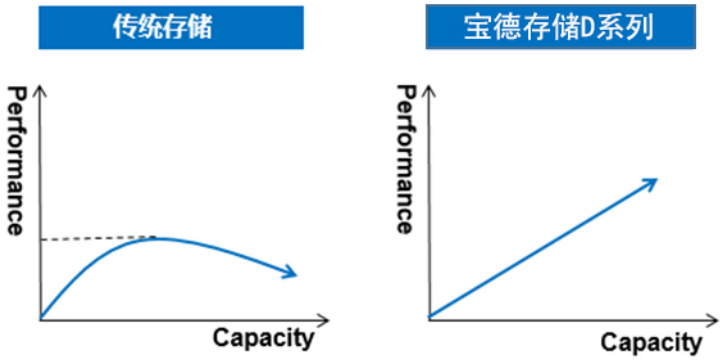
图3-1 统一命名空间功能示意图



上图中所示例的 PLStor D 系列系统由 3 个节点组成，这 3 个节点对用户都是透明的，用户并不会感知到是哪个节点在提供服务。假如用户访问不同的文件，实际上是由不同的节点在提供服务。

PLStor D 系列支持无缝横向扩展，系统支持 3 节点至 288 节点弹性无缝扩展，整个扩容过程业务无中断。由于采用了可灵活扩展的 Share Nothing 的全对称分布式架构，元数据和数据平均分散存储到所有节点上，不存在性能瓶颈，随着节点数的增加，存储容量和计算能力线性增加，最终给用户提供呈线性递增的吞吐及并发能力。并且天然支持了自动精简配置功能，根据应用实际所需要的容量分配，当该应用所产生的数据增长，分配的容量空间已不够时，系统会再次从后端存储池中补充分配一部分存储空间，使得存储资源得到充分利用。

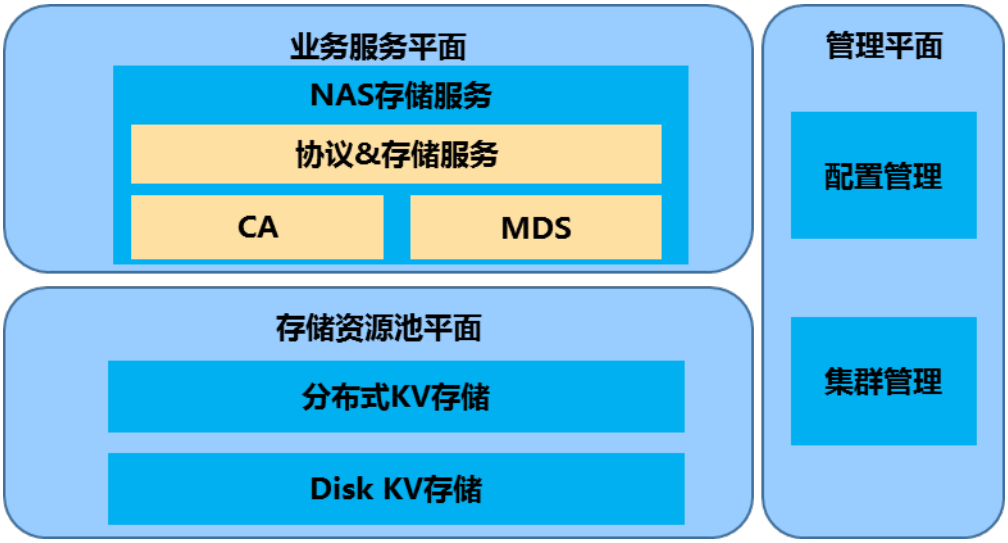
图3-2 容量和性能线性增长



PLStor D 系列对外提供 CIFS、NFS、FTP 接入功能，并提供统一命名空间，让用户业务轻松接入存储系统。访问时支持集群节点间负载均衡及管理功能。在这些功能特性基础上，由于全对称的架构设计使得 PLStor D 系列每个节点都可以对外提供全局的业务访问，且任何单节点故障时可自动切换。

PLStor D 系列逻辑层次如下图所示：

图3-3 PLStor D 系列逻辑架构视图



PLStor D 系列从架构层面分为三层：业务服务平面，存储资源池平面和管理平面。

- 业务服务平面： PLStor D 系列业务服务平面对外提供分布式文件系统服务。

  分布式文件系统服务对外提供 NAS 协议访问、文件系统相关的增值特性；对外提供统一命名空间、提供存储协议访问，支撑包括 NDMP、远程服务 FTP 等增值服务。

- 存储资源池平面：存储资源池平面负责管理存储集群节点的所有物理存储的分配和管理，数据统一存储在统一的存储资源池。存储资源池通过分布式技术，为业务服务平面提供强一致、跨节点可靠性的 KV 存储服务。存储资源池平面还负责节点间的负载均衡和数据自动修复能力。负载均衡使得在系统扩展的同时，能够充分利用新节点的 CPU 处理能力，内存缓存能力和磁盘能力，使得整个系统的吞吐量和 IOPS 伴随节点的扩容而线性增长。

  存储资源池平面可以为业务服务平面的分布式文件系统服务提供数据读写访问；这使得 PLStor D 系列可以在同一套物理集群上提供 NAS 服务，共享物理存储空间。

- 管理平面：管理平面对外为用户提供可视化图形界面以及命令行管理工具，对内提供集群状态管理以及配置数据的管理功能。提供硬件资源配置，提供性能监控，存储系统参数配置，用户管理，硬件节点状态管理，软件升级等功能。

3.1.1 分布式文件系统服务

分布式文件系统服务层主要由协议＆增值服务模块、CA 模块和 MDS 模块构成；

- 协议＆增值服务模块负责 NAS 协议的语义解析和执行；

- CA 模块为协议＆增值服务模块提供标准的文件系统读写访问等接口；
- MDS 模块负责文件系统的元数据，负责文件系统命名空间的目录树管理。

PLStor D 系列支持最大 140PB 全局命名空间，用户不用管理多个命名空间，从而减轻管理复杂度。消除多个命名空间，也消除了多个命名空间带来的数据孤岛。

分布式文件系统服务层分布在集群的每个节点上，采用全对称的分布式技术，提供全局统一命名空间，允许从系统任何节点接入访问整系统的任何文件；并且支持文件内的细粒度的全局锁，提供从多个节点并发访问相同文件的不同区域，实现高并发读写，最终达到高性能访问系统。

3.1.2 存储资源池平面

存储资源池平面负责管理存储集群节点的所有物理存储的分配和管理，通过切分节点池的方法将集群节点划分为多个节点池。

PLStor D 系列的 InfoProtector 技术，提供 N＋M 的保护能力，N 代表数据被切割到多少个节点上，M 代表能够抵御的同时故障的节点和磁盘的数量；用户可以配置 M 值，N 值由系统自己根据集群大小来确定，伴随着集群节点数量的增长，N 则不断增长，从而在数据保护能力不下降的情况下提供更大的存储利用率。当数据被配置为＋M 保护时，只有同一个节点池内的大于等于 M＋1 个节点故障或者大于等于 M＋1 个硬盘故障时，才会造成数据损毁。而将整个存储集群节点划分为多个节点池的方法，使得数据损坏的概率得到极大的降低。这种保护方式使得文件能够散列到整个集群中，从而提供更高的数据并行访问能力和重建并行能力，在磁盘或者节点故障时，系统能够发现哪些文件的哪些部分受到影响，并让多个节点参与重建过程，这样参与重建的磁盘数量和 CPU 数量远远超越传统 RAID 技术，使得故障重构时间得到飞跃的进步。

同时，PLStor D 系列支持针对不同的应用需求提供不同的硬件节点类型，支持混合使用不同的硬件类型，通过划分不同节点池的概念将不同的硬件统一在单一的文件系统下，满足客户同时多层次的容量和性能级别的需求，并通过分级存储特性适应业务数据在不同层级之间流动。

3.1.3 管理平面

随着越来越多的数据和越来越大规模的设备需要管理，减少管理复杂性成为关键点。PLStor D 系列支持一站式的全系统管理，支持在线的扩容、升级等维护活动让客户轻松掌控整系统。PLStor D 系列不需要单独的管理服务器，节省了硬件成本开支。PLStor D 系列的管理服务支持通过 SNMP 与用户管理服务对接。

