TengLing
腾凌科技

# 腾凌 T2000-N-24S 全闪阵列
# 技术白皮书

文档版本　V2.0

发布日期　2024-11-04

北京腾凌科技有限公司

## 商标声明

TengLing和其他腾凌商标均为北京腾凌科技有限公司的商标。本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

## 注意

您购买的产品、服务或特性等应受腾凌科技公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，腾凌科技公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

**北京腾凌科技有限公司**

总部：北京 邮编：100095

地址：北京市海淀区地锦路 9 号院 14 号楼四层

电话：400-6506106

网址：www. bjtengling. com

# 前言

随着软件定义存储解决方案的部署越来越深入行业用户的应用场景，技术供应商正在积极按需调整产品配置，尤其在文件细分市场，正在向大容量存储和数据高效处理两个方向发展。一方面，用户考虑到成本效益和可扩展性，正在采购中低端价格段的软件定义文件存储，用于将生产、业务系统之外的数据归档到能够提供大容量存储能力的系统上来，同时可以简便的管理和做到灾难恢复的功能，确保数据安全以及业务连续性，以及在落地 AI 等新兴应用之前进行数据的池化工作。另一方面，由于越来越多的软件定义供货以全闪存的方式提供，可以在高性能计算等场景中提供极低延迟和极高吞吐量，又同时具备灵活扩展的能力，确保了企业投资的长期价值，成为企业用户的投资重点。

过去的几年，数据量的快速增加和数据价值的应用挖掘，催生了 IT 的技术革新，尤其是存储设备的革新。

- 主存介质从 HDD 向 SSD 的切换；
- 主存协议从 SAS 向 NVMe 的切换；
- 对存储系统访问端到端的时延也由 10ms 向 1ms 演进；
- 对数据存储要求逐步从单纯的块/文件向容器、虚拟化、云等多业务负载演进；

一系列的需求发展和关联技术的变革对客户 IT 基础设施的设计和构建提出了更高的要求，要构建一个现代化的 IT 基础设计，选择一个合适的存储系统是最为关键的环节。

多协议、多业务、温冷数据融合的新型存储是构建 TCO 和效率最优的 IT 系统的基础。存储系统持续的高性能以及叠加各种增值高可靠配置后依然稳定的性能是构建一个智能 IT、规模扩展能力的 IT 系统的基础。高效存储是 IT 系统构建的成本关键因素，存储数据

的高效流动、智能运维、不中断升级、长期供应保障则是 IT 系统长期演进和发展的基础。

# 目录

1 全闪阵列存储背景介绍 ..... 5
2 腾凌自主创新存储系统简介 ..... 7
2.1 腾凌自主创新技术概况 ..... 7
2.2 腾凌自主创新存储产品 ..... 8
3 技术方案 ..... 9
3.1 设备组成 ..... 9
3.2 设备工作原理 ..... 9
3.3 关键技术分析 ..... 11
3.3.1 双控 active-active 工作模式 ..... 11
3.3.2 双机双活 ..... 12
3.3.3 SAN/NAS 统一存储 ..... 14
3.3.4 RAID 技术 ..... 15
3.3.5 数据快照技术 ..... 16
3.3.6 自动精简配置 ..... 18
3.3.7 NFS 权限管理 ..... 20
3.3.8 主机多路径支持 ..... 20
3.3.9 带外管理技术 ..... 24
3.3.10 存储管理功能 ..... 25
3.3.11 智能分层存储 ..... 25
3.3.12 LUN 拷贝技术 ..... 27
3.3.13 掉电保护 ..... 30
3.3.14 数据镜像 ..... 31
3.3.15 异构数据迁移 ..... 33
3.3.16 一键销毁 ..... 33
3.3.17 服务质量控制 ..... 34

# 1 全闪阵列存储背景介绍

全闪阵列存储是一种将多个闪存驱动器组合成一个逻辑单元的数据存储技术，它通过RAID(独立冗余全闪阵列)技术提高数据的可靠性和性能。以下是全闪阵列存储的背景介绍：

1.全闪阵列的组成：全闪阵列由多个硬盘组成，这些硬盘通过 RAID 技术组合起来，提供比单一硬盘更高的数据读写速度和数据冗余保护。全闪阵列的组成主要包括 CPU、控制器、网卡、硬盘和缓存等部件。

2.RAID 技术的基本功能：RAID 技术主要提供三个基本功能：通过对全闪存上的数据进行条带化，实现对数据成块存取，提高数据存取速度；通过对一个阵列中的几块闪存盘同时读取，提高数据存取速度；通过镜像或者存储奇偶校验信息的方式，实现了对数据的冗余保护。

3.全闪阵列的优势：

1）高可靠性：全闪阵列采用 RAID 技术，可以在部分硬盘故障的情况下，保证数据的完整性。这使得全闪阵列适合存储关键业务数据。

2）高性能：通过多硬盘并行读写，全闪阵列可以提供较高的 I/O 性能。适用于需要处理大量数据的应用场景，如数据库、数据分析等。

3）集中管理：全闪阵列通常具备管理软件，可以集中管理存储空间、快照、双活等任务。这降低了管理员的工作负担，提高了管理效率。

4.适用场景：

1）数据中心：企业级数据中心通常使用全闪阵列作为核心存储设备，以满足高并发访

问、持久性存储和共享访问的需求。

2）党政建设：政务云数据中心建设依托国家电子政务外网构建政务云平台体系，整合算力资源，支撑大数据、人工智能、区块链等新技术创新应用，面向政务部门提供绿色集约、共享共用、安全可靠的一体化存储服务。

3）医疗领域：全闪阵列为医院信息系统提供数据处理和安全，存储大量医疗数据如病历和影像，确保数据可用性和快速访问。支持高并发请求和数据增长时的存储扩展。在 PACS 系统中，处理小文件保持性能，应对影像数据增长。总体上，为医疗领域提供稳定、高效的存储解决方案。

4）教育科研：全闪阵列在教育科研中提高数据存储效率，满足高性能计算、AI 和大数据需求。它分散数据至多硬盘，增强安全性和可用性，支持复杂数据模型，适应数据增长。灵活配置支持项目需求，促进数据共享和科研合作，推动创新。

5）能源电力：全闪阵列在能源电力领域的应用通过提供高可用性和容灾解决方案，确保电力系统的稳定运行，同时采用节能技术和绿色存储设计，降低能耗并提高能源利用率，支持新能源驱动的海量存储系统，从而在保障电力行业信息化和数据安全的同时，促进能源的高效利用和可持续发展。

5.市场趋势：随着数据量的爆炸性增长和云计算的兴起，全闪阵列在企业级存储中的地位愈发重要。未来，全闪阵列将更多地融入云存储架构，支持数据的实时同步和异地备份，提高数据的可用性和灾备能力。

# 2 腾凌自主创新存储系统简介

## 2.1 腾凌自主创新技术概况

腾凌科技掌握存储底层核心技术，是目前国内唯一一家深入研究存储底层核心协议的厂商，基于核心协议和技术的深入研究，自主研发出腾凌存储引擎，以专用硬件管理存储业务，成功突破了 SAS 控制和 RAID 控制等存储核心技术领域的壁垒，实现了 SAS 控制和 RAID 控制的国产化，并独创了芯片级的 iSCSI 和 LUN 加速技术，深度优化 IO 处理流程，极大地提高了设备的整体性能。

1.领先的架构

先进异构计算架构，适配飞腾、海光、龙芯、Intel 等 CPU，管理业务分离，更高的性能支撑。

2.自主核心技术

深度掌握存储底层协议和核心技术，国内唯一完全自主研发的 SAS 芯片。

3.硬件加速技术

ASIC 级别解析 iSCSI 协议，基于 TOE 技术的硬件级 iSCSI 协议解析引擎，高效解析 iSCSI 协议数据，进一步提升数据处理效率，加速整机性能；ASIC 级别 RAID，完全自主研发 ASIC 级别的硬 RAID 引擎，从底层实现完整 RAID 协议，填补国内空白。

4.自主存储平台

完全自主研发的存储平台，简洁、完善的图形化管理界面，丰富的软件特性，支撑全系列存储产品，确保了功能和用户界面的一致性，便于快速推出新产品。

5.完全自主知识产权

软件、硬件核心技术完全自主知识产权，腾凌 T2000 存储阵列采用 Intel Xeon 处理器，极大提升业务处理能力，消除传统架构的性能瓶颈。

## 2.2 腾凌自主创新存储产品

腾凌 T2000 存储阵列是腾凌科技基于 Intel Xeon 处理器，自主研发的国内首个企业级自主创新存储系统，基于 CPU+存储专用芯片的先进架构优势，使得 T2000 的性能提高 50%，功耗降低 60%，真正实现自主创新的目标。腾凌 T2000 采用双控制器架构，支持主备、双主工作模式；模块化设计，硬件完全冗余设计；并且支持 1/10/40GbE 接口、8/16/32Gb FC、10/40FCoE、4*12GbMINI SAS 接口。支持链路聚合和链路冗余，保障系统高可用性和高性能；具备专用管理网口，具备串口；

