**MK9000系列存储产品技术白皮书**

****

**北京智云创新科技股份有限公司**

**2019.1**

**目录**

[1 产品概述 3](#_Toc481699343)

[2 产品定位 3](#_Toc481699344)

[3 系统架构 3](#_Toc481699345)

[3.1 专业的存储管理系统 3](#_Toc481699346)

[3.2 领先的硬件架构 4](#_Toc481699347)

[4 产品功能 6](#_Toc481699348)

[4.1 高效的闪存/磁盘一体化混合存储优势 6](#_Toc481699349)

[4.2 性能均衡 7](#_Toc481699350)

[4.3 灵活卷管理 13](#_Toc481699351)

[4.4 多重校验的高可靠Raid技术 18](#_Toc481699352)

[4.5 异步容灾技术优势 22](#_Toc481699353)

[4.6 双机集群OmniHA模型 25](#_Toc481699354)

[4.7 重复数据删除 28](#_Toc481699355)

[4.8 维护管理 32](#_Toc481699356)

[5 产品特性 34](#_Toc481699357)

[6 MK9000系列产品技术规格 36](#_Toc481699358)

[7 名词解释 37](#_Toc481699359)

#

# 产品概述

MK9000存储系统是智云存储推出的一款双控制器架构多协议统一存储系统，可提供卓越的 ROI，降低用户的运营成本，高性能与高可靠性的完美结合，最大限度地减少业务增长所带来的风险。其融合统一架构能够以更快的速度支持更加广泛的 SAN 和 NAS 业务生产。

# 产品定位

MK9000系列产品适合各种对数据可靠性及性能要求较高的存储应用，如中小型数据库、中小型医院核心业务系统（HIS、LIS等）、新媒体制作、视频非线编、高清视频监控、高清制作网、高性能计算、私有云环境下云计算等，并广泛应用于政府、中大型企业、能源、交通物流、医疗、国防、新媒体、广电、影视特效制作等行业。

# 系统架构

## 专业的存储管理系统

智云存储MK9000系列存储产品是基于自主研发专业存储操作系统的IPSAN/FCSAN/NAS存储系统，通过新一代专用硬件可以让您在极短时间内搭建高稳定性、高性能的网络存储系统，构建NAS存储系统、ISCSI存储设备、FC存储设备一体化存储系统。

智云的MK9000系列存储系统集成了COS2.0存储系统包括（操作系统内核、各种硬件驱动程序和存储管理软件），不需要安装任何其他操作系统和管理软件。

**软件特点：**

* 内置功能丰富的管理界面；
* 支持ISCSI 1Gb/10Gb、FC 8Gb/16Gb、InfiniBand等主机端口；
* 简易的安装配置后，通过浏览器进行简单的配置和管理；

**软件特色功能：**

* 灵活的多协议存储架构，同时支持FC/FCoE/iSCSI/InfiniBand/CIFS/NFS/HTTP/FTP多协议访问；
* PCIE NVME 闪存卡与大容量磁盘的完美融合，满足用户在性能和成本上的最优均衡；
* 性能均衡，允许用户对不同数据卷设置优先级，并为每个数据卷设置不同的属性，包括：用户优先级、系统优先级和Cache策略。
* 灵活卷、重复数据删除、数据压缩等功能提高空间利用率；
* 通过数据快照、卷克隆、卷镜像、RAID多重校验等数据保护机制，有效保证用户数据安全；
* 基于ACL的细粒度的权限控制，可以达到文件级别的权限控制；
* 支持WORM（一次写入，多次读出）；
* 支持文件过滤，防止病毒扩散；
* 支持CIFS/NFS 统一命名空间集群。

## 领先的硬件架构

MK9000系列产品采用控制器和磁盘分离的领先硬件架构，实现了系统的高性能和高扩展性，可以满足复杂的应用环境和不断增加的容量需求。

* **Active-Active双控双活架构**

MK9000系列一款独立控制器架构产品，两个控制器模块之间通过万兆互连，实现缓存镜像功能；整个系统采用无线缆、模块化设计。

****

* **强大处理能力**

系统是基于Intel X86架构设计，每个控制器搭配2颗Intel最新一代多核CPU，兼容BROADWELL CPU 。最大支持32个DDR4 DIMM，可扩展至1024GB内存容量。端到端的12Gb解决方案，将系统I/O性能发挥到极致。

* **丰富的I/O接口**

每控制器板载2个iSCSI主机接口，1个千兆以太网管理网口（与IPMI共用），1个Console接口（RJ45方式，兼容CISCO标准）。每个控制器可选扩展32个10Gb iSCSI，或32个16Gb FC/8Gb FC。

后端磁盘扩展接口，每个控制器标配支持2个4x12Gb SFF8644 Mini SAS HD。

* **灵活的SAS Controller扣板**

主板采用支持Mezzanine Card设计方式，支持LSI 3008和PMC8068两款SAS 控制芯片扣板，客户可以根据自己的需求，选择不同的SAS Controller扣板。