PLStor D 系列提供 GUI 和 CLI 两种管理界面。用户使用管理界面可以完成状态、容量、资源使用率、告警等信息查询，也可以完成系统各种配置和操作。访问管理界面的用户分为超级管理员、管理员和只读用户，满足不同级别用户访问的需要。GUI 界面集成了用户常用的各种功能，CLI 除了支持 GUI 具备的功能外，面向高级管理维护人员的高级功能及非常用系统配置功能也通过 CLI 提供。

集群管理子系统设计实现了一致性选举算法，使节点状态的变化在整系统所有节点上是统一的，为了保证监控元数据集群的可靠性，系统在所有的节点上启动监控进程，这些监控进程之间组成一个集群，负责监控和同步节点和软件模块的状态，当系统中添加节点或节点/软件模块故障的时候，会通过事件的方式通知关注集群状态变化的子系统或模块。

配置管理集群子系统负责整个系统的业务管理、业务监控、业务状态及设备状态监控等功能。系统中正常情况下只有一台节点对外提供服务，当该设备故障后，管理服务可以自动切换到其它正常的设备上。管理服务在切换的过程中，对于客户端透明，即管理服务切换成功后，对外提供服务的 IP 地址仍为原来的 IP 地址。

3.1.4 网络平面

PLStor D 系列具有安全的物理组网结构，根据业务类型可划分为管理网络、BMC 网络、业务网络和存储网络，可以支持租户通过设置 VLAN 进行逻辑隔离，也可以支持独立网口和独立交换机的物理隔离，保护系统运行的安全。

3.2 元数据管理

PLStor D 系列支持超大目录，单目录可达百万数量级，大目录访问和普通目录访问无响应差别。

PLStor D 系列的元数据管理架构采用业界先进的动态子树方式，如下图所示。元数据和数据一样都是使用节点间冗余方式，不同在于，元数据使用的是节点间 mirror 方式，每一份都是一个独立完整的拷贝。默认情况下，PLStor D 系列的元数据比数据保护级别高一级。

PLStor D 系列采用统一命名空间，整个文件系统的目录文件层次结构是一个树型结构，而系统集群是有很多平等的物理节点构成，所以，需要将整个文件系统这个大的树进行切割，使得每个物理节点上的 MDS（元数据管理）模块分别管理不同的子树。

整系统目录树结构可以划分为若干子树，每个子树属于一个 MDS，一个 MDS 可以包含若干个子树。

子树的分裂依赖文件夹的分裂，文件夹分裂的原则有 2 个：

- 原则 1：按照热度进行分裂

每次元数据访问时，都会在内存中根据访问类型来增加文件夹的加权访问热度，并且热度会根据时间延长而进行衰减。当文件夹热度超过阈值时，文件夹进行分裂。

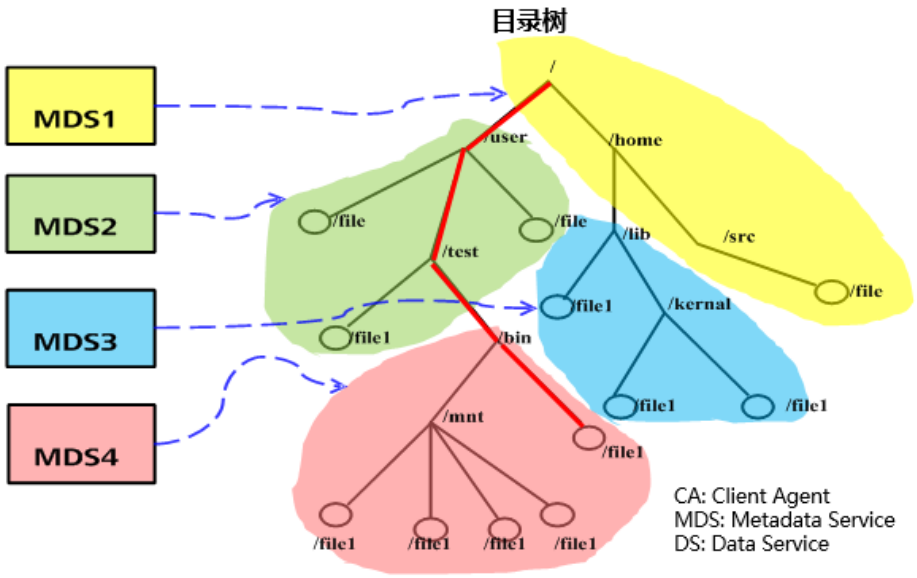
- 原则 2：单文件夹下文件数过多时，文件夹分裂

满足任一条件时，文件夹会被分裂，记做 dir_frag（目录分片）。 当如上原则不满足的时候，文件夹会合并，避免过多文件夹碎片。

当分裂的文件夹为子树的根时，文件夹分裂即为子树分裂。分裂后的子树仍存放在同一个 MDS 上，定期进行负载均衡检测。发现负载不均衡的情况，会在 MDS 之间进行子树的迁移。

根据上面原则，频繁访问超大目录时，大目录会分裂成多个 dir_frag，对应分裂成多个子树，并会负载到多个 MDS 服务器上提供业务，不会产生元数据访问瓶颈。

图3-4 命名空间子树划分示意图



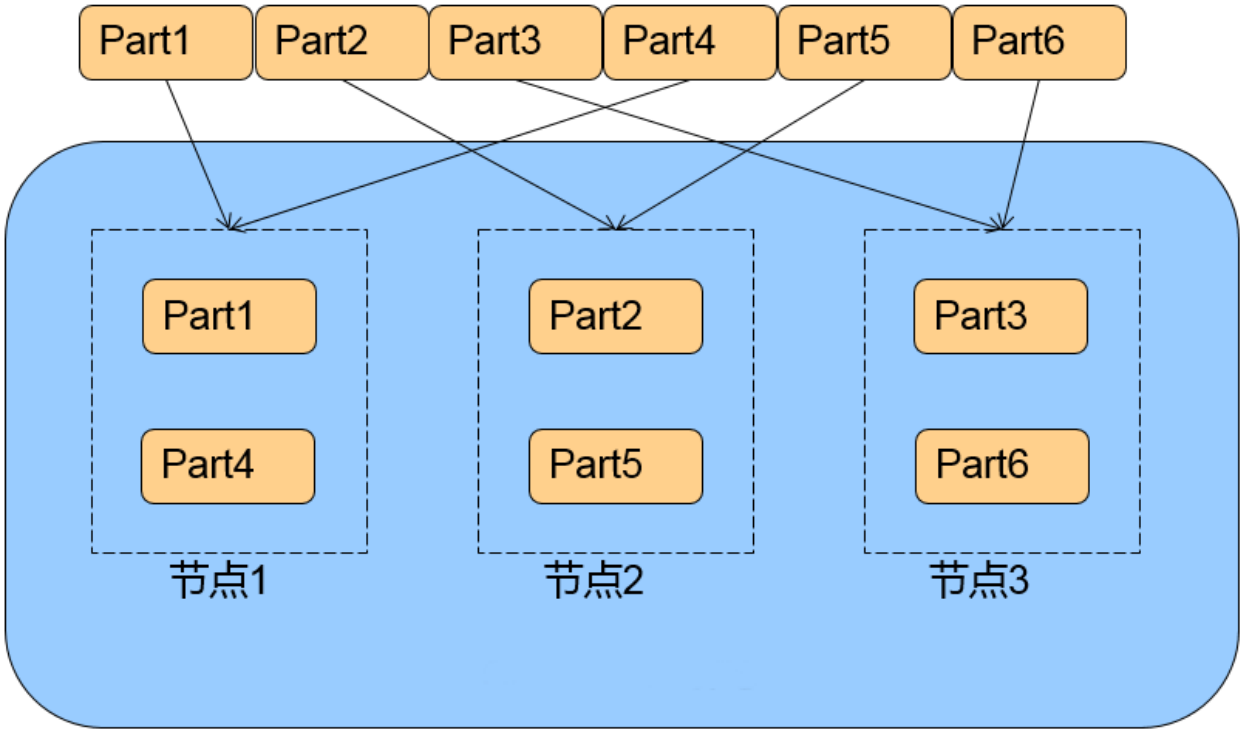
3.3 分布式数据可靠性技术

PLStor D 系列的 InfoProtector 技术提供了数据跨节点的保护能力，在多个硬盘或者节点故障时也能够继续提供服务，将数据放置到同一个节点池内不同节点的不同硬盘上，数据获得了跨节点的可靠性和故障快速重构的能力。