全闪阵列设备支持的数据保护功能：精简配置、动态扩容、LUN 快照、LUN 拷贝、远程同步、远程镜像等高级容灾功能，支持在线数据压缩，采用专用芯片加速压缩，提升压缩效率；支持界面统一管理，支持 UNIX，Linux，Windows，AIX、HP-UX、Solaris 等；支持与 OpenStack 云计算平台对接，由 OpenStack 统一管理存储资源。

腾凌 T2000 设备不仅从硬件、软件、设计三方面都实现了完全的自主创新，并且具备高性能、高稳定与高可靠性，满足数据库 OLTP/OLAP、Exchange、服务器虚拟化和视频监控等各种应用需求，可以为政府、企业、医疗、教育等多个重要行业用户的关键业务提供存储能力的支撑，确保信息安全，规避安全隐患，助力中国向着自主、安全的信息化发展方向迈进。

# 3 技术方案

## 3.1 设备组成

设备主要由控制器、机箱、背板、电源、硬盘等组成。总体部件框图如图 3-1。

主控板负责运行全闪阵列存储系统，对外提供存储业务，采用 Intel Xeon 自主处理器。

背板采用国产高速连接器和国产电源连接器把各硬件模块链接起来，从电源模块获取电力提供给主控板和硬盘，提供高速数据通道用于主控板与硬盘连接以及主控板之间的双控数据传输和管理。

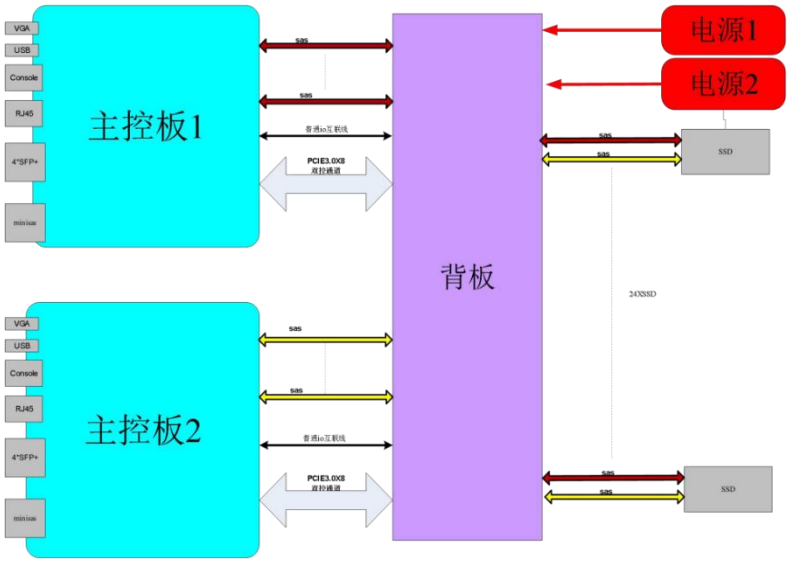


图 3-1 总体部件框图

## 3.2 设备工作原理

存储阵列为 19 英寸，高度为 4U、24 盘位全闪阵列，可通过扩展坞扩展至 28 盘位，采用 Intel Xeon 处理器，支持千兆网络、万兆网络、FC 网络。双控制器架构，所有模块都有冗余设计，无单点故障，支持不停机维护，在线调整，电源、控制器主板、硬盘等都支持热插拔。

整机提供 2 个电源模块槽位，2 个存储控制器槽位，主控板内 Intel Xeon 处理器运行我司自研统一存储软件平台。后端采用 SAS 技术访问硬盘，前端采用基于 TOE(TCP offload engine)技术的万兆网络以及 FC 网络提供存储服务。

其中统一存储软件平台由我司自主研发，该平台利用 RAID 技术将 SSD、HDD 等存储介质抽象成资源池，并通过 FC、iSCSI、CIFS、NFS、FTP 等协议对上层应用提供服务，实现 SAN、NAS 一体化。该软件平台已经稳定运行在飞腾、龙芯、海光、Intel 等硬件设备上，支撑公司推出了全系列存储产品，对外通过国家权威测试机构多次严苛测试，设备可用性达到 99.9999%，并在各行业用户稳定运行多年。

统一存储软件平台主要软件功能如下表:

| 关键软件特性 | |
|---|---|
| RAID 等级 | 0,1,5,6,10,50,60,线性 RAID |
| 系统兼容性 | UNIX、Linux,Windows 等 |
| 虚拟化支持 | 支持 VMWare,Hyper-V,XenServer,KVM 等 |
| OpenStack 支持 | 支持与 OpenStack 云计算平台对接，由 OpenStack 统一管理存储资源 |
| 支持 LUN 数目 | 4096 |
| 管理软件 | 具备基于 Web 的图形化管理界面，支持统一管理，支持操作日志、故障统计、邮件告警、性能监控、系统配置导出和导入、系统诊断 |
| 效率安全存储池重构 | 可根据业务繁忙程度智能调度重建速率 |
| 资源效率提升 | 智能 RAID 校验、RAID 自动重建、RAID 在线扩容、介质扫描、数据完整性检测、硬盘定位 |
| 双活控制器架构 | 双控间高速通道数据交换，请求无痕切换，延迟大大降低 |
| 智能硬盘管理 | 支持智能硬盘休眠 |
| 数据保护 | 数据复制、文件迁移、数据镜像、克隆、自动精简、快照、压缩、重删、自动分层、本地逻辑卷镜像、回收站等 |

## 3.3 关键技术分析

### 3.3.1 双控 active-active 工作模式

我司双控之间包含管理和业务双通道，管理通道主要用于同步配置修改、双控协商等管理相关数据通信，业务通道主要用于内存镜像、读写请求转发等用户数据传输。双控通道采用 PCIE 互连方式，且管理通道优先处理，确保即时响应各类事件。其中一块控制器故障则另一块控制器自动接管全部功能。

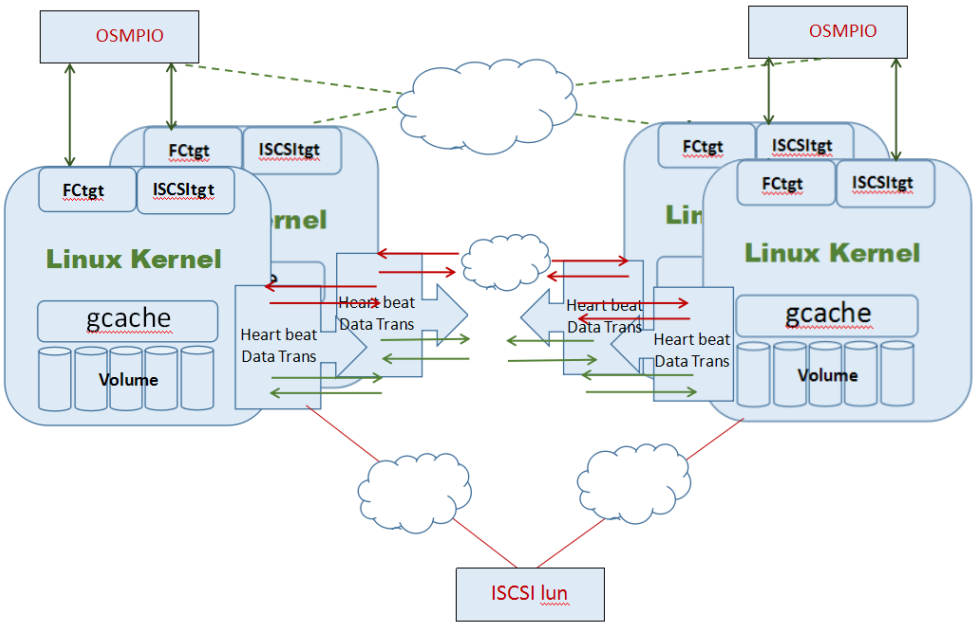


图 3-2 存储双活内部结构示意图

双控管理上具有主备之分，分别为 Backup（管理备机）和 Master（管理主机），所有管理请求由主控发起，如果管理终端的管理请求发送至备控，备控会转发至主控，并由主控处理。双控通过硬件互连 IO 线可以自动判断在位状态和主备优先级。各控制器同时通过管理通道传输心跳报文和状态报文，确保各自运行状态，及时处理异常事件。