* **完备数据保护**

该系统中设计了1＋1 BBU冗余电池备份系统，满足在AC意外掉电后，对数据进行保护。

* **节能环保**

电源模块采用1+1冗余电源，基于PID闭环控制的智能风扇控制以及高转换效率的DC-DC开关电源等，极大降低系统功耗。

# 产品功能

## 高效的闪存/磁盘一体化混合存储优势

采用海量和安全的128位存储专用系统（OFS）是一个革命性的全新的文件系统，它从根本上改变了文件系统的管理方式，这个文件系统的特色和其带来的好处至今没有其他文件系统可以与之媲美，OFS被设计成可升级并易于管理的。其内置的OmniCache功能，能够结合最新的闪存和通用磁盘的优点，做到最佳的性能、容量、价格的最佳匹配。保证大容量数据采集、制作和编辑业务的高效运转。工作如图：



用5%的闪存，解决95%传统存储的并发瓶颈问题。

## 性能均衡

**技术原理**

OmniShare允许用户对不同数据卷设置优先级，并为每个数据卷设置不同的属性，包括：用户优先级、系统优先级和Cache策略。

**OFS操作**

来自任何访问协议（NFS/CIFS/FCP/ISCSI）的读或写操作请求进入MK9000存储系统的操作系统后，都会被转化成为OFS的读写操作。同样，一个来自系统本身的请求也会最终被转化为OFS操作，这些操作最终由OFS完成。

MK9000存储系统把每个OFS操作分类成用户发起或者系统发起。举例来说，来自与客户端的请求被认为是一个用户操作，而一个OmniMirror操作则被认为是一个系统操作。

**处理队列**

OmniShare为每个设置了优先级的数据卷保持不同的数据处理队列。这些队列仅当OmniShare服务启动时才发生作用，当OmniShare服务停止时，数据将不会进入这些处理队列而是直接被送去OFS处理。

MK9000存储系统会维护一个缺省队列。当OmniShare服务启动时，所有未设置优先级的数据卷的数据会被送到这个缺省队列，而设置了优先级的数据卷的数据则会进入OmniShare为其保持的专用队列。

vol1 和 vol2设置了优先级，因此各自拥有独立的处理队列。

**用户操作与系统操作**

OmniShare区分用户操作与系统操作，可以为每个数据卷分别配置其优先级。

OmniShare依据数据请求的发生源来确定一个读写操作是用户请求或系统请求。如果该请求来自数据访问协议，则OFS认为这是一个用户操作，而其他所有操作则被认为是系统操作。

下表给出了一些主要的用户与系统操作：

|  |  |
| --- | --- |
| **用户操作** | **系统操作** |
| 􀂃NFS 􀂃CIFS 􀂃iSCSI􀂃FCP 􀂃HTTP 􀂃FTP  | 􀂃OmniHA􀂃OmniCache􀂃OmniMirror􀂃OmniVol􀂃OmniClone􀂃OmniDedup􀂃OmniCompress􀂃OmniSplit􀂃OmniSnap􀂃OmniRestore􀂃OmniMrmniro􀂃NDMP  |

**Cache策略**

与其他系统相似，MK9000存储系统同样使用Cache缓存数据，用于数据的快速访问。正常情况下，当Cache已经满并且需要空间来缓存新数据的时候，MK9000存储系统使用一个改进的最近最少使用（least-recently-used）的Cache算法来决定哪些数据需要被清除出Cache以腾出空间。

OmniShare能够改变这种缺省的Cache策略。通过配置，OmniShare可以决定哪些数据需要保持在Cache中，而哪些数据空间可以被重用。

OmniSharecache策略，如果依据应用负载正确配置，可以显著增强系统的负载，这种Cache策略是按每数据卷来配置的。

**操作方式**

OmniShare有下列操作方式，

1. OmniShare服务停止或未运行

缺省状态下，OmniShare服务未运行。系统行为没有任何改变。

2. OmniShare服务启动，没有为任何数据卷设置优先级策略。

当OmniShare服务启动运行时，OmniShare为每个数据卷提供相同的优先级，用户操作与系统操作也具有相同的优先级，OmniShare仍然使用缺省的Cache策略。

3. OmniShare服务启动，并且有为数据卷设置优先级策略。

当OmniShare服务启动运行，并且有为数据卷设置优先级策略，OmniShare会依据这些优先级策略对读写IO依据优先级顺序进行处理，并且根据设置改变Cache策略。

**OmniShare与系统资源**

OmniShare依据数据卷优先级设置与Cache策略自动调配系统资源的使用，系统管理员不再需要做其他配置来管理OmniShare所调配的系统资源。OmniShare依据优先级设置来调整系统资源，使其优先服务于高级别的应用，但是并不确保系统资源的绝对可用性，因此，这是Cos(Class of Service)服务，而不是Qos(Quality of Service)服务。