3.3.1 数据条带化

为实现数据保护和高性能读写，PLStor D 系列对数据进行按节点条带化处理，首先，创建新文件时文件系统会按照配置的保护级别挑选符合要求的节点，然后写数据时文件系统将用户的数据平均分布在各节点上，读数据时文件系统从所有节点并行读取。

图3-5 文件条带化示意图



上图所示例的 PLStor D 系列系统由 3 个节点组成，用户的数据平均分布在 3 个节点上。实际使用中用户的数据分布需要根据配置而定。

PLStor D 系列系统使用 Erasure Code(纠删码)方式存储数据，可以针对目录/文件配置不同的数据保护方式。不同的数据保护方式，是通过不同的数据条带化方式实现的。

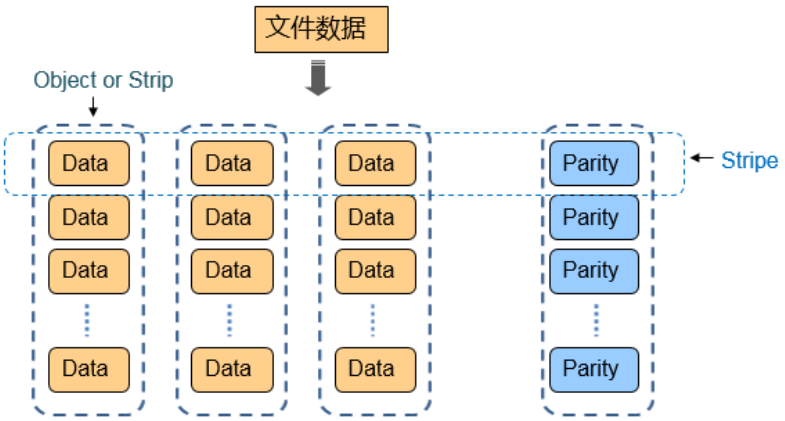
写入 PLStor D 系列系统的数据，会按照固定大小划分为一个条带（(NAS：512KB/256KB/128KB/32KB/16KB 可选；对象 512K)，可以按照目录配置冗余配比，切分为多个原数据条带，然后对每 N 个原数据条带，计算得到 M 个冗余数据条带，最终这 N+M 个条带组成一个分条，写入到系统中。当系统出现故障，丢失了其中的某些条带时，只要一个分条中丢失的条带数目不超过 M，就可进行正常的数据读写。通过数据重构算法，丢失的条带可从剩余条带中计算得到。在这种方式下，空间的利用率约为 N/(N+M)，数据的可靠性由 M 值的大小决定，M 越大可靠性越高。

3.3.2 集群对象存储系统

PLStor D 系列的分布式文件系统，是以底层的集群对象存储系统为基础的，文件系统的数据和元数据经过数据条带化后生成条带和分条，最终以对象的形式存储到硬盘中。以一个 3+1 方式保护的文件数据举例，如下图所示。

其中纵向的虚线框代表不同硬盘，横向虚线框代表一个数据分条(Stripe)，每个分条落在单个硬盘上的部分我们称之为对象或者条带(Strip)。

图3-6 条带与对象



在 PLStor D 系列内部存储资源池，所有数据存储实现为一种基于对象（此处的对象不同于对象存储服务）的分布式存储系统。PLStor D 系列的对象存储系统，是将系统中所有的设备格式化成对象存储设备后，通过网络连接组成的一个集群系统。

PLStor D 系列不间断的监视着系统内的节点、硬盘的状况。

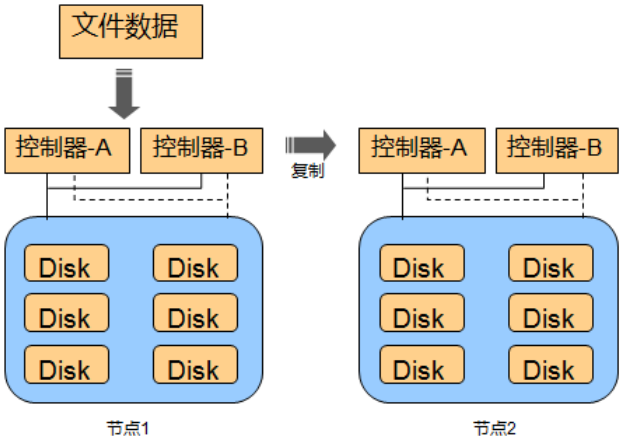
- 当有硬盘坏道时，能自动检测并在后台自动进行修复，在内存中重构对应坏道的数据并重新写入磁盘。

- 当有硬盘或者节点损坏时，能自动发现故障，并自动发起数据重构。这种重构只重构真正的数据，不会像传统 RAID 那样进行全盘重构，因此具有更高的重构效率。另外，在重构过程中，会选择不同的节点和硬盘作为重构目标，并发地执行重构过程，相对于传统 RAID 只能重构到一块热备盘上的方式，这种重构可以达到非常高的重构速度。

3.3.3 N+M 数据保护

相比于传统的 RAID 方式，PLStor D 系列在提供高可靠性的同时也能够提供更高的磁盘利用率。传统 RAID 把数据存放在一个 RAID 组内的不同硬盘上，当其中有硬盘损坏时，通过 RAID 重构，重构坏盘上的数据。

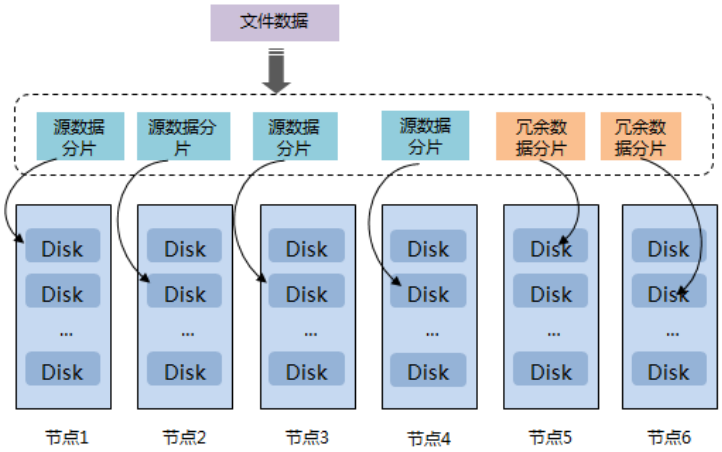
图3-7 传统 RAID 数据保护技术



这类存储系统常用的 RAID 方式有 RAID-0\1\5\6 等，其中可靠性最高的 RAID-6 最多只能支持 2 块硬盘同时发生故障。另外一方面，这类存储系统使用控制器执行 RAID 数据存储，为了预防控制器故障，它们通常使用双控制器的方式来保证服务的可用性，但当 2 个控制器同时发生故障时，还是会导致服务中断。虽然这类系统还可以通过在多个节点间进行同步/异步的数据复制，进一步提高系统可靠性，但这会导致硬盘利用率很低，让用户承担较高的 TCO（总体拥有成本）。

PLStor D 系列的数据保护技术，是建立在分布式、节点间冗余的基础上的。数据进入系统之后，首先被切分为 N 个数据条带，然后计算出 M 个冗余条带，并最终保存在 N+M 个不同的节点中。

图3-8 PLStor D 系列 N+M 数据保护技术

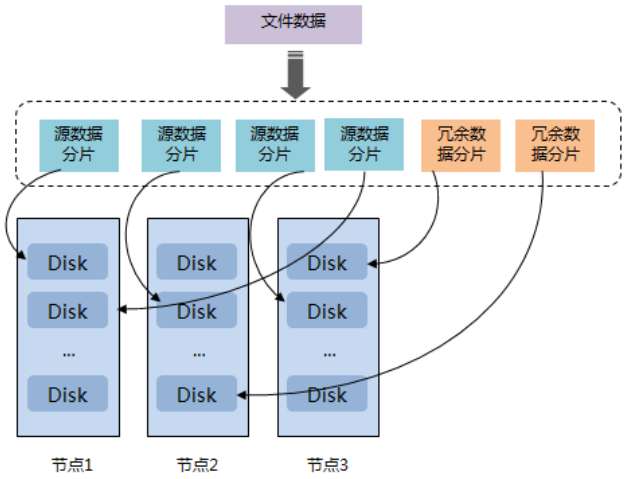


图示以4份数据切片2份冗余切片存储在6个节点上举例

由于同一条带的数据保存在不同节点中，所以 PLStor D 系列存储系统中的数据不仅能支持硬盘级的故障，而且能够支持节点级的故障，保证数据不丢失；单节点故障时，EC 可以自动重新配比，确保数据可靠性不降级。只要系统中同时故障的节点数不超过 M，系统就可以持续提供服务。通过数据重构过程，系统可以重构出损坏的数据，重构整系统的数据可靠性。