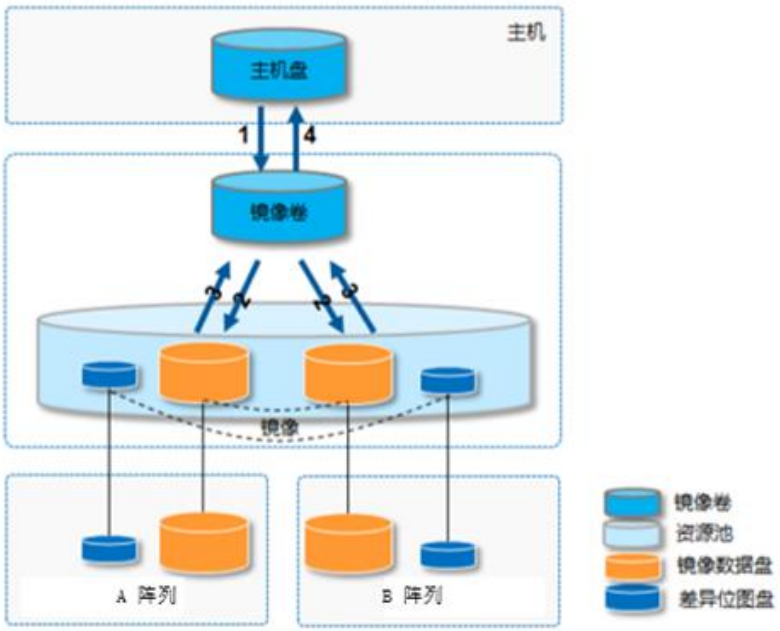


图 3-3 存储双活写 IO 流程

双控的数据访问采用 active-active 工作模式，两个控制器同时响应用户 IO 请求。应用服务器可以根据多路径的配置，决定服务器每次 IO 派发控制器，每个 IO 请求只会发送到单个控制器，接收到 IO 请求的控制器为访问控制器。另外，所有 LUN 都具有一个归属控制器，所有 LUN 的访问都需要经过归属控制器进行处理。

### 3.3.2 双机双活

我司双机双活实现 AA 双活，两套阵列的 LUN 实时镜像同步，且两端能够同时处理应用服务器的请求，提高资源的利用率和系统的效率、性能，让客户从双机双活系统中获得最大价值。

在链路故障或单套存储阵列故障时，服务器多路径软件与存储阵列的心跳检查中断，服务器则会自动将所有请求切换到另外一套存储阵列。在单台存储阵列故障时，另外一套存储阵列可以单独提供服务，故障恢复后自动进行数据一致性增量同步，快速恢复数据一致性。

存储阵列间互联支持 IP、FC 链路，可根据用户应用场景选择链路类型及连接方式。设备间支持单条或多条链路，建议双机双控进行交叉互联，最大限度保证安全性。交叉互联示意图如下两种，分别为直连交叉互联和交换机交叉互联。

双机间存在任意一条链路则双机之间就可以正常数据同步。

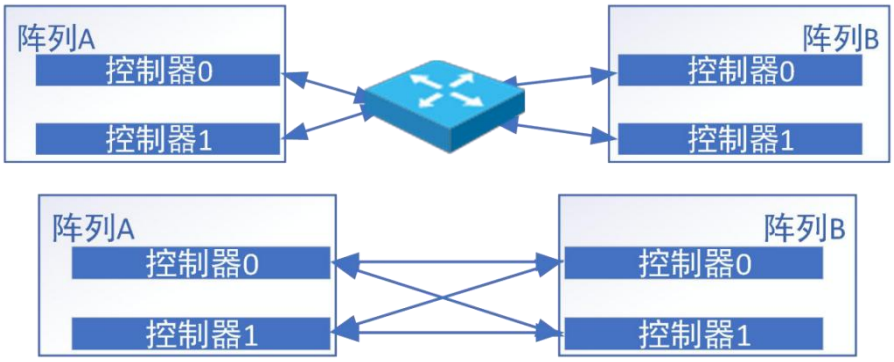


图 3-4 双机双活互连方式

在双机之间链路故障时，阵列之间无法实时同步，为保障数据的一致性，我司具备完善的仲裁机制决定那套阵列提供服务，解决脑裂问题。双机双活组网示意图如下图。

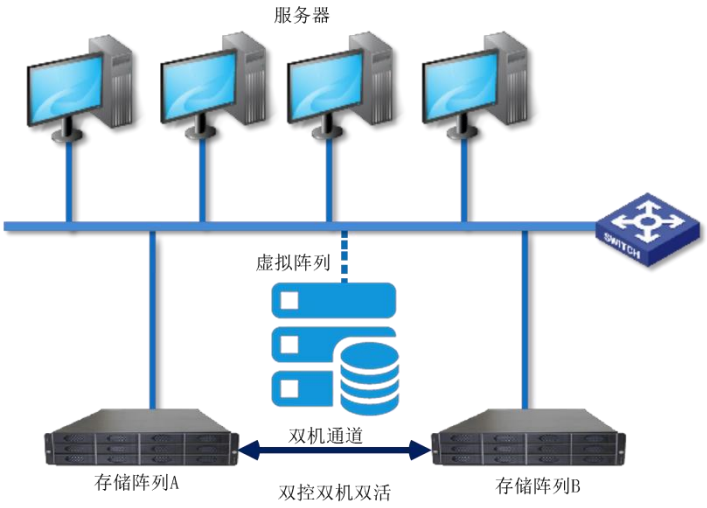


图 3-5 双机双活示意图

服务器端基于 ALUA 方式多路径软件，将双机映射硬盘在服务器中自动虚拟机成单块硬盘。在正常情况下，服务器写入数据时，每条 IO 写请求只会向单条路径发送写数据指令，

收到写指令阵列在写入本地同时发送写指令到另外一台阵列，双机都执行成功后才会通知服务器命令执行成功，确保数据实时一致性；服务读取数据时，收到读取数据指令的阵列只会读取本设备以响应读取请求，提供更快的响应速率。

在其中一台阵列不允许访问情况下，服务器写入数据时，读写指令只会发送到正常阵列，数据仅写入正常阵列，正常阵列同时记录双机差异位置，待异常阵列恢复后，通过读取差异来执行增量同步，快速恢复数据一致性。

当双机间链路故障时数据无法实时镜像，为保证数据一致性，只能由一台阵列向服务器提供服务。为防止双机脑裂，我司提供两种仲裁机制：静态优先级机制和仲裁服务器机制。

静态优先级机制无需第三方仲裁服务器，用户配置一台阵列为优先站点，另一台阵列为非优先站点。该模式下阵列间链路故障时非优先站点不允许访问，仲裁服务器机制在阵列间链路正常时，双机可正常访问。当阵列间链路异常时，为保证数据一致性，双机只能有一台阵列提供服务，另外一台阵列不允许访问。阵列间链路异常时由仲裁服务器进行仲裁，仲裁胜利的阵列提供服务，无法与仲裁服务器通信的阵列仲裁失败，双机都能与仲裁服务器通信时，优先站点仲裁胜利。

### 3.3.3 SAN/NAS 统一存储

我司软件平台中实现 SAN 和 NAS 功能融合，在一套软件平台中提供了原生的文件共享接口和块映射接口，为客户提供了一套同时满足文件访问和块访问的统一存储产品，保护了客户投资。

我司 SAN/NAS 统一存储的架构如下图所示,LNU 和文件系统是平行的运行在智能分层存

储池之上，存储池之下为 RIAD 子系统，LUN 可以通过 FC 或 iSCSI 协议实现快存储映射，文件系统可以通过 CIFS、NFS 或 FTP 协议实现文件共享。

### 3.3.4 RAID 技术

我司统一存储软件平台支持 RAID 包括 RAID0、1、10、5、6、50、60，可以根据用户具体应用场景决定使用 RAID 协议，下面简单介绍一下部分 RAID 的实现原理。

RAID0：RAID0 在结构形式上将多个物理硬盘切成等大小的 CHUNK 块，取每块硬盘的一个 CHUNK 组成一个条带，所有条带组成一个大的逻辑盘。每个条带分布在所有硬盘上，条带中所有 CHUNK 都用作存储数据，不具有数据副本功能，这样的结构形式使得 RAID0 也叫作为条带 RAID。因为 RAID0 提前将硬盘分化成 CHUNK 块大小组织了条带，在存储数据时，其将按照数据在条带中的不同的 CHUNK 块所在的盘来进行 IO 映射，然后将数据写入映射后的具体硬盘中。

RAID1：RAID1 实现了数据镜像，将组成 RAID1 的硬盘分为主盘和镜像盘，在写 RAID1 时，需要在主盘上存储数据，也必须同时在镜像硬盘上写一样的数据，这样当主盘发生故障时，镜像盘就可代替主盘继续工作。因为 RAID1 将数据数据保存了双份，所以 RAID1 的数据安全级别较高，但其闪存空间利用率最低的，因为无论用多少闪存做 RAID1，仅算主盘的容量。