**OmniShare操作示例**

正常情况下，OmniShare没有启动的情况下，IO操作的顺序，OFS根据IO到达的先后次序进行处理。

当对数据卷进行优先级设置后，OmniShare会根据优先级设置对进入的IO进行重新排队，再将排序后的IO送到OFS进行处理，经过OmniShare排序后的IO处理次序。OmniShare并没有改变OFS本身的数据处理方式，只改变了读写IO提交给OFS的次序。可以认为OmniShare是一个前置的优先级队列处理。

**技术特点**

OmniShare赋予了MK9000存储系统下列能力，

* 不同数据卷之间的优先级设置；
* 每数据卷的用户操作与系统操作的优先级设置；
* 每数据卷的cache策略设置；

这些能力让存储管理员在系统超负荷的情况下，可以调整与优化系统资源的使用，是存储系统达到更高的CPU利用率和对关键应用更好的响应。

**技术优势**

**简化存储管理**

* 允许更高效的存储整合，从而减少了需要管理的存储系统；
* 提供了整合环境下简单的性能管理机制；
* 即可以使用CLI，也可以通过MK9000存储系统 API进行管理；

**降低成本**

* 在不牺牲关键应用的前提下，允许整合更多的容量并提高存储的CPU利用率；
* 不需要额外的硬件或软件；
* 不需要软件许可，免费使用；

**灵活性**

* 可灵活定制以满足不同应用环境下的负载情况。

## 灵活卷管理

OmniVol是一个划时代的新技术，所有的卷被视为逻辑的数据容器，同时可以在不破坏底层物理存储结构的前提下被独立的设置、调整大小、管理和移动。

MK9000存储系统OmniVol技术提供了真正的存储虚拟化解决方案，能够缩减开销和资本费用，减少业务中断并降低风险，同时还具有很高的灵活性，可以快速方便地适应企业不断变化的需求。OmniVol技术可以自动集中存储资源，便于用户在一个大型磁盘池中创建多个灵活的卷。有了这一灵活性，就可以简化操作，最大限度地提高利用率和效率，并可以快速、无缝地进行修改。利用MK9000存储系统，用户可以随时随地根据需要以最低成本增加存储容量，并且无需中断业务运作。

如下图所示，一个存储池被定义为给卷分配空间的许多磁盘的池。从管理者的角度来看，卷仍保持着对主要项的数据管理。但由于是对管理者透明，MK9000的卷现在面向的是逻辑项而不是传统的物理存储了。

因此MK9000的卷不再受限于它们存在的磁盘的限制了。用户可以在一个存储的“存储池”中根据需要定义一个相应空间的卷，不再由其创建卷所在的磁盘个数硬性规定。一个卷也可以不需要停机任意收缩或者扩展。同时可以在存储池所拥有的所有物理磁盘上实现条带化。对于那些对性能敏感的应用来说，MK9000的卷比那些由磁盘个数限定大小的卷拥有更好的性能。



* **性能的提高**

在MK9000中，RAID组被整合起来创建出一个存储池。由于卷仍然是存储管理的基本单位，它将跨越存在于一个MK9000存储系统中组成存储池的所有磁盘，同时在这个大的存储池上也可以存在多个卷。 这将使该卷充分利用所有磁盘的并行性能，满足在系统中某些比较繁忙的卷对于性能的需求。OmniVol是灵活的，因为底层的存储的物理结构不需被预分区。



图：OmniVol比传统卷拥有更好的性能

* **OmniVol的容量担保**

在MK9000中引入了一个新的存储管理的概念——担保（guarantees）。担保的概念不同于以前用户们在使用iSCSI和Fibre Channel中所熟悉的“空间预留（space reservations）”。担保扩展了管理员的权限，使其在卷或者文件创建前可以决定预分配的策略。使其充分贯彻所谓的“自动精简配置（thin provisioning）”的概念。

担保，在卷一级设置，用以决定在存储池上给一个OMNI volume预分配多少空间。当用户在一个存储池上创建一个OmniVol，将指定其容量，同时也可以指定担保的类型。

有三种担保的类型：

* 卷（Volume）：卷担保类型确保在存储池上为OMNI volume分配的总空间总是可用的，这也是OMNI volume的默认设置。
* 文件（File）：在文件担保类型中，存储池确保为可重写的LUN或者文件保留的空间总是可用。
* None：对于一个无预留空间担保类型的OMNI volume来说，不管给这给卷中的LUN设置了多少预留空间，当容纳它的存储池没有足够的可用空间时，对预留空间的LUN的写操作都将失败。
* **灵活的容量规划**

对于一个OmniVol的大小本质上是没有约束的，同时volumes可以动态的调整大小。管理员可以将volumes作为一个强大的工具为不同的用户、组和项目分配和提供存储资源。举例来说，假设一个数据库比原先预计增长快很多时，管理员可以在系统运行时随时重新配置相关的volumes。重新分配存储资源的过程不需要任何宕机过程，而且它是对用户透明的。