另外 PLStor D 系列存储系统还提供一种 N+M:B 的保护方式，可以支持 M 个硬盘故障或者 B 个节点故障。这种保护方式在节点数量少于 N+M 的小容量系统中非常有效。同时 PLStor D 系列存储系统支持在节点数量满足的条件下从 N+M:B 的保护方式升级为同等的 N+M 保护方式。

图3-9 PLStor D 系列 N+M:B 数据保护技术



图示以4份数据切片2份冗余切片存储在3个节点上举例

PLStor D 系列的数据保护方式与传统 RAID 相比，能达到类似于传统 RAID 在多节点数据复制的高可靠性，同时仍可保持 N/(N+M)的高硬盘利用率。另外，在 PLStor D 系列系统中，任意可用空间都可以作为“热备”空间使用，不需要像传统 RAID 那样预先划分独立的热备盘，因此可进一步提高存储利用率。

PLStor D 系列存储系统提供多种 N+M（或者 N+M:B）的冗余比配置，用户可根据业务需求在管理界面上进行配置。配置的范围可以是任意目录，对目录配置冗余后，目录下的文件都采用该冗余配比保存；用户甚至可以对目录与此目录下的子目录配置不同的冗余比。这意味着用户可以灵活多变地根据自己的实际需求来指定数据冗余，从而设置最适合的可靠性。

PLStor D 系列系统内的节点可划分为多个 Node Pool（节点池），每个节点池的节点最少为 3 个，最多为 20 个，在部署和扩容时可根据需要来划分节点池。

在实际配置中，PLStor D 系列提供智能配置，用户只需要指定其数据的可靠性（支持几个节点同时故障，或者支持几块硬盘同时故障），即只需对目录/文件设置相应的+M（或者+M:B）即可。PLStor D 系列系统会根据系统当前 Node Pool(节点池)的节点数量，自动选取最合适的冗余比。目前 PLStor D 系列系统支持的 M 为 1 到 4（对象存储 M 为 1 到 3，当配置为+M:B 时，B 可选为 1）。在不同的节点数目下，不同的配置对应的实际 N+M（或 N+M:B）如下表所示，其中括号内为存储利用率：

表3-1 PLStor D 系列冗余配比对照表

节点\ 配置	1	2	3	4	+2:1	+3:1
3	2+1 (66.66%)	4+2(:1) (66.66%)	6+3(:1) (66.66%)	6+4(:1) (60%)	4+2:1 (66.66%)	6+3:1 (66.66%)
4	3+1 (75%)	4+2(:1) (66.66%)	6+3(:1) (66.66%)	6+4(:1) (60%)	6+2:1 (75%)	8+3:1 (72.72%)
5	4+1(80%)	4+2(:1) (66.66%)	6+3(:1) (66.66%)	6+4(:1) (60%)	8+2:1 (80%)	12+3:1 (80%)
6	4+1(80%)	4+2(66.66%)	6+3(:1) (66.66%)	6+4(:1) (60%)	10+2:1 (83.33%)	14+3:1 (82.35%)
7	6+1(85.71%)	4+2(66.66%)	6+3(:1) (66.66%)	6+4(:1) (60%)	12+2:1 (85.71%)	16+3:1 (84.21%)
8	6+1(85.71%)	6+2(75%)	6+3(:1) (66.66%)	6+4(:1) (60%)	14+2:1 (87.50%)	16+3:1 (84.21%)

9	8+1(88.88%)	6+2(75%)	6+3(66.66%)	6+4(:1) (60%)	16+2:1 (88.88%)	16+3:1 (84.21%)
10	8+1(88.88%)	8+2(80%)	6+3(66.66%)	6+4 (60%)	16+2:1 (88.88%)	16+3:1 (84.21%)
11	10+1(90.90%)	8+2(80%)	8+3(72.72%)	6+4 (60%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
12	10+1(90.90%)	10+2(83.33%)	8+3(72.72%)	8+4 (66.66%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
13	12+1(92.30%)	10+2(83.33%)	10+3(76.92%)	8+4 (66.66%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
14	12+1(92.30%)	12+2(85.71%)	10+3(76.92%)	10+4 (71.42%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
15	14+1(93.33%)	12+2(85.71%)	12+3(80%)	10+4 (71.42%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
16	14+1(93.33%)	14+2(87.50%)	12+3(80%)	12+4 (75%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
17	16+1(94.11%)	14+2(87.50%)	14+3(82.35%)	12+4 (75%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
18	16+1(94.11%)	16+2(88.88%)	14+3(82.35%)	14+4 (77.77%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
19	16+1(94.11%)	16+2(88.88%)	16+3(84.21%)	14+4 (77.77%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)
20	16+1(94.11%)	16+2(88.88%)	16+3(84.21%)	16+4 (80%)	16+2:1 (88.88%)/18+2:1 (90%)	16+3:1 (84.21%)

3.3.4 多副本

传统的硬盘级 RAID 模式将数据存放于单节点内的不同硬盘，当整节点发生故障时，无法有效恢复数据。PLStor D 系列存储系统将数据在节点间进行多副本构建，有效避免数据丢失。多副本是通过将相同的数据在不同的节点上存储多份来实现数据保护的一种技术，支持三副本和两副本，推荐三副本。三副本的空间利用率为 33.3%，两副本的空间利用率 50%。

3.4 全局缓存

PLStor D 系列提供了可以全局访问的一致缓存，所有存储服务器上的内存空间在逻辑组成一个统一内存资源池，缓存在任何一个存储服务器上的数据，在后续其它任何存储服务器接收到访问该数据请求时都可以在全局缓存中命中，同时所有用户数据在整个集群系统中只缓存一份（校验数据不缓存）。

PLStor D 系列中的缓存容量随着节点增加而线性增长，随着节点数目的增加，全局缓存的容量也增加，更多的热点数据可以被命中，大大减少硬盘的 I/O 访问，满足各种应用场景下的高性能和低时延要求。

3.4.1 全局缓存组成要素

一级缓存

一级缓存是位于与协议服务对接的分布式文件系统客户端引擎(Client Agent)层，该客户端引擎代表用户访问文件系统，该层以文件数据作为缓存对象，以文件分条（Stripe）为缓存单位，一级缓存主要用于针对文件的访问模型预测后用于文件数据预取和加速热点文件分条的缓存。该级缓存是整系统全局共享的，即只要缓存在任意一个节点上的文件分条数据，其它任意节点再次收到该数据的访问请求后都可以从一级缓存中命中该数据。

通常在大规模分布文件系统中只有少量文件是热度比较高的文件，大部分都是冷数据。因此，缓存热点文件数据和对数据进行预取是充分发挥缓存的优势，降低后端存储硬盘访问的压力提高业务的响应速度的最有效方法。

二级缓存

二级缓存有 SSD 盘和数据块元数据及数据块缓存，该级缓存只用于缓存本节点所有硬盘上的热点数据。主要用于加速该节点上条带（Strip）或分条（Stripe）的访问速度，减少频繁访问的热点数据对硬盘的压力及加快数据块请求的响应速度。如：每块硬盘上的超级块、对象集到对象的描述符及关键对象描述符等数据。