RAID10：RAID1+0 这种结构是两个 RAID1 再组成 RAID0 的结构，产生的原因是为了结合起来，补足 RAID-0 和 RAID-1 各自的缺点，同时能够做到安全又高性能，RAID1+0 也被叫做 RAID-10 标准，构成 RAID10 至少需要四块盘，它的优点是同时拥有 RAID-0 的高性能和 RAID-1 的镜像数据可靠性，缺点是硬盘空间利用率只有一半。所以它主要用于对容量需求不大，对

性能、响应时间和安全性要求较高的数据库等时时在线业务中。

RAID5：带循环奇偶校验的条带 RAID，和 RAID0 相同的是都是条带 RAID,但和 RAID0 不同的是，RAID5 需要在每个条带中选择一个 CHUNK 存储冗余信息，RAID5 的方式是存储奇偶校验信息，其组成形式是把数据和相对应的奇偶校验信息存储到组成 RAID5 的各个闪存上，并且奇偶校验信息和相对应的数据分别存储于不同的闪存上。当 RAID5 的一个数据发生损坏后，可以利用其他完整的数据和相应的奇偶校验信息去恢复被损坏的数据继续工作。

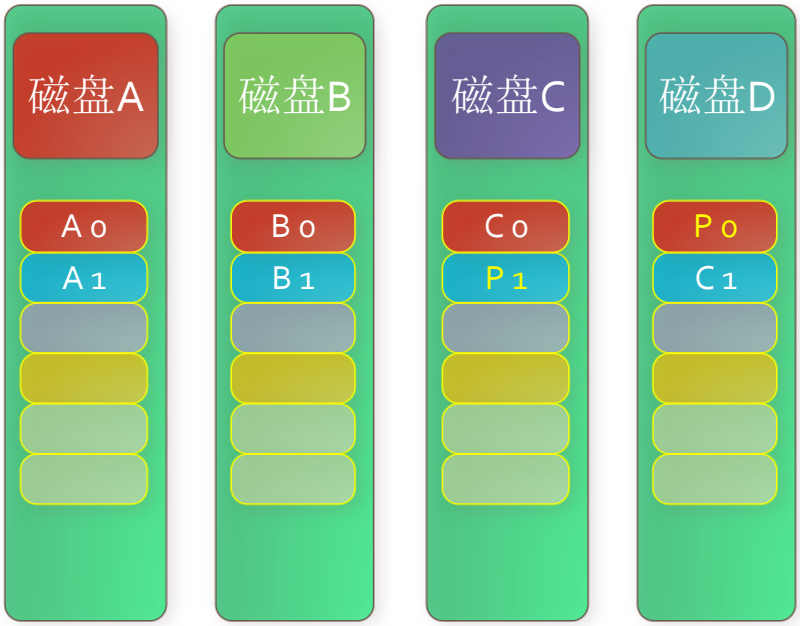


图 3-5 RAID5 结构图

RAID6:RAID-6 是 RAID5 类似,它在 RAID-5 基础上,进一步加强冗余数据保护数量,RAID-6 设计采用双重校验方式，能够防止双盘故障失效导致 RAID 数据丢失，RAID-6 的数据保护能力很强，比 RAID5 和 RAID10 都要好。但是 RAID-6 又增加了一份校验数据，增加了数据写入时校验值计算量，所以数据存储的效率比 RAID-5 还要低。

### 3.3.5 数据快照技术

统一存储软件平台在创建存储池时，会将所包含的 RAID 按照固定大小划分成若干物理块，当基于存储池创建 LUN 时，LUN 会根据用户配置从存储池划分若干物理块，并记录物理块与 LUN 的逻辑位置的关系，用户访问 LUN 时会根据逻辑位置查找对应物理块进行读写。统一存储软件平台的快照就是将源 LUN 逻辑块与物理块关系克隆一份，并在后续数据写入 LUN 时申请新的物理块，保证原有物理块数据不发生改变。下面简单介绍快照创建、具有快照的源 LUN 数据读写流程、快照回滚。

1.快照创建：

快照创建时首先在存储池中创建新的逻辑卷，然后将源 LUN 的逻辑块与物理块关系复制给该逻辑卷，使得该逻辑卷与源 LUN 共享所有物理块。因为逻辑卷与源 LUN 映射关系完全相同，所以该逻辑卷包含了与源 LUN 相同的数据，该逻辑卷就是源 LUN 当前时刻的快照。快照创建过程无需拷贝任何用户数据，只需要复制逻辑关系，快照可以瞬间创建成功。

2.具有快照的源 LUN 读写流程：

源 LUN 进行数据写入时，存储池会检查物理块是否共享，如果物理块为共享块，则为源 LUN 申请新的物理块，并将旧物理块数据拷贝到新物理块，之后再将新数据写入新物理块，最后修改源 LUN 逻辑块与物理块的对应关系，使得源 LUN 逻辑块位置指向新的物理块位置。如果物理块并未与快照共享，则直接写入数据。

传统存储池中 LUN 没有逻辑块与物理块的关系，在写时拷贝时无法为源 LUN 申请新的物理块，只能为快照申请新的物理块。在快照较多时，该方案在写时拷贝需要为每个快照申请一个物理块，并执行多次写时拷贝，浪费较多存储空间且写入缓慢。

我司快照方案，即使快照个数很多，也只需进行一次块申请及一次数据拷贝，减少了存

储空间占用，挺高了写入效率。

3.快照回滚：

传统写时拷贝方案在快照回滚时需要将快照中数据拷贝至源 LUN 中，拷贝完成后才能完成回滚。我司快照回滚时，只将快照的逻辑块与物理块的对应关系复制给源 LUN 即可恢复原有数据。回滚过程无需拷贝任何用户数据，只需拷贝逻辑关系，快照也可以瞬间完成。

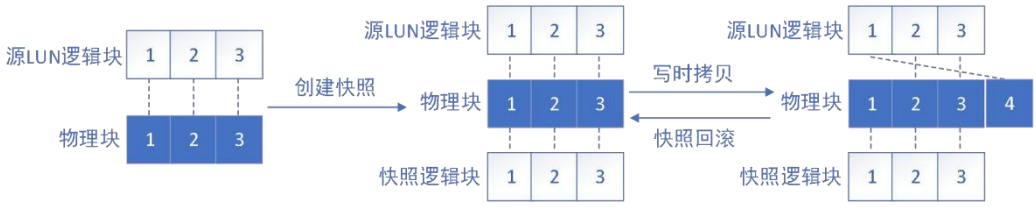


图 3-6 快照映射关系图

### 3.3.6 自动精简配置

我司统一存储软件平台自主研发的分层存储池同时支持精简 LUN 和非精简 LUN。非精简 LUN 在创建时按照用户配置容量分配所有存储空间，提高存储资源使用效率，更大限度满足业务的性能需求。精简 LUN 创建时不会预先分配存储空间，但会将用户配置容量形态呈现给用户，使用户看到的存储空间等于用户配置存储空间。精简 LUN 配置的容量为逻辑空间，实际占用的空间为实际存储空间。

在精简 LUN 使用时进行按需分配，用户用多少则分配多少，存储池中未使用的空间可以分配给任何需要存储空间的 LUN。因为精简 LUN 在创建时并不分配空间，所以允许用户创建超过存储池容量的 LUN，如果数据容量超过存储池容量时，可以动态扩展存储池容量。

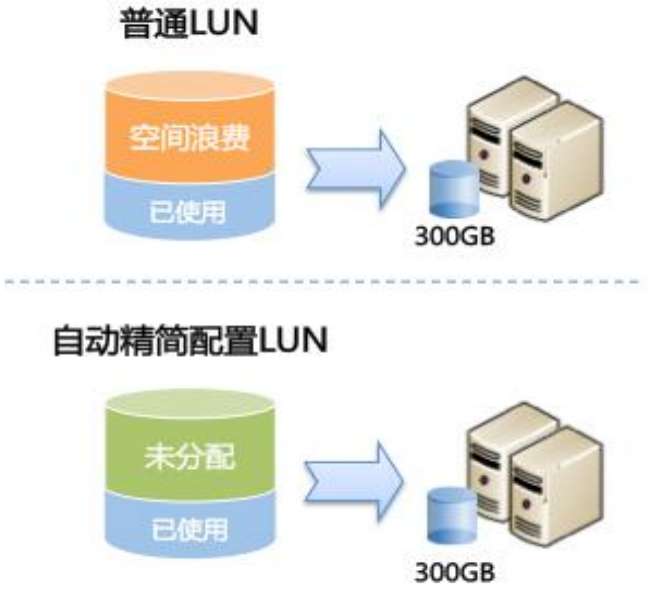


图 3-7 自动精简技术原理

因为精简 LUN 的实际存储空间并未完全分配，在数据读写时需先查询 LUN 的读写位置是否已经分配存储空间，根据查询结果不同处理方式不同。

1.数据读取：

精简 LUN 收到服务器读 IO 请求后，若未查询到已分配的存储空间，则将该区域设置为全零并返回给服务器，此流程只是返回给服务器全零数据，并不会分配存储空间；若查询到已分配的存储空间，则将该存储位置的数据读取后返回至服务器。