当需要额外的物理空间时，管理员可以通过添加指定磁盘给存储池以提高其容量，新的磁盘将成为存储池的一部分，同时其容量和I/O带宽将对存在于该存储池中的所有的volumes生效。

所有存在于同一个存储池上的volumes的总容量也可以超过存储池的实际物理总容量。增加某一个OmniVol容量时也不需要改变存在相同存储池其它的OmniVol的容量或者存储池本身的容量。

* **OmniVol的技术限定**
* MK9000中最大的存储池个数（包括存储池和传统卷）= 1000。
* MK9000中最大的卷的个数（包括传统卷和灵活卷）= 5000。
* 最大的存储池/传统卷的大小 = 1024TB。
* **OmniVol的技术优势**
* 灵活的尺寸：MK9000存储系统的灵活卷的最小颗粒度是4KB，可以精确的适合各种大小的存储应用的需求。其它的系统需要卷或者文件系统的大小要基于底层物理或者逻辑磁盘的整数倍而定，显然会浪费数百兆甚至数G以至更多的物理容量。
* 灵活地改变尺寸（扩展和收缩）：MK9000存储系统OmniVol技术允许OFS文件系统在线且安全的进行扩展和收缩，以使文件系统精确的适合用户的数据需求。其它的存储提供商基本能够提供不停机的容量扩展机制，但它们却不提供类似MK9000存储系统OmniVol的无缝和简单的容量扩展，并且不能提供在线且安全的文件系统收缩的功能。
* 空闲空间池：由于MK9000存储系统OmniVol可以像扩展一样简单的实现容量的收缩，已分配给OmniVol的未使用空间可以非常简单且安全的从OmniVol移出并重新分配给其它即将被填满需要更多空间的OmniVol。通过MK9000存储系统OmniVol技术，未使用或者未分配的空间在存储池容器层被视为一个池（pool），可以被任意的OmniVol进行配置使用或者从OmniVol移出至池中。
* 瘦容量提供（Thin capacity provisioning）：MK9000存储系统OmniVol可以实现瘦容量提供，即OmniVol仅仅占用已经被写入实际数据的物理存储空间，其他在OmniVol中已定义但未被使用的容量并不占用实际的物理磁盘空间。该功能在其它存储系统中是无法实现的，无论采用的是文件、LUN或整个文件系统。
* 磁盘聚合提高性能：MK9000存储系统OmniVol技术使用一个新的智能数据容器对象——存储池。通过使用存储池，OmniVol可以充分利用一个含有大量磁盘的条带化的全部I/O对各种大小和种类的数据对象进行访问。

简单化的存储系统管理：通过使用存储池，MK9000存储系统OmniVol实际上已经消除了在存储产品上令数据管理员头痛的所有低级别的条带化的管理工作。存储池将注意力集中于数据的管理，而不是磁盘的管理，将他们从一味的磁盘驱动器的配置的噩梦中解脱出来。

## 多重校验的高可靠Raid技术

传统的单一奇偶校验RAID技术，能为单个故障磁盘驱动器提供保护。需要警告的是，在重建故障磁盘时必须确保不能发生其他磁盘故障，而且不能在读操作过程中出现不可修复的误码。如果在重建过程中出现上面任何一种事件，那么保留在RAID阵列或卷中的部分或所有数据就可能会丢失。随着现代更大的磁盘存储介质的出现，由于磁盘容量显著增加，而误码率仍是原来的水平，这就使得出现不可修复的误码的可能性相当高。因此，传统的单一奇偶校验RAID保护数据的方法已经超出了其保护极限。更高一级的RAID数据保护已经发展到RAID双奇偶校验，也称作RAID PX，它可以应用于整个网络应用数据存储产品线。

智云存储MK9000存储系统推出了双奇偶校验RAID，叫做RAID PX。本节给出RAID PX的概览和RAID PX如何在各种磁盘故障情况下动态提高数据的容错性。其他所涉及的关键领域包括RAID PX的是免费的、无特殊硬件要求，以及能够非常容易的将现存的RAID 5卷升级到RAID PX。这里将借助对双磁盘故障的恢复情形来说明RAID PX如何做到允许卷既能不间断地提供数据，又能同时对两块故障磁盘上损失的数据进行重建。

在最基本的层面上，RAID PX为卷中的每个RAID组增加了第二个奇偶校验磁盘。RAID组是建立卷的基本结构。每一个传统的MK9000存储系统RAID PX组都有一些数据磁盘和一个奇偶校验磁盘，而一个卷通常会包括一个或多个RAID PX组。而RAID PX卷上的奇偶校验磁盘则通过RAID PX组上的磁盘对行奇偶校验进行存储，额外的RAID PX奇偶校验磁盘则通过RAID PX组上的磁盘存储对角线上的奇偶校验。通过RAID PX上的这两个奇偶校验条，一个传统的水平奇偶校验和一个对角线奇偶校验，即使同一RAID组上的两个磁盘发生故障时也能得到数据保护。