数据掉电保护缓存

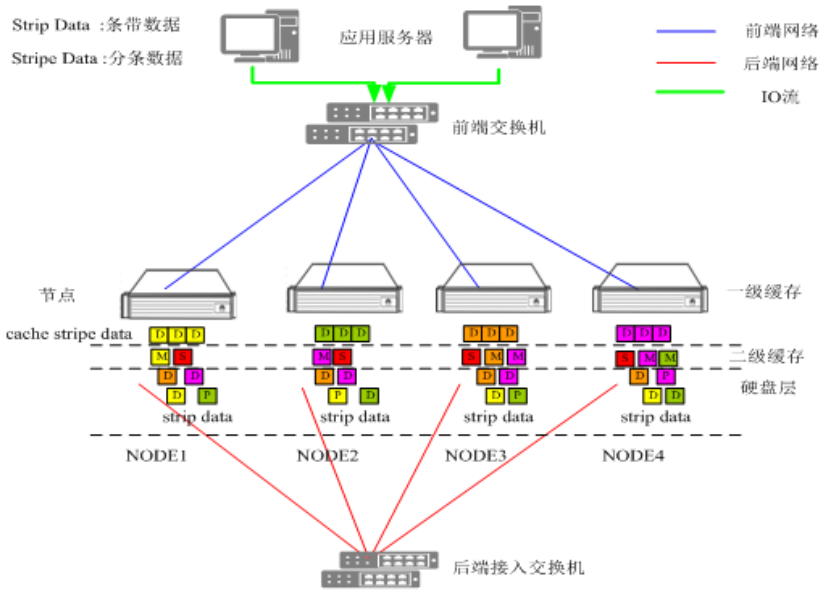
数据掉电保护缓存主要作为写缓存使用，是通过系统内存中规划固定大小的缓存空间，用以保存用户写入数据，来自客户端的写操作数据刷新一级缓存后，对数据切片并进行数据冗余计算后将所有数据片通过存储后端网络发送到每个节点的掉电保护缓存中，便可立即响应客户端写操作成功。数据进入掉电保护缓存就表示数据已经安全，因此，不需要立即刷到对应的硬盘中，这些驻留到掉电保护内存中的数据便可以进行去重和合并处理，即如果对一个数据进行多次修改，则只需要将最新修改的数据刷到硬盘中，对该数据块之前的修改则可以直接丢弃；如果多个数据块是属于一个对象的并且在逻辑上是连续的，则可以将这些逻辑上连续的数据写入到物理连续的硬盘上，这样在数据被访问的时候便可以提高数据的在硬盘的顺序性和连续性，从而提高数据的访问性能。

分布式锁管理

分布式锁管理（DLM）是用于保证全局缓存有效运行，保证全局缓存共享性、一致性的基础。分布式锁管理负责创建分布式锁管理数据结构，该数据结构包括共享资源锁请求、存储共享资源的内存以及锁类型等其它相关内容。只要有进程对该资源有加锁请求，共享资源就始终存在，如果没有任何进程对该资源有加锁请求，分布式锁管理器才能删除该资源。如果进程异常退出。与该资源相关的锁也就被异常退出，与该资源相关的锁也就被异常释放。

3.4.2 实现原理

图3-10 全局缓存原理示意



说明：

D：代表用户原始文件切成的数据条带

S：代表文件系统超级块

M：代表硬盘上管理数据条带块的元数据

P：代表文件条带化时，每个分条（Stripe）中的校验条带数据块（Strip Data）。

缓存与读取

当 Node1 上的文件系统服务收到数据读请求时，首先向分布式锁服务器申请分条资源读锁，加锁成功后，会检查所读数据的缓存是否在全局缓存中以及缓存在哪个节点上，如果该文件分条资源在 Node2 节点上的缓存中，则直接从 Node2 节点上的全局缓存中获取数据并返回客户端，如果不在全局缓存中，则 Node1 上的文件系统服务直接从各个节点上读取该分条数据的所有条带数据后构造出分条数据后再返回给客户端。

缓存与写入

当 Node1 上的客户端 CA 收到数据写请求时，CA 首先向分布式锁服务器申请分条资源写锁，加锁成功后，CA 首先将用户数据接收到本节点上的全局缓存中，然后将该条带数据根据该文件指定的保护级别进行切片处理，对所有切片后的原始数据通过 ErasureCode 进行计算生成校验数据片，最后将包括校验数据片在的数据片写到对应的节点上的保电内存中，写各节点的保电内存成功后则本次写操作成功。

当其它节点上的客户端再次访问该文件分条时，可以直接从该节点的全局缓存中直接读取，而不需要从分条所在的所有节点上的硬盘中读取数据。

缓存释放

- 数据召回

缓存的数据被客户端修改，该客户端的 CA 会加写锁，其他缓存该数据的节点读锁被召回，相应的缓存区数据被释放。

- 数据老化

当节点缓存空间达到老化阈值时，会按照 LRU 来释放最长时间未被访问的缓存数据。

PLStor D 系列中的 Global Cache 将所在存储服务器上的内存空间在逻辑上整合为一个系统全局统一的内存资源池，所有用户数据在整个分布式存储系统中只缓存一份并且对于一个文件分条来说，只在内存中缓存用户的数据条带不缓存校验数据条带。同时，只要位于分布式存储系统中任意一个存储服务器内存中的数据，CA 无论通过哪个存储服务器访问该文件分条数据，都能够从缓存该分条数据的存储服务器内存中命中该数据，从而保证优先访问缓存在全局缓存在中的数据，如果在全局缓存中不命中才从硬盘上读取数据。

这一技术和现有技术相比，PLStor D 系列的全局缓存技术大大提高了整系统内存空间的利用率，对于系统内的热点数据尽可能的避免了不必要硬盘 IO 与网络 IO，充分利用缓存技术提升系统的访问性能。

3.5 文件写示意流程

PLStor D 系列的软件运行在每个节点上，每个节点之间都是平等的，每个文件读写操作都可能同集群中的多个节点进行交互，为了方便表述，我们可以将集群中任意一个节点涉及到业务 IO 的模块表述为以下的模型。

图3-11 I/O 模型

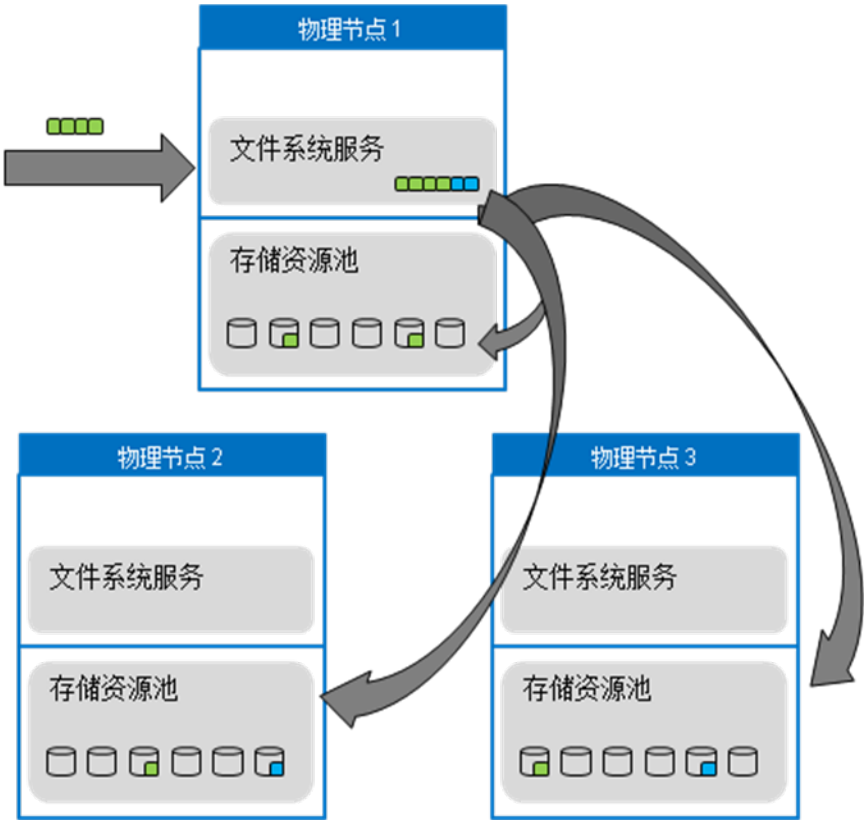
如上图所示，软件层面，分为两层，上层为文件系统服务，下层为存储资源池；上层文件系统服务层负责 NAS 协议的解析，文件操作语义的解析，文件系统元数据管理；下层存储资源池层负责节点上磁盘资源的分配管理和数据的持久化；

当一个客户端连接到一个物理节点上进行一个写文件操作。首先这个操作被文件系统服务层处理，文件系统服务层首先从文件路径＋文件名查找到文件的元数据信息，从而得到文件的布局信息和保护级别。

PLStor D 系列对于文件的数据保护是采用跨节点跨硬盘保护的，一个文件首先被切分为分条（Stripe），每个 Stripe 是由 N 个条带＋M 个冗余校验条带构成；一个分条中的不同条带是放置在不同节点不同硬盘上的。

参考下图，文件系统服务层获得了文件的布局信息和保护级别后，就将收到的数据按照分条的粒度，计算出冗余数据条带来，然后并发通过后端网络写入不同节点的不同硬盘中，每个硬盘上只写入一个条带的数据。