2.数据写入：

精简 LUN 收到服务器写 IO 请求后，若未查询到已分配的存储空间，则先从存储池中分配存储空间，每次最少分配空间为 4MB，然后将数据写入分配好的存储空间，并记录 LUN 的逻辑块与物理块关系；若查询到已分配的存储空间，则将该存储位置的数据读取后返回至服务器。

另外，精简 LUN 支持标准的 unmapSCSI 命令以进行空间回收。在虚拟硬盘空间回收等应

用场景时，精简 LUN 在收到 unmap 命令后，若查询到已分配存储空间，会直接释放存储池中存储空间，并删除精简 LUN 中逻辑空间与实际空间的映射关系。

### 3.3.7 NFS 权限管理

我司统一存储软件平台支持 NFSv2、NFSv3、NFSv4、NFSv4.1 版本，可根据用户需求使用特定版本的 NFS 协议。为保证数据的安全，我司支持 NFS 多重权限管理，包括访问网段控制、访问读写权限控制、普通用户是否匿名控制以及 root 用户是否匿名控制。

访问网段控制可以限制允许访问的客户端 IP 或者客户端网段，限制其他网段或 IP 访问 NFS。

访问读写权限控制可以限制某些客户端以读写权限访问，而其他客户端只能以只读权限访问。

普通用户是否匿名控制用来控制访问 NFS 的用户身份信息，如果配置为不匿名访问，NFS 使用客户端原有的用户 ID 和用户组 ID 检查共享文件的权限；如果以匿名用户访问 NFS，则用户使用匿名用户 ID 和用户组 ID 检查共享文件的权限。

由于 root 用户权限较大，因此允许专门限制 root 用户是否匿名访问。如果配置为不匿名访问，则 NFS 使用 root 用户检查共享文件的权限；如果 root 用户匿名访问，则 NFS 使用匿名用户检查共享文件的权限。

### 3.3.8 主机多路径支持

普通的服务器上的硬盘是一块硬盘挂到一个总线上，当总线或者硬盘故障时则数据无法访问或丢失。服务器使用存储阵列时，服务器可以通过多网卡与存储阵列进行连接，服务器

到服务器就可以有多条路径进行选择。如下图中组网所示，主机到存储之间的 IO 由多条路径可以选择。

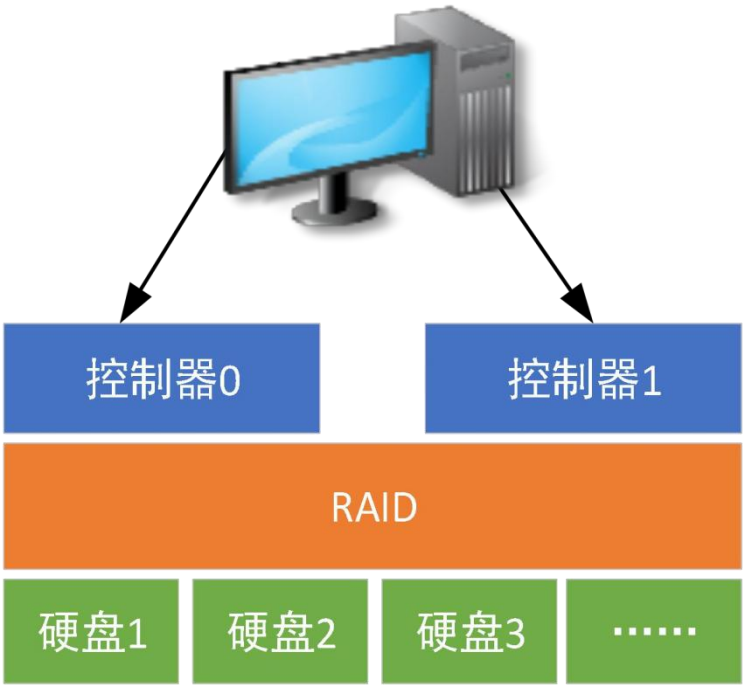


图 3-8 主机多路径示意图

根据上图所示，服务器到所对应的存储可以经过两条不同的路径，如果是同时使用的话，读写流量如何分配？其中一条路径坏掉了，如何处理？还有在操作系统的角度来看，每条路径，操作系统会认为是一个实际存在的物理盘，但实际上只是通向同一个物理盘的不同路径而已，这样是在使用的时候，就给用户带来了困惑。多路径软件就是为了解决上面的问题应运而生的，通过硬盘虚拟化解决每条路径一块硬盘问题，通过路径管理分配 I/O 派发路经并解决路径故障问题。

硬盘虚拟化的目的是将多条路径映射的同一个 LUN 虚拟化成一个盘，方便服务器直接访问而不关心每条链路的状态。每条路径都可以映射一块 SCSI 硬盘并都具有一个 SCSI 地址，该地址由 initiator ID、bus ID、target ID 以及 LUN ID 组成，在实际组网中，initiator

ID 一般对应主机 HBA 端口，bus ID 在 SAN 映射硬盘中一般固定为 0，target ID 一般对应存储阵列控制器端口,LUN ID 表示 LUN 在 target 中的序号。多路径软件为区分不同路径映射硬盘是否为相同的 LUN，都借用 SCSI 设备的 WWN，当 WWN 相同时表示两条路径映射的 SCSI 设备为同一个 LUN。

我司统一存储软件平台中，每个 LUN 在创建时就会生成唯一的 WWN，直到该 LUN 被删除前该 WWN 不会发生改变。

通过检查映射 SCSI 硬盘的 WWN 是否相同，多路径软件将两条或多条路径映射硬盘虚拟化为一块硬盘。如下图所示，LUN1 通过路径 1 和路径 2 在服务器中映射出 DISK1 和 DISK2，多路径软件检查两块硬盘的 WWN 相同，将两块硬盘虚拟化为 mpatha。同样的，LUN2、LUN3 分别虚拟化为 mpathb、mpathc。

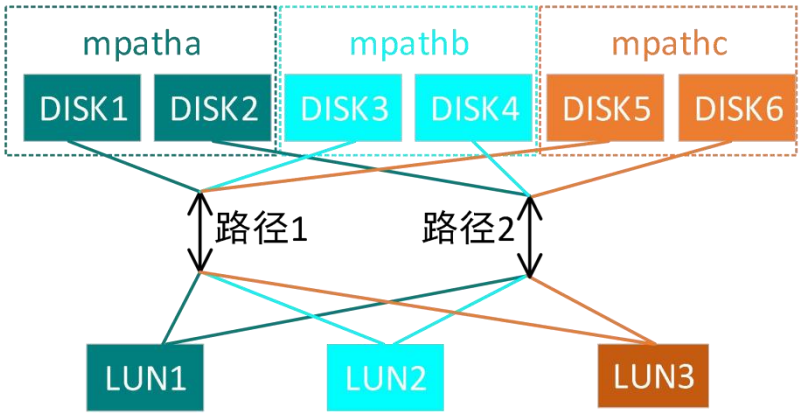


图 3-9 多路径映射

服务器通过多路径软件可以将多条路径的 SCSI 设备映射为一块虚拟的硬盘，但虚拟硬盘读写数据时每条读写指令只需向一条路径发送,多路径软件需要选择每条执行发送的控制器。如下图所示，服务器访问 LUN 共包含 4 条路径，分别为：