在RAID PX中，仍然使用了传统的RAID PX水平奇偶校验结构，并且这种结构已成为RAID PX结构的子集，换句话说，RAID PX并没有改变RAID PX在MK9000存储系统存储上的工作方式。数据通过每一行的奇偶校验计算被写入水平行，并被视作双奇偶校验中的行，这个同样的过程仍会保留到RAID PX中。实际上，如果是单一磁盘故障，或是从坏数据块中的读取错误，或者出现误码，那么RAID PX的行奇偶校验方法就是进行数据恢复的唯一手段，无需采用RAID PX。在这种情况下，RAID PX的对角线奇偶校验组件仅仅是行奇偶校验组件周围的保护层。

**RAID PX水平行奇偶校验**

下列图表说明了在传统MK9000存储系统RAID PX解决方案中使用水平行奇偶校验的方法，也是您了解RAID PX和双奇偶校验的第一步。



图1：RAID P1水平行奇偶校验

图表代表了传统的RAID PX组，使用包括4个数据磁盘（前4列，标注为“D”）和一列奇偶校验磁盘（最后一列，标注为“P”）的单向奇偶校验。上面图表中的行显示了实施传统MK9000存储系统RAID PX时使用的标准4KB块。上面图表中的第二行在每个4KB 块和行中数据的奇偶校验计算中添加了一些样例数据，随后存储在奇偶校验磁盘上相应的块中。在这种情况下，计算奇偶校验的方式是向每个水平块中添加值，然后将这些值的和存储为奇偶校验值（3 + 1 + 2 + 3 = 9）。实际上，奇偶校验是通过专门的OR（XOR）过程进行计算，但其它情况也与此例非常相似。如果需要从单一故障中重建数据，则产生奇偶校验的过程将与此相反。例如，当RAID PX恢复上面第一列的数据值3时，如果第一个磁盘发生故障，就会从奇偶校验中存储的值中减去剩余磁盘上的值（9- 3 - 2 - 1 = 3）。这个通过单一奇偶校验RAID重建的例子，应该有助于从概念上理解为什么能够对单一磁盘故障提供数据保护。

**添加RAID PX双奇偶校验条**

下表将一个以蓝色阴影块表示的对角线奇偶校验条和一个在第六列中以“DP”表示的第二个奇偶校验磁盘添加到前面的RAID PX组中。这表示RAID PX结构是RAID PX水平行奇偶校验解决方案的超集。



图2：RAID PX双奇偶校验条

在这个例子中对角线奇偶校验条通过相加的方法计算，并且存储在第二个奇偶校验磁盘上，而不是上面所说的实际使用的XOR（1 + 2 + 2 + 7 = 12）。此时，需要注意的最重要的问题是，对角线奇偶校验条包括一个来自行奇偶校验的元素作为其对角线奇偶校验和的一部分。RAID PX将把最初在RAID PX结构中的磁盘（包括数据和行奇偶校验磁盘）视作相同的磁盘。下一个图表为每个块加入了余下的数据，并且建立了相应的行和对角线奇偶校验条。



图3：RAID PX

从上图可以明显看出，在RAID PX中，对角线奇偶校验条都在行奇偶校验结构的边缘。在这个例子中，对于RAID PX对双磁盘故障进行恢复的能力，有两种重要情况可能不是很明显。第一种情况是，每一个对角线条会错过唯一的一个磁盘，但是每个对角线错过的磁盘都不一样。第二种情况的结果是，在第二个对角线奇偶校验磁盘上会产生或存储一个没有得到奇偶校验的对角线条。在这个例子中，被遗漏的对角线条是白色的块。在之后的重建例子中会明显的看出，遗漏一个对角线条不会影响RAID PX在双磁盘故障中恢复所有数据的能力。

值得重视的是，在本例中介绍的许多同样的RAID PX对角线奇偶校验都是真实的存储情况；其中涉及到在RAID组中的许多磁盘上，以及通过RAID PX组水平写入成数百万的数据行。并且，通过上面较小的例子我们很容易地例证了RAID PX，而且不管RAID 组中的有多少磁盘，对于大尺寸的RAID组，其恢复情况也与此相同。

由于RAID PX可以通过两种方式对双磁盘故障进行所有数据的恢复：一种是使用数学原理并进行证明，另一种则是简单地通过双磁盘故障并随后进行恢复。上述示例使用后面的方法来证明RAID PX的双奇偶校验保护的概念。

## 异步容灾技术优势

数据是现代企业正常运转的重要依据和企业发展的宝贵资源。如何确保数据安全成为了企业是否具有核心竞争力的体现。

企业需要在发生自然或人为灾难、操作员出错或是技术和应用出现故障的情况下，保护数据并快速进行恢复。用户还需要一个有效的方式将数据复制到远程位置，如果没有有效的数据保护和发布策略，运作过程可能会被迫中断，导致严重的经济损失。

OmniMirror软件具有强大的功能，而且易于使用和管理，它满足了当前的企业对灾难恢复和数据分布解决方案的需要。通过在 LAN 或 WAN 上高速复制数据，OmniMirror软件将尽可能地为关键应用程序提供最高的数据可用性和最快的恢复速度。OmniMirror架构示意图如下：