图3-12 写数据流程



从上面的写流程中，我们可以看到：

1：PLStor D 系列具备很好的并发性，每个物理节点都可以同时接入很多的客户端；这些客户端可以并发的进行文件访问操作；

2：高带宽特征，PLStor D 系列将一个文件切分为分条，每个分条交给不同的节点，不同的硬盘来存储，并通过高效的布局，使得即使同一个文件的不同分条也分布在不同的硬盘上，可以充分的发挥集群中多节点多硬盘的能力，大大增强文件访问的性能。

3.6 负载均衡

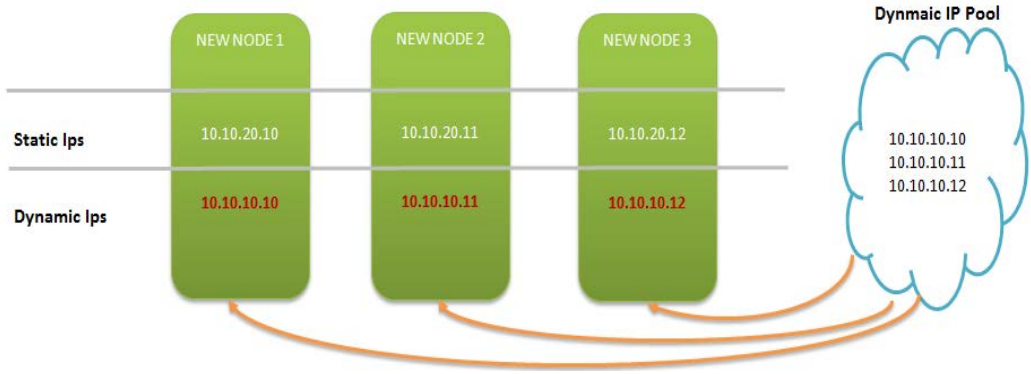
PLStor D 系列负载均衡服务是基于域名请求的负载均衡，仅在域名请求的时候进行干预，不会参加到实际的数据流业务中，对整系统性能影响较小，避免成为整系统性能的瓶颈。区别于大多数 DNS 负载技术，PLStor D 系列负载均衡服务直接集成了响应 DNS 查询的功能，用户不需要额外部署 DNS 服务。

3.6.1 智能 IP 管理

PLStor D 系列对于集群内节点对外提供的接入 IP 进行了统一管理，支持新加入节点的 IP 自动分配，也支持节点 IP 的故障切换(failover)和故障恢复(failback)，用户仅需要配置一个 PLStor D 系列的 IP 地址池即可，不需要对每个节点进行逐一的配置，整个 IP 的管理对于用户会变得更加简单，而且在集群扩容时也容易处理。

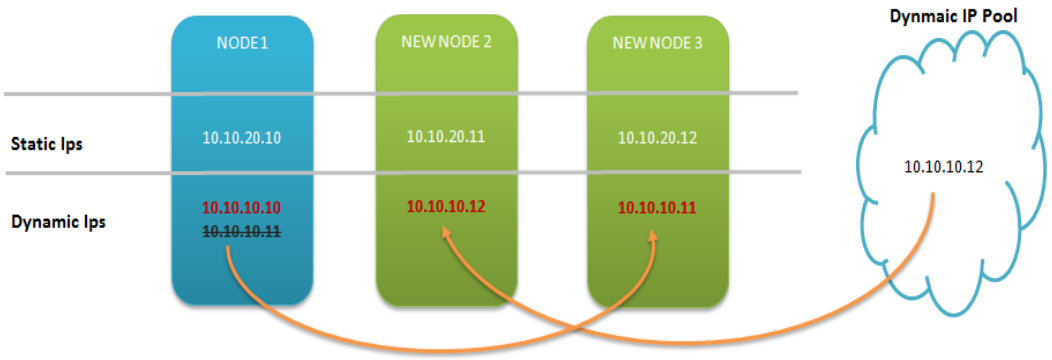
- 每个 PLStor D 系列节点会持有静态 IP 和动态 IP，客户可以使用这些 IP 来访问 PLStor D 系列服务。其中静态 IP 在节点故障后恢复可以保持 IP 不变，动态 IP 则会在节点故障时丢失，并在恢复时重新分配。静态 IP 在环境部署的时候通过部署工具统一配置，而动态 IP 则通过负载均衡服务使用 IP 地址池中 IP 来统一分配。节点的 IP 分配示意如下图所示。

图3-13 节点 IP 分配示意图



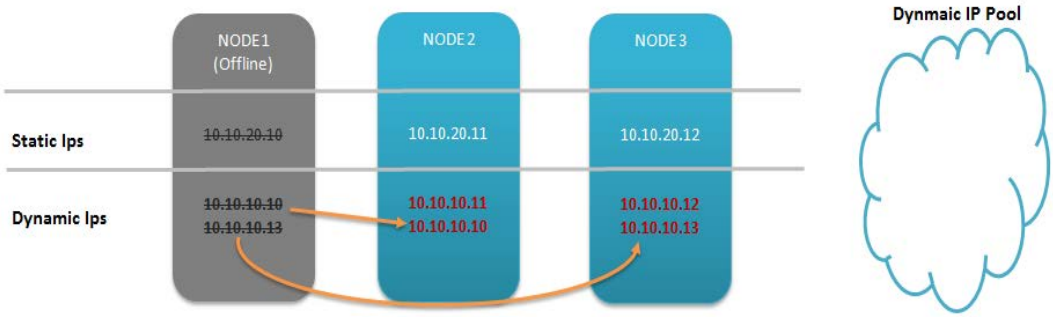
- 节点新加入时刻，负载均衡服务会从 IP 地址池中获取空闲 IP 分配给新加入的节点，如果无空闲 IP，会判断当前集群内节点是否有节点持有多个 IP 的情况，如果有，会从持有多个 IP 的节点强制分配一个 IP 到新加入的节点，确保节点可以参与到负载均衡中来。如果没有，则会通过告警通知 PLStor D 系列系统管理员去为 IP 地址池加入新的空闲 IP，如下图所示。

图3-14 节点加入时 IP 分配示意图



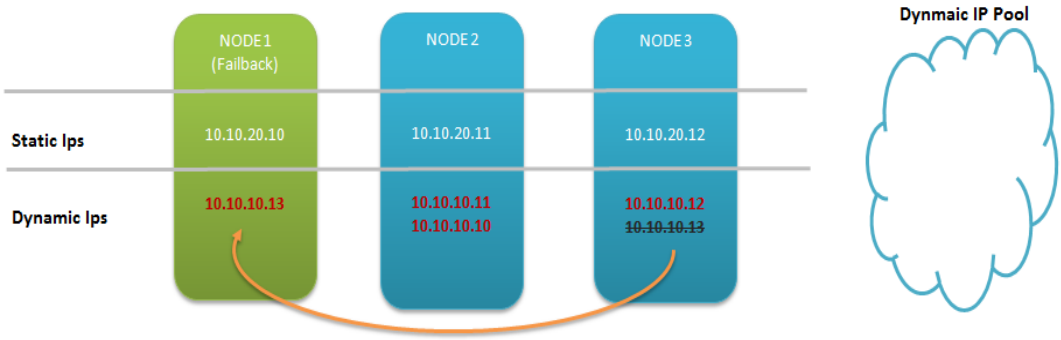
- 节点故障时，如果该节点存在多张网卡，只是部分网卡故障造成 IP 故障，此时会在节点内进行 IP 的故障切换，会把 IP 从故障的网卡切换到正常的网卡，如果一个节点存在多张网卡，在 IP 分配的时候会进行均衡分配。如果该节点故障，则会从集群中选取一个当前负载最低的节点进行接管，如下图所示。

图3-15 节点故障时 IP 切换示意图



- 节点故障恢复时刻，负载均衡服务首先从 IP 地址池中获取空闲 IP 分配给该节点，如果无空闲 IP，会判断当前集群内节点是否有节点持有多个 IP 的情况，如果有，会从持有多个 IP 的节点强制分配一个 IP 到该节点，如果没有，则会通过告警通知用户去为 IP 地址池加入新的空闲 IP，如下图所示。