1.服务器端口 1<->交换机 1<->控制器 1 端口 1

2. 服务器端口 1<->交换机 1<->控制器 2 端口 1

3. 服务器端口 2<->交换机 2<->控制器 1 端口 2

4. 服务器端口 2<->交换机 2<->控制器 2 端口 2

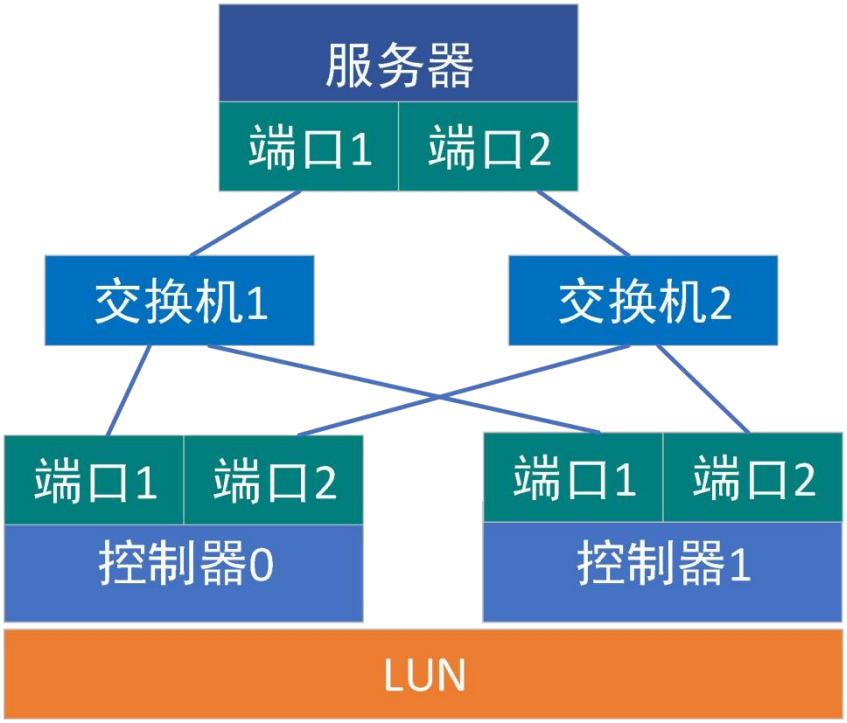


图 3-10 服务器访问路径

我司统一存储软件平台的控制器支持 Active-Active 模式，每条路径都可以访问任意 LUN，但通过不同的控制器访问其效率是不同的，通常一个 LUN 会有一个优选控制器，通过该控制器来访问该 LUN 效率最高，通过其他控制器来访问该 LUN 则会有效率的损失。如果某路径访问 LUN 归属控制器，则通知服务器该路径为活动优化（Active/Optimized，简称 AO 路径），如果某路径访问 LUN 非归属控制器，则通知服务器该路径为活动未优化（Active/NonOptimized，简称 AN 路径）。多路径软件将所有 AO 路径归为一个路径组，所有 AN 路径归为一个路径组。

当一个路径组内包含多个路径时，读写指令在路径组内进行负载均衡多路径软，充分利

用多条链路的带宽，提高系统整体的吞吐能力。负载均衡的方式有三种，分别为轮询算法、当前 IO 最小路径算法、IO 服务时间最短算法。

如果存在 AO 路径组，则优先访问 AO 路径，只要存在 AO 路径，多路径软件将不会访问 AN 路径。当所有 AO 路径故障后，多路径软件将自动切换到 AN 路径进行数据访问。

当 AO 路径恢复后，多路径软件会自动恢复到 AO 路径进行数据访问。

### 3.3.9 带外管理技术

本设备使用 AST2500 作为 BMC（Baseboard Manager Controller）的核心芯片，BMC 实现了设备的温度、电压、风扇、电源等一系列监控，并基于监控结果进行智能调节工作，以保证系统处于健康的。用户也可以通过 BMC 将设备硬件信息和日志记录导出，用于后续问题定位。

BMC 是一个独立的系统，它不依赖于设备的 CPU、内存等硬件，也不依赖于 BIOS、存储 OS 等统一存储软件平台系统，当统一存储软件平台故障时，BMC 可以通过复位的方式重启存储系统。

BMC 采用 IPMI（Intelligent Platform Management Interface，智能平台管理接口）作为管理接口，IPMI 良好的自治特性便克服了以往基于操作系统的管理方式所受的限制，例如操作系统不响应或未加载的情况下其仍然可以进行开关机、信息提取等操作。

在工作时，所有 IPMI 功能都是向 BMC 发送命令来完成的，BMC 基于 IPMI 规范接收并处理系统事件消息，维护描述系统中传感器情况的穿管器记录。在需要远程访问系统时，IPMI 基于 LUN 上串行（SOL）改变 IPMI 会话过程中本地串口发送方向，可以为本设备提供紧急管

理服务、串口控制台的远程访问，实现远程查看 BOOT、OS 系统加载器或紧急事件管理控制台以诊断并修复设备相关问题的标准方法，并可以配置系统引导阶段各种组件。

### 3.3.10 存储管理功能

我司统一存储软件平台支持 Web 界面、命令行、SNMP 等多种配置管理方式，并且支持对外提供 WebService、Restful 等访问接口，可以方便灵活的与第三方存储管理平台或业务系统集成。

Web 管理支持全中文图形界面，同一个管理界面同时管理 SAN 和 NAS 的所有功能和特性，针对系统的使用和操作有专门的描述，功能操作有有帮助，有完善的提示信息，清晰易懂。平台图形界面不仅可以配置设备，还通过图表、列表形式展示阵列的配置、状态、映射关系、性能监控等信息。另外，平台可记录丰富的日志信息，包括用户修改配置的操作日志、管理员登录日志、系统报警日志、映射用户登录日志等等，日志也可以通过邮件、SNMP 等方式及时通知到第三方监控平台。

命令行访问方式有两种，分别为串口访问和 telnet 访问，命令行支持了基础业务的功能配置，可以满足用户的基础业务配置需求。平台支持 SNMPv2、SNMPv3 协议，可以通过 SNMP 查看端口、LUN、存储池等设备信息，还可以监控系统日志信息。

另外，系统提供了丰富的 WebService、Restful 访问接口，便于统一存储软件平台被第三方管理平台或业务系统集中管理。

### 3.3.11 智能分层存储

我司统一存储软件平台自主研发的智能分层存储池可以智能分析数据的访问频率，将访

问频率较高的数据存放到高性能硬盘，将访问频率较低的历史数据存放到低性能硬盘。智能分层存储池能够同时满足用户性能和容量需求，并且降低用户成本。

智能分层存储池将存储空间按照 4M 划分若干数据块，LUN 的空间分配也会按照块大小进行分配的。如下图所示，智能分层存储池将各层 RAID 划分成固定大小的物理块，每个 LUN 也会划分相同大小的逻辑块，每个逻辑块指向一个物理块，逻辑块对应的物理块迁移就可以改变数据存储的位置。在数据访问时，存储池以块为单位进行统计和分析数据的访问频率，并通过数据迁移将访问频率较高的数据迁移到具有较高性能的硬盘中，将访问频率较低的数据迁移到具有高大容量且更低容量成本的硬盘中。

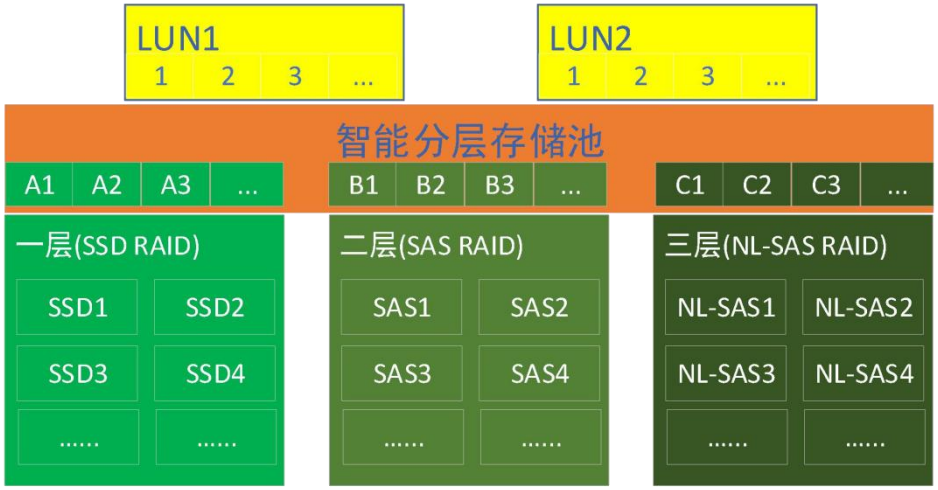


图 3-1 智能分层存储

智能分层存储池优化存储空间需要三个步骤，分别为统计 IO 访问频率、访问频率排序、数据迁移。

统计 IO 访问频率是基于存储池的块进行统计的，在存储池存在的生命周期内，对每个下发的 IO 都会进行记录，用于后续访问频率排序。IO 访问统计信息全部记录在内存中，设备启动后重新统计 IO 访问频率，不统计设备重启前 IO 访问频率。每次数据迁移完成后会将

IO 访问频率进行加权，弱化历史访问频率对后续统计排序的影响。

访问频率排序是根据 IO 访问频率进行排序，在同一个存储池中将所有数据块访问频率从高到低进行排序，并根据存储池各层容量大小确定数据迁移的阈值。按照访问频率越高的数据块存放到越高性能硬盘中，确定每个块应该映射到那种硬盘中。

根据 IO 访问频率，迁移数据块在分层池存储位置，使得访问频率较高的数据尽可能的存放到较高层，访问频率低的数据尽可能存放到较低层。数据迁移支持两种触发方式：手动触发数据迁移和定时触发数据迁移。手动触发迁移可以根据用户需要进行手动触发，定时触发可以设定好固定的迁移开始时间。

### 3.3.12 LUN 拷贝技术

随着各行业数字化的推进，产生了因设备升级或数据备份而进行数据迁移的需求。传统的数据迁移过程是存储系统—应用服务器—存储系统。这样的数据迁移速度较慢，且迁移过程中还会占用应用服务器的网络资源和系统资源。如何提升数据迁移速度成为了急需解决的问题。LUN 拷贝技术应运而生，在拷贝过程中的数据直接在存储系统间或存储系统内传输，并可同时在多个存储系统间迁移多份数据，满足了用户快速进行数据迁移、数据分发及数据集中备份的需求。