**…...**

Win

Linux

UNIX

**变化的数据块**

**服务器**

*上海*

*北京*

**LAN/WAN**

**…...**

Win

Linux

UNIX

**OmniMirror技术优势**

OmniMirror技术将数据镜像复制到一个或多个东华软件网络存储上，OmniMirror不断地更新镜像数据，以确保数据是最新的，并且能够用于进行灾难恢复、减少磁带备份、发布只读数据、在非生产性存储系统上进行测试、执行联机数据迁移等等。如果您的企业分布在不同的地点OmniMirror可以能够将同一数据发布到所有地点。通过自动更新这些数据，并支持对镜像数据的本地访问方式，OmniMirror可以大大提高员工的工作效率。

**节省宝贵的网络带宽资源**

OmniMirror软件具有许多节省带宽的功能，可以降低数据复制和灾难恢复的基础设施成本。您可以先执行一次整卷数据传输，以后每次数据同步均为基于4KB数据块进行增量传输，您只需要通过网络对新的数据块和更改的块进行增量更新。由于只需复制所有数据的一小部分，OmniMirror大大降低了对网络带宽的需求。此外，OmniMirror在数据传输过程中会设置检查点，如果系统出现故障，传输会从最近的检查点重新开始。OmniMirror 还会执行智能的重新同步过程，这样，在从镜像损坏的状态或不同步的状态进行恢复时，无需传输整卷数据。如果在应用程序测试期间修改了镜像副本上的数据，可以将生产系统上的新数据块和更改的数据块复制到镜像副本，使镜像副本快速实现与生产数据的重新同步。

**灵活的配置**

OmniMirror可以方便地部署到任何具有足够的带宽来处理数据传输的网络基础设施中。它能够支持多种传输方式（FC 和 IP），这一点确保可以更好地利用现有设备，并且还因能够实现路径间的容错而提高了可用性。

OmniMirror允许客户选择适当的同步级别（半同步和异步），从而实现了较高的灾难恢复保护级别。例如，如果选择 “半同步”允许客户根据站点的需要，确定有多少 I/O 操作可以失去同步，或者是复制站点可以在多长时间内与数据源失去同步。而“异步”选项则允许您根据需要（每分钟、每小时或每天）任意安排传输过程。您可以设置最适合每一站点的频率。时间安排很容易修改，而且修改后可以立即生效。您还可以为源系统和镜像系统选择不同的存储系统配置。

OmniMirror目标卷可以作为其他目标的源卷，而且每个“镜像对”都可以有自己的时间安排以满足各个站点的特定需求。如果需要远距离复制数据，级联镜像是非常理想的解决方案。例如，如果是从北京将数据复制到南京、广州和西安，您可以将数据从北京复制到南京，然后使用低成本链接将数据从上海复制到广州和西安。

**易于使用，经济合算**

用户在几分钟内就可以设置好OmniMirror，并能通过易于使用的GUI图形用户界面进行管理和操作，存储系统支持OmniMirror软件，您不需要再安装其他软件。

**OmniMirror技术方案组成**

1）快速数据复制和容错——将主站点发生故障的宕机成本降至最低。

2）访问镜像数据——减少磁带备份，让您在灾难恢复上的投资物超所值。

3）卷复制——镜像选定的数据集，大大降低了对网络基础设施的要求。

4）选择复制过程的同步级别——采用一个产品即可控制复制的频率（异步、准同步）。

5）更有效地利用网络——降低了数据复制和灾难恢复的成本。

6）设置简便——几乎不需要增加 IT 资源；可以经常测试灾难恢复计划。

**OmniMirror可以用于：**

1）灾难恢复——将数据镜像到远程位置以实现容错。

2）数据分布——使用级联或多跃点镜像方式，将数据集发送到世界各地。

3）远程数据访问——应用程序能够以只读模式访问镜像数据。

4）联机数据迁移——将数据迁移导致的宕机时间减到最少。

5）数据复制——使用镜像数据可以进行隔离测试。

6）负载均衡——通过访问镜像数据，在更多的客户端之间分摊负载。

## 双机集群OmniHA模型

智云存储的OmniHA为关键业务环境提供可靠的高可用数据服务。OmniHA安装在一对MK9000存储系统上，它通过将不可用A控制器的数据服务转移到集群中的另一个B控制器来确保数据可用性。通常，数据服务的转移对最终用户和应用程序是透明的，数据服务将快速恢复，不会出现明显的业务运转中断。

OmniHA提供大于 99.999% 的数据可用性。任何对运行时间有严格要求的企业、工作组或电子商务应用都可从中受益。OMNIHA的管理任务简单、直观，并且易于使用，可以降低管理开销并减少人为操作错误。总而言之，它们可以进一步提高整个环境的总体可用性。