图3-16 节点恢复时 IP 切换示意图



3.6.2 多样的负载策略

PLStor D 系列的负载均衡服务目前支持多种负载均衡策略，可以供用户按照实际的环境进行配置。

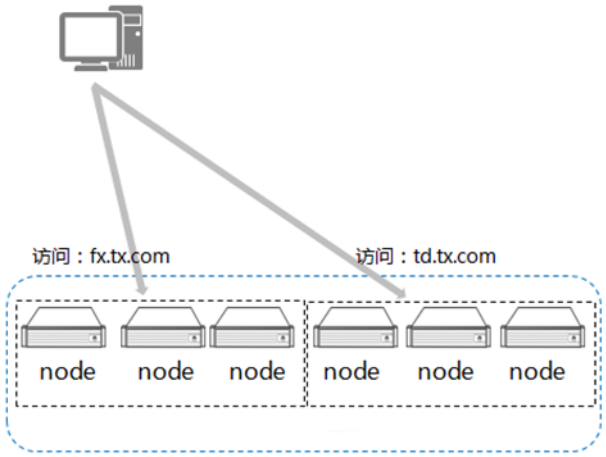
- 轮询方式（默认策略）
- CPU 使用率
- 节点连接数
- 节点吞吐量
- 节点能力值

负载均衡的节点能力由静态能力值和动态负载状态决定。如果某节点负载过重则自动减少该节点的能力值，如果当前负载比较轻，节点能力值也会得到相应的增加。按节点能力值选择节点处理客户端连接请求，节点能力值越高则被选中的机率越高。如果一个节点存在多个 IP，则在多个 IP 间也需要做基于 IP 能力值的选择，能力值高的 IP 会被优先选中。

3.6.3 节点分区管理

PLStor D 系列对所有节点进行分区管理，可以按照需要对节点进行分类，对于每个分区负载均衡系统支持配置独立的负载策略、独立的访问域名。例如，IT 管理员可以划分高性能分区和高容量分区，并给各个分区分配指定能力的节点，配置独立的域名。客户按照需要使用不一样的域名访问不一样的分区。如下图所示，4 个节点可以划分为 2 个分区。客户可以按照需要使用不同的域名访问不同的分区，如下图所示。

图3-17 节点分区访问示意图



PLStor D 系列负载均衡系统提供智能客户端连接管理、负载均衡和故障切换，提高了 PLStor D 系列系统的可用性，保证了 PLStor D 系列系统的高性能。

负载均衡系统通过提供智能的 IP 管理，节点加入时自动分配 IP，节点退出时 IP 自动迁移。使得 PLStor D 系列节点的增加或退出，从客户连接的角度来看是没有变化的，除了性能方面的改善。

3.7 数据重构

当系统中的磁盘或者节点损坏时，系统就会启动数据重构的流程，对未损坏的数据块做 Erasure code 算法的运算，计算出需要重构的数据块，并将该数据块写入到其他正常工作的磁盘上。根据配置的冗余级别的不同，PLStor D 系列最多支持对同一块数据配 4 份冗余数据，所以最多可以容忍 4 块磁盘同时损坏而数据不丢失。

PLStor D 系列在做数据重构的时候，可以将一块磁盘上的数据按照数据对象做划分，把这些数据对象分别重构到其他不同的磁盘上，从而实现了并发重构的功能，将重构时间缩至最短，达到 1 小时重构 2TB 数据的速度。

4 系统特点

4.1 卓越性能

PLStor D 系列系统采用全互联全冗余的组网机制，全对称分布式集群设计，实现存储系统节点的全局统一命名空间，从而允许系统中任何节点并发访问整系统的任何文件；并且支持文件内的细粒度的全局锁，提供从多个节点并发访问相同文件的不同区域，实现高并发读写，最终达到高性能访问系统。

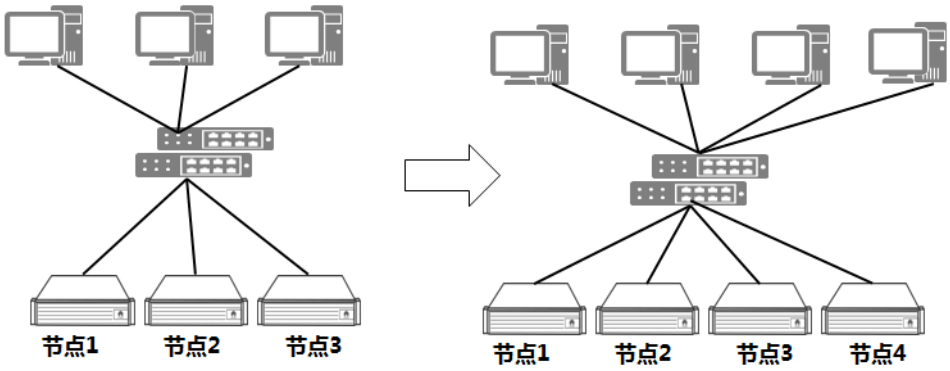
众所周知，数据从缓存中读写远比从硬盘直接读写数据速度快。PLStor D 系列缓存技术，为业务系统提供超大缓存池，有效提高数据访问命中率，提升整体系统性能；硬件采用 SSD 盘存储元数据，加速元数据访问效率，有效提升系统小文件处理能力。同时独有的 InfoTurbo 技术，单客户端带宽高达 2.5GB/s。系统内部采用以太网络全 IP 互联，把系统内部网络延迟降到最小，最终向上层业务提供极低时延的响应。整系统向客户提供业界领先的超过 500 万的 OPS，超过 700GB/s 的系统总带宽，极低的时延，充分满足高性能计算、媒体编辑等场景的高性能要求;不仅单节点可输出高性能，整系统性能也会随着节点扩容线性增长，从容满足业务的更高性能要求。

4.2 灵活扩展

PLStor D 系列支持节点动态扩展，节点数目从 3~288 随需而定，而且节点扩展中业务不中断。随着节点数的增加，存储容量和计算能力线性增加，最终给用户提供呈线性递增的带宽、并发数。PLStor D 系列提供了全局一致的缓存，缓存容量随着节点增加而线性增长，随着节点数目的增加，越来越多的热点数据可以被缓存命中，大大减少硬盘随机 I/O，提高整系统性能。

传统的存储系统需要耗时的规划，升级和维护活动，增加容量或者性能往往需要横向扩展和重新配置应用程序，从而导致中断用户活动，并最终损失工作效率和收入；PLStor D 系列提供全局命名空间，对外呈现为单一文件系统，在扩容时也保持这个特征，分钟级的扩容能力，自动负载均衡，不需要更改配置，不要更改服务器或者客户端的挂载点，不需要更改应用程序，对客户业务无中断。

图4-1 无缝扩展功能示意

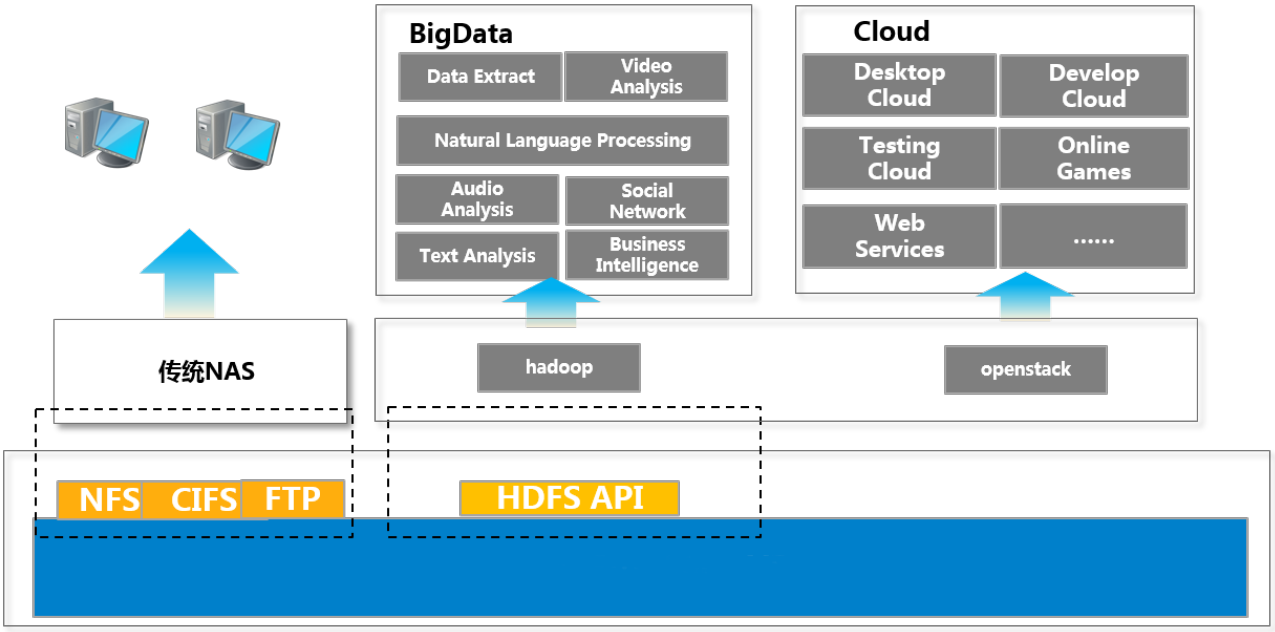


为了既满足客户对性能和容量的需求，又最大程度的为客户节约投资，PLStor D 系列提供了满足不同应用场景的节点类型，客户按需配置，节点统一管理，资源统一调度，轻松管理。