定义：LUN 拷贝是一种数据拷贝技术，可以同时在设备内或设备间快速地进行数据的传输，从而扩展了系统存储的复制功能，会将所有数据进行完整地复制。

LUN 拷贝特性的目的和给用户带来的好处：

在多个设备之间同时对多份数据进行拷贝：用户可根据需求在多个设备内

创建多个目标 LUN,同时进行数据拷贝。

即时调整拷贝速率当业务繁忙时，可降低 LUN 拷贝的拷贝速率，减小对业务的影响；当业务空闲时，可提升 LUN 拷贝的拷贝速率，加快拷贝进度。

#### 3.3.12.1 LUN 拷贝

LUN 拷贝每次启动都会把源 LUN 数据完整地拷贝至目标 LUN,在数据较大的情况下，LUN 拷贝的备份窗口很长。

腾凌提出 LUN 拷贝功能，客户可以在大量的应用中使用 LUN 拷贝以随需应变的方式创建逻辑驱动器实现完整拷贝。为帮助防范本地断电，可以使用 LUN 拷贝将数据从一个阵列转移到另一个阵列，从而为防止数据丢失提供了更多的手段。通过以不同方式创建数据的第二份拷贝，可以帮助防范各种存储介质的故障，并实现数据恢复提供一份随时可用的数据拷贝。

这一能力是优化存储系统和应用性能的关键。使用 LUN 拷贝来优化性能和数据保护作为一种优化工具，客户可以为提高性能或数据保护而使用 LUN 拷贝进行数据移植。配置为阵列形式的逻辑驱动器在出现性能问题时可以使用LUN拷贝将一个或更多的逻辑LUN转移到统一存储系统的内部的另一个阵列上，这样可以减轻原始阵列上对逻辑驱动器进行平衡的压力。与原始阵列相比，腾凌产品的目标阵列还可以提供更大的性能空间，从而帮助平衡负载。例如驻留在 RAID5 阵列上的逻辑驱动器可以使用 LUN 拷贝将数据移植到一个 RAID10 阵列中，这将可以帮助提供数据保护和提高吞吐量。

同样的，如果一个逻辑驱动器不再需要更高级别的性能和数据保护，也可以使用 LUN 拷贝将其转移到一个性能或数据保护功能的较低阵列上。

### 3.3.12.2 远程 LUN 拷贝

远程 LUN 拷贝以前只是一项一对一的、完全复制型的技术，用于离站灾难恢复。将一个 LUN 从这儿复制到那儿。可能是块级的，也可能是文件级的，但都是整个 LUN,把所有东西从一处复制到另一处。但是，随着技术的不断成熟，如今它已应用于整个数据中心完成无数个任务，包括连续数据保护和热故障转移支持。

由于它是一项完全复制技术，有些厂商开始优化拷贝技术，以便实现异步拷贝，更好地管理带宽利用和连接中断。然后，拷贝技术才真正地成为一项大家负担得起的基于主机的技术。这也成为主机故障转移系统或基于主机集群的基础。

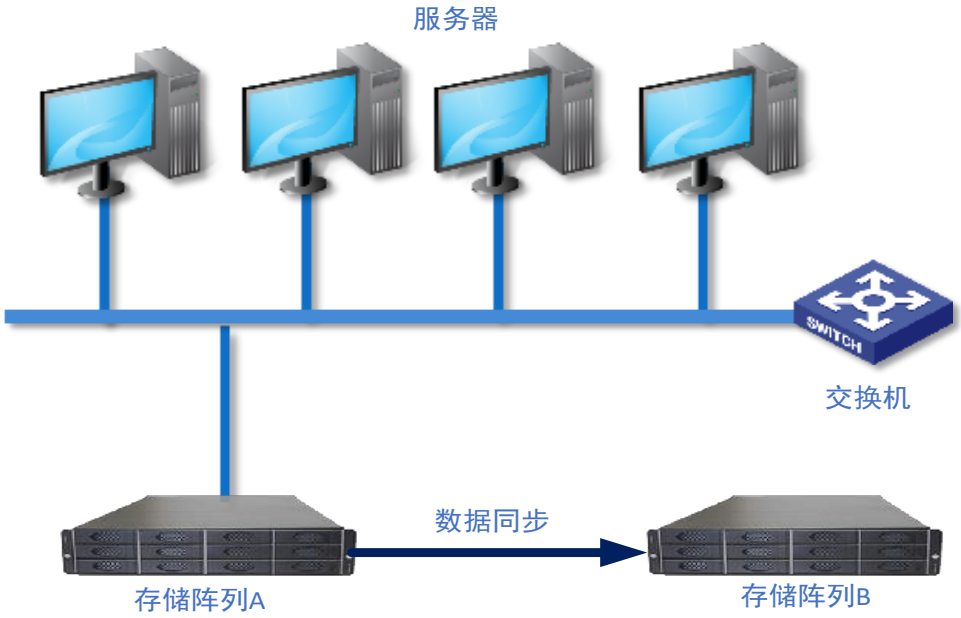


图 3-212 远程复制组网示意图

当前的远程拷贝实现方式可以分为阵列级和主机级。基于阵列的存储拷贝通常是相同的阵列。阵列之间需要专用的通道和设备进行连接。这种方式的优点是可靠性强、速度快。但是此类配置需要昂贵的投入，并且不具备通用性。

而腾凌存储设备是基于主机级存储拷贝，在一定程度上屏蔽了底层硬件的差别，主机之

间的硬件不需要完全相同，这种方案通常可以架构在已有的设备之上，它不仅性能好，而且成本低，具有很强的通用性。

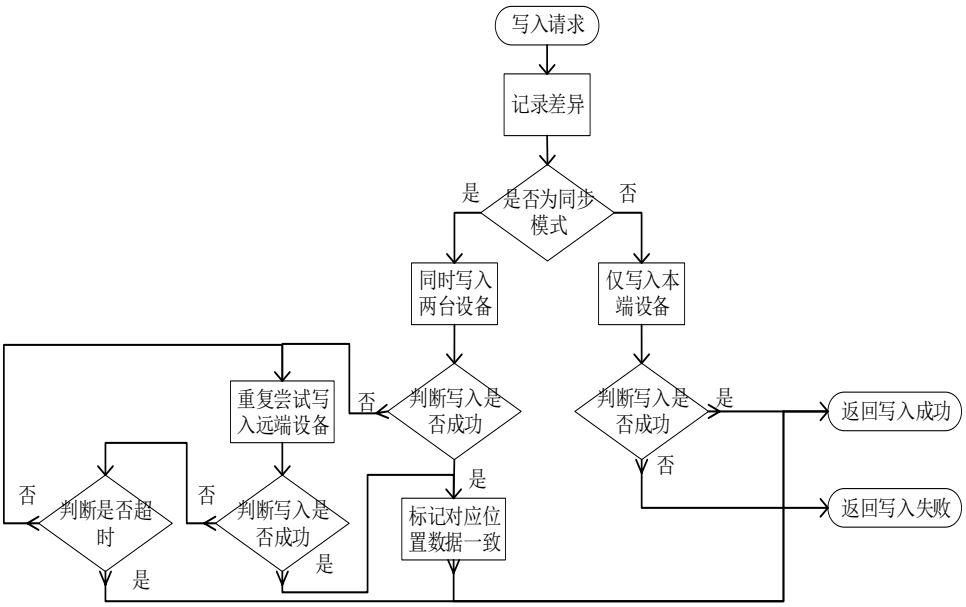


图 3-13 远程复制写入流程图

### 3.3.13 掉电保护

腾凌 T2000 可内置 BBU 模块（备电）,当存储阵列在供电故障后，能利用 BBU 模块将各控制器内存中的缓存数据刷入到缓存中。待阵列电力恢复后，系统在启动时，再将缓存中的缓存数据恢复到内存，保证数据不丢失。

BBU 模块位于控制器上，即误操作拨出控制器，BBU 模块也会随着控制器一同被拨出。控制器上软件模块在监测到控制器拨出后，能利用控制器上的 BBU 模块(备电),将用户缓存数据备份到缓存中,保证数据的完整性。系统的掉电刷缓存流程，由底层系统完成，不依赖上层软件单元，确保了刷盘过程不会受业务影响，进一步提升了用户数据的可靠性。