MK9000存储系统支持通过群集技术（OmniHA），实现存储系统间互为热备失败切换的高可用功能，从而提供更强的防范硬件故障的能力。集群的控制器之间通过内部适配器和线缆加以连接。并配置为共享一套光纤通道磁盘阵列和网络连接。MK9000群集的主控单元（存储主机）通过心跳线互相侦测对方的健康状态，当对方的状态出现异常时（如：主控单元掉电、网络连线中断），可以接管对方的读写操作（包括对方的硬盘组、尚未写盘的数据、IP地址和MAC地址），继续数据的IO读写，该群集的失败切换接管（Cluster FailOver）操作完全由存储系统自动完成，对于前端的服务器和用户完全透明。



OmniHA存储模型

在这种配置下，每一个控制器存储引擎对一组磁盘负主要责任，并能分别独立工作。MK9000集群结构采用active-active的工作模式。在正常操作期间，两个控制器分别对它们各自的磁盘阵列加以操作并提供数据服务。当集群结构中的一个控制器发生故障的时候，另一个控制器机头将会自动接管发生故障的机头，并且响应那些访问故障机头所对应盘阵的请求。在整个故障转移操作过程中，接管控制器的数据服务不会受到任何影响，并且完全可用。接管控制器将一直保持这种双数据服务模式，直到管理员采取措施将数据服务恢复到其原始状态。整个故障转移过程是自动完成的，不需要任何人为操作。

| **功能** | **优点** |
| --- | --- |
| 自动执行故障转移，并向管理员发出通知 | 确保不可用控制器的数据服务不会中断，并通知管理员发生了故障转移。 |
| 手动故障转移和故障恢复 | 可以有计划地对控制器进行维护，而不会影响数据可用性；管理员完全可以控制何时将集群控制器恢复到其正常状态。 |
| 故障转移完全透明 | 最终用户不会受到故障转移的影响；客户端在故障转移后可以继续以相同的方式访问数据。 |
| 双活动配置 | 两个集群控制器都可分别提供有用的服务；集群操作不需要空闲资源。 |
| 简单易懂的设计和实施 | 管理员在不到 30 分钟的时间内就可以安装并配置好集群控制器；需要的日常维护很少。 |
| 长距离集群 | 集群控制器的地理位置可以非常灵活。 |

OmniHA的优点

除了故障时的自动接管外，OmniHA也同时支持手工接管，这项功能对于计划中的存储设备停机维护是十分有帮助的，可以在设备维护期间仍然提供不间断的数据服务。管理员最常利用OMNIHA来预定的存储设备维护和升级期间通过手动启动故障转移操作来确保数据服务不间断。在配对存储控制器接管了数据服务后，管理员就可以执行计划的任务，而不会影响到最终用户。当故障控制器可以恢复正常的数据服务时，管理员可以选择在最佳时间和条件下执行故障恢复操作 — 将数据服务从接管控制器A转移回原控制器B。也可以在OmniHA检测到故障控制器B可以恢复操作时，由OmniHA向原控制器执行自动故障恢复。这样，管理员即可进一步减少额外的人为操作，同时增加系统正常工作时间。

应用集群软件提供了一个经实践证明的成熟解决方案以确保应用的可用性。结合OmniHA，应用与数据联合集群解决方案可以为关键业务环境提供空前水平的总运行时间。智云存储已经与业内领先的集群应用提供商建立了战略合作伙伴关系，可以提供完全集成、完全合格的集群解决方案。

## 重复数据删除

**技术原理**

MK9000存储系统允许用户在任何存储环境中使用数据重复删除技术，包括生产应用环境、与备份归档环境。这将帮助用户以更少的存储空间来存储更多数据。

MK9000存储系统的重复数据删除技术OmniDedup与应用无关，无论是文件存储、数据库、SAN访问方式或者NAS访问方式，OmniDedup都能够找出相同数据块并予以合并。

在存储系统中实现OmniDedup的关键是采用了Dedup卷。Dedup卷是一个特殊的灵活卷，它允许卷中的包含相同内容的数据块共享物理存储空间。而MK9000的核心技术OFS，在结构上支持这样的共享，从而能够最终实现存储空间的节省。如下图所示，OFS的数据块索引结构支持多个索引指针指向同一个物理数据块，从而实现空间共享。



为跟踪这些指向自己的索引指针，每一个数据块还需要保持一个计数器，当指向自己的指针增加或者减少，计数器要相应地增减，以反应指针的变化。

OmniDedup在Dedup卷内数据块级别实现重复删除，下图给出了其工作原理的逻辑示意



Dedup的工作流程大致如下：当在一个数据卷上启用Dedup之后，存储系统仍然会以正常的方式保存新写入的数据块。系统为每个新数据块生成一个对应的指纹信息，用于与数据卷中其他数据块的指纹信息进行比对。如果发现两个数据块的指纹信息相同，则会对这两个数据块进行完全的字节比对，如果仍然相同，则存储系统就会丢弃重复的数据块并且回收对应的存储空间。