企业，尤其是新兴的企业，在起步阶段，业务量不大，需要的 IT 设施规模也不大，更不可能有大手笔的 IT 预算，但对性能方面的要求又很高。宝德自强 PLStor D 系列大数据存储系统的起步配置可以以低的 TCO，满足企业在容量和性能方面的需求;随着企业发展壮大，对 IT 的需求也在不断攀升，这时可以对原有的 IT 投资保护，只需简单的扩容系统，即可轻松应对存储空间和性能的需求。

4.3 开放融合

图4-2 多生态接入能力



PLStor D 系列同时支持支持 NFS，CIFS，NDMP，FTP 等多种接口，一个系统承载多业务应用，实现数据的全生命周期管理。PLStor D 系列开放的系统，能够很好的对接 Openstack manila 公有云生态系统，其插件化的应用特性组合，在基础架构上加载不同的特性满足客户多方面需求，对同一基础架构上的不同应用之间的数据实现统一调度管理。

4.4 高级特性

4.4.1 多租户

PLStor D 系列支持提供多租户管理能力，不同租户之间的数据逻辑隔离，便于资源划分；统一管理，资源在多个租户间共享，提升租户的安全性，减少初始投资。

一个租户可以同时属于多个资源分组，每个资源分组可以同时提供大数据、对象和文件服务，一个资源分组只能属于一个租户。多租户管理提供了一个通用的可扩展的多租户管理模型，通过多租户管理可实现租户级的配置。多租户能力支持基于租户单元的统一资源管理，以租户为单位分配和管理资源。多个租户共享同一套物理存储系统，租户间资源隔离，确保安全性和隐私。

租户级的配置类型

租户级配置	配置说明
用户管理	支持 AD 域的用户，LDAP 用户管理。
Qos 管理	支持租户级的 Qos 配置。
网络管理	支持 VLAN 配置管理，实现网络隔离。
协议类型	支持 HDFS、对象、文件管理。
审计日志	支持租户级日志配置。
元数据检索	支持租户级元数据配置。

4.4.2 分级功能

PLStor D 系列支持非结构化服务提供的分级功能，实现在一个存储池内的不同类型物理节点划分成不同的硬盘池。允许用户通过策略定义，将高使用量的文件放置在高可用性、高性能的存储设备上，低使用量的文件放置在成本较低的、性能和可用性规格较低的设备上。同时，也允许对接资源池外的支持对象协议的异构设备。

PLStor D 系列存储分级功能被分为热、温、冷三个等级，每个硬盘池都有对应一种分级等级，如 SSD 节点硬盘池为“热”，SAS 节点硬盘池为“温”，SATA 或 NL-SAS 节点硬盘池为“冷”。每个等级可包括若干个硬盘池。同一存储等级内部多个硬盘池间系统自动实现负载均衡，其中包括压力均衡和容量均衡。

PLStor D 系列存储分级功能分级策略分为：写入策略、迁移策略和删除策略

- 写入策略
  - 用于存放文件数据初始写入的位置，该类策略条件包括：文件名、FS/DTree、UID/GID。如果不在以前策略，根据系统默认的放置策略决定放置位置。该策略只支持将数据放置在 PLStor D 存储池。
- 迁移策略
  - 该类策略条件包括以下属性内容：文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、FS/DTree、UID/GID。该策略用于周期性数据迁移确定文件数据迁移的具体

位置。按照策略系统会自动将数据搬迁到指定等级的硬盘池中。数据搬迁的时间默认在每日凌晨零时，可根据具体情况自定义。可以迁移到内部硬盘池或异构设备。

- 该类策略条件包括以下属性内容：文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、UID/GID。按照策略一次性数据迁移系统会自动将数据搬迁到指定等级的硬盘池中。该搬迁策略只有执行当时符合策略的文件数据，并且只会在在配置后一次性招执行。也可以从异构设备中回迁数据。

- 删除策略

  - 该类策略用于周期性删除指定的文件。该类策略条件包括以下属性内容：文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、UID/GID。周期性删除策略对于启用了 S3 服务的命名空间，还可以自定义前缀、标签、过期策略、非当前版本过期天数等条件。系统自动按照删除策略，无需管理员操作，直接删除文件。

  - 该策略用于用户一次性删除指定的文件。与周期性删除策略不同，该一次性删除策略任务只会执行一次。

4.4.3 QoS 功能

PLStor D 系列存储支持非结构化服务能力或结构化服务能力提供配置 QoS 策略：

- 非结构化：PLStor D 系列存储非结构化服务能力以租户、用户、命名空间（文件系统/桶）或客户端粒度配置策略，可以基于容量或上限配置；支持文件、HDFS、对象服务设置租户级 QoS;租户级粒度 QoS 支持 OPS 和带宽的上限控制，这个租户内所有的命名空间（文件系统/桶）共享 QoS。租户级 QoS 控制与命名空间级（文件系统/桶）QoS 控制采取叠加式控制原则；客户端级粒度的 QoS 支持以客户端 IP 为粒度进行 OPS 和带宽的上限控制，同样，客户端级根据叠加式控制原则；用户级 QoS 支持以用户粒度设置带宽和 OPS 上限。SMB 协议和对象服务支持用户 QoS，但是需要管理员配置 Windows 用户和 S3 用户映射给 UNIX 用户。

- 结构化：PLStor D 系列存储结构化服务的 QoS 支持基于卷或池设置性能目标，支持按带宽和 IOPS 进行配置最大性能目标，可分别限定读写总性能、读性能和写性能；定时策略可以基于业务繁忙情况下避免各种 IO 风暴影响生产业务。

4.4.4 生态兼容性

PLStor D 系列支持 NFS、SMB、HDFS、S3、iSCSI 等多种协议接口，承载多业务应用。作为分布式存储系统，PLStor D 系列支持主流的开放接口，提供丰富的兼容性，与客户的云数据中心和应用平台平滑集成。

- 存储协议兼容性：

  PLStor D 系列提供的各种类型存储服务均遵循业界主流的协议和标准。

  - 块服务接口：遵循标准的 SCSI 语义和 iSCSI 协议，提供分布式块存储业务。
  - 文件服务接口：遵循标准的 NFS/SMB 协议和 POSIX 语义的并行文件客户端，提供分布式文件存储业务。
  - 对象服务接口：遵循 S3 接口标准，实现了对 S3 主要功能兼容。
  - 大数据服务接口：遵循原生的 HDFS 协议。

HDFS 大数据服务支持对文件进行以下接口调用：

HDFS	**命令作用**
append	追加一个文件到已经存在的文件末尾
setacl	用于设置文件或目录的访问控制列表（ACL）权限
setxattr	用于设置文件或目录的扩展属性（.xattr）
rm	删除文件或文件夹
du	统计文件夹的大小信息
count	统计一个指定目录下的文件节点数量
mv	在 HDFS 目录中移动文件

5 缩略语和术语

表5-1 缩略语清单

AD	Active Directory
CLI	命令行界面
DAS	Direct-Attached Storage
DNS	Domain Name System，域名系统
Erasure Code	纠删码
FTP	File Transfer Protocol
GID	组 ID
GUI	图形化用户界面
HTTP	Hypertext Transport Protocol
IPMI	Intelligent Platform Management Interface
LDAP	Lightweight Directory Access Protocol
NAS	Network Attached Storage
NIS	Network Information Service
NVDIMM	非易失性内存
RAID	Redundant Arrays of Inexpensive Disks
RDMA	Remote Direct Memory Access
SAN	Storage Area Network
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment
SSD	Solid State Disk
TCP	Transmission Control Protocol
UID	用户 ID