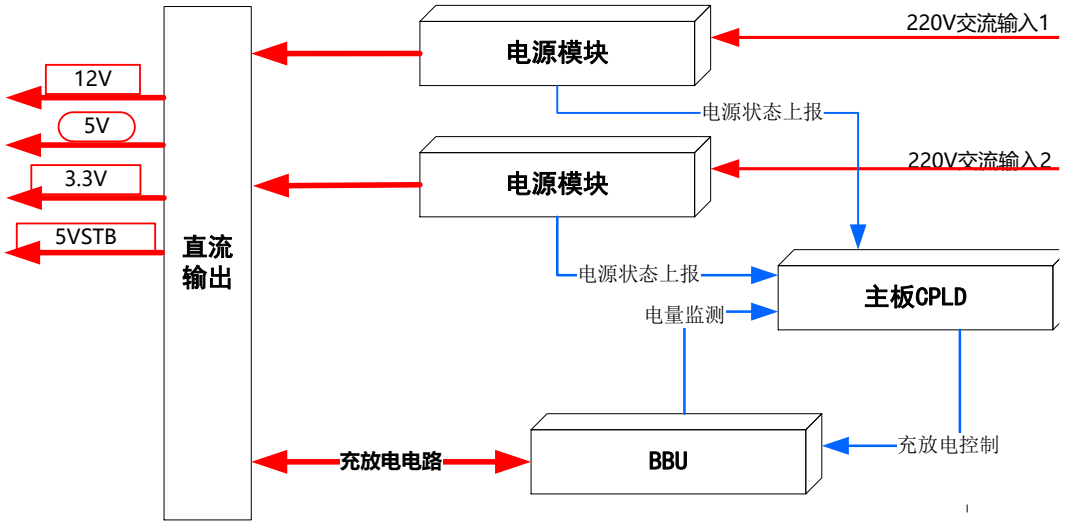


图 3-14 BBU 工作原理示意图

### 3.3.14 数据镜像

我司统一存储软件平台是一种数据保护技术，可以为一个 LUN 创建两个物理副本，在应用侧无感知的情况下加强对 LUN 提供持续的冗余备份保护。

数据镜像要求一个 LUN 的两个镜像物理副本分配在两个存储池中，这种情况下其中一个存储池故障，也可以保障 LUN 上的数据不丢失且可以正常工作，应用侧可以毫无感知的正常工作，故障镜像物理副本恢复后，可以通过正常镜像副本自动增量同步。

如下图所示，数据镜像功能共包含一个镜像 LUN、两个副本 LUN，三个 LUN 逻辑容量相同，但镜像 LUN 并不占用实际的存储空间，副本 LUN 分别在两个存储池中占用相同。下面分别介绍数据镜像的读写流程及故障恢复流程。

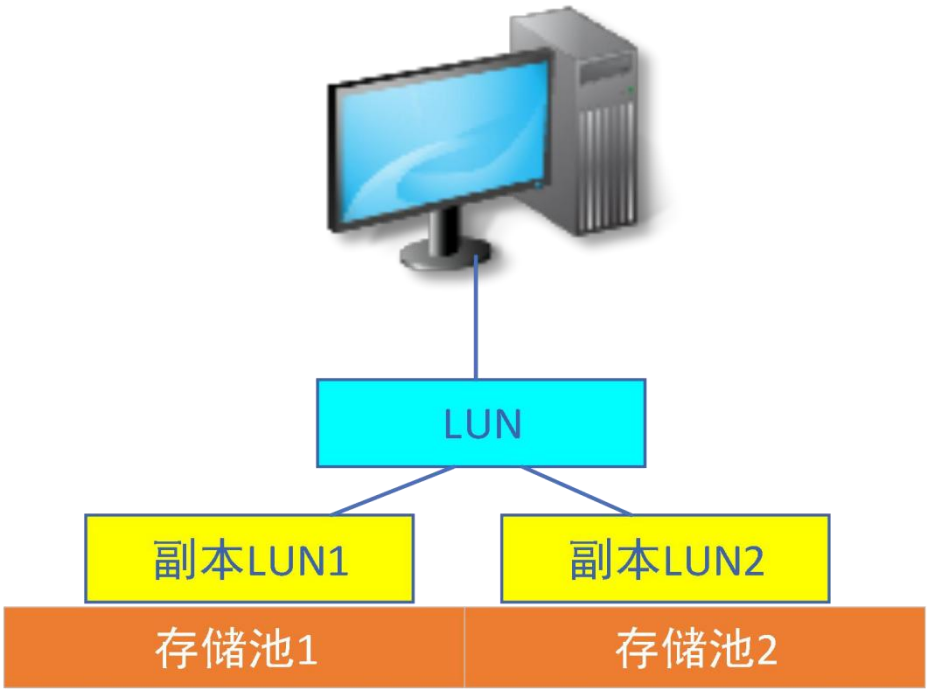


图 3-15 智能分层存储

1. 读写流程

在数据写入时，服务器写执行发送至镜像 LUN，数据镜像功能模块首先检查两个副本 LUN 是否正常，若两个副本 LUN 都正常，则同时向两个副本 LUN 写入数据，若两个副本 LUN 都写入成功后则直接通知服务器写入成功，若其中一个副本 LUN 写入失败则记录镜像副本差异位置且标记写入失败副本 LUN 故障，并通知服务器写入成功；若一个副本 LUN 故障且另外一个副本 LUN 正常，则向正常副本 LUN 写入数据并记录两个镜像副本 LUN 差异位置，写入成功后通知服务器写入成功。

在数据读取时，服务器读指令发送至镜像 LUN，数据镜像功能模块两个副本 LUN 状态，选择一个数据完整的副本 LUN 读取数据，若数据读取成功后将读取数据返回服务器，若数据读取失败则标记当前副本 LUN 故障，并尝试通过另外一个正常副本 LUN 读取数据。

2. 故障恢复

当标记一个镜像副本 LUN 故障后，会立即启动副本 LUN 状态检查机制。检查机制通过定时将正常副本 LUN 数据拷贝到故障副本 LUN 中，检查是否故障副本 LUN 是否能够成功写入数据，如果无法正常写入数据则认为副本仍处于故障状态，若故障副本 LUN 能够正常写入数据则认为该副本 LUN 已经恢复正常。

故障副本 LUN 恢复正常后，系统将启动数据同步机制，将正常副本 LUN 中数据增量拷贝到故障副本 LUN，尽快恢复两副本数据一致。

### 3.3.15 异构数据迁移

我司统一存储软件平台支持智能化的数据迁移，可以在不中断原有业务的情况下实现将源 LUN 上的数据全量迁移到目标 LUN 上，实现业务无感知的情况下完成数据迁移。数据迁移不仅支持本机数据迁移，还支持异构系统之间的数据迁移。

数据迁移实现了源 LUN 的数据完全复制到目标 LUN，并在复制结束后使用目标 LUN 代替源 LUN。数据迁移包含业务数据同步和继承源 LUN 标识信息两个步骤。

1.业务数据同步

源 LUN 中所有数据全量复制到目标 LUN 中，在迁移过程中若源 LUN 数据发生改变，则修改数据也会复制到目标 LUN 中。

2.继承源 LUN 标识信息

数据迁移完成后，目标 LUN 将继承源 LUN 的 WWN 等信息，确保服务器业务无感知。

### 3.3.16 一键销毁

我司统一存储软件平台具备一键销毁功能，可以一键销毁存储阵列的所有配置信息及数

据信息。

销毁包含快速销毁和安全销毁功能，快速销毁会删除配置信息及硬盘管理数据，安全销毁功能不仅会删除配置信息及硬盘管理数据，而且会将所有硬盘写零，确保数据无法恢复。

### 3.3.17 服务质量控制

我司统一存储软件平台的 QOS 机制允许用户对存储系统的带宽资源、计算资源、缓存资源等进行控制及调节，动态调配系统的资源来改变特定应用的服务级别，限制非关键应用的资源，优先保证关键应用能够获取更多的需求，保证关键应用的性能要求。

我司 QOS 机制通过 I/O 流量控制策略和 I/O 优先级调度策略来保障高性能应用的需求。

1.I/O 流量控制策略

I/O 流量控制策略是基于传统的令牌桶机制,通过针对某个 LUN 或文件系统的 IOPS 或带宽进行限制，来限制非关键业务流量的上限，避免这些应用出现突发流量过大，造成关键业务的性能需求无法满足。

根据 I/O 的方向分为控制读 I/O，写 I/O 和读写 I/O，根据控制的资源分为 IOPS 控制和带宽控制。

控制策略的主要实现机制为增加 I/O 延迟时间，从而降低非关键业务的流量。

2.I/O 优先级调度策略

I/O 优先级调度策略是基于 LUN、文件系统的优先级来进行分队列，关键业务具有较高优先级，非关键业务具有较低优先级，在资源紧张的情况下优先保障高优先级的业务。用户可以对 LUN 和文件系统配置为高、中、低三个等级。

系统包含高、中、低三个 I/O 队列，访问 LUN 或文件系统为高优先级时 I/O 进入高优先级 I/O 队列，访问 LUN 或文件系统为低优先级时 I/O 进入低优先级队列。在系统资源紧张时会优先处理高优先级队列中 I/O，从而尽可能保障高优先级业务。