Dedup进程以以下三种方式之一运行：

设定计划表，进程将按时启动运行；

通过命令行手动运行；

当累计有20％的新数据被写入到数据卷时，自动运行；

下图给出了Dedup进程的不同阶段的状态，



Gathering：对数据卷中所有数据进行扫描并生成指纹信息，该过程仅对在已有数据的数据卷上激活的Dedup功能时才发生；如果在激活Dedup功能后写入的新数据，在其写入的同时系统指纹信息已经生成。

Sorting：扫描指纹信息库并进行比对。

Deduplicating：根据指纹信息和字节比对结果去除重复数据块，并修改数据块指针。

Checking：检查指纹信息库，并去除其中的冗余信息（即：已合并的数据块的指纹信息）。

**技术特点**

粒度细，在数据块级别；

作为后台进程操作，并可配置成自动运行，手动运行或计划表运行多种方式；

对应用透明，可用于数据中心的各个层面；

可以对卷中已有的数据进行重复删除操作；

US全系列支持；

对于不同的应用环境，节省的空间比可能达到20:1甚至更高;

支持双控制器cluster结构；

**技术优势**

**与存储协议无关**

OmniDedup对数据卷中的数据块进行比对操作并删除重复数据，与通过何种存储协议写入数据无关，无论是FCP、NFS、CIFS、ISCSI的数据均可以。因此从Dedup的操作层面看，对应用完全透明。

**可直接用于生产环境**

OmniDedup技术是OMNIOS系统的一个软件特性，只要一个license即可激活使用（而且免费）。因此，可以方便地用于生产环境，当然也可以用于备份、归档等应用环境。特别是在服务器虚拟化等存储消耗型应用中，使用OmniDedup可以节省大量存储成本和空间。

**性能损失最小**

OMNIOS存储系统的OmniDedup技术与OFS结构紧密集成，因此具有极高的效率。不需要复杂的哈希算法和数据表查询，相反OmniDedup能够利用到OFS自身的数据结构特性来生成和比对指纹信息，重定向数据块指针，以及释放冗余数据块空间，因此将对性能的影响减至最小。

**与SnapMirror结合，减少WAN流量**

OmniDedup可以与OMNIMirror结合，将OmniDedup数据卷的数据OMNIMirror到远端MK9000。由于数据在源端已经OmniDedup过，通过网络传输的数据量将大大减少，从而进一步降低WAN带宽的需求，节省成本并且能够实现更好的RPO。

## 维护管理

* **技术原理**

每个MK9000存储系统设备上都提供内置的 MK9000存储系统GUI工具，IT 管理员可以通过 Web 浏览器进入一个一致、易用的图形用户界面 (GUI)，进行日常管理。

管理员可以从任何地方，从任何计算平台远程设置并控制所有 MK9000存储系统 存储设备，而无需中断关键业务操作。安装后，管理员可以有效地监视状态、满足增加存储容量的请求，并对文件系统配置进行更改。管理员还可以在不中断用户访问文件系统的情况下扩展卷，或在连接的 Windows域中添加新的通用 InternOmni 文件系统 (Common InternOmni File System, CIFS) 共享。

MK9000存储系统GUI是用于执行从安装到完成配置的管理任务的主界面。利用 MK9000存储系统View，管理员不必再学习特殊的语言或命令行界面来配置和维护 MK9000存储系统 存储设备。只需一个 Web 浏览器，管理员就可以进入易用的图形用户界面，并且可以通过浏览各个屏幕来执行任务。界面上将始终显示一个浏览树，以便于在各个任务间切换。

OmniView 将 MK9000的安装过程简化到只需执行几个简单的步骤。通过控制台或动态主机配置协议 (Dynamic Host Configuration Protocol, DHCP) 服务器将一个 IP 地址分配给 US 后，管理员可以通过安装向导快速完成基本配置设置。

 由于 OmniView 采用基于 Web 的方法，因此无需学习和使用复杂难懂的命令行界面执行管理任务。通过一个图形用户界面，管理员可以配置和监视任何系统资源或服务：

文件系统： NFS 与 CIFS 实施的磁盘、卷限额；OmniView卷报告显示所有卷的当前配置状态。

数据保护：OMNIShot和OMNIMirror工具、网络数据管理协议 (NOmniwork Data Management Protocol, NDMP)

可用性：集群容错软件

联网：接口卡；HTTP、DNS（域名服务器）和 NIS（网络信息系统）服务器

管理：安全功能、autosupport和OmniView软件、SNMP 实用程序、资源监视工具

OmniView 的安全模式确保所有管理和配置命令都严格保密，无论通过专用网络或公共 InternOmni 执行。安全模式需要OMNICLI软件，该软件在 上实施安全套接字层 (Secure SockOmni Layer, SSL 3.0) 服务，用于进行用户验证和保密通过 HTTPS 处理的命令。

* **技术特点**

完整的基于Web的管理工具集，提供完善的存储设备配置管理功能，监视功能。支持基于SSL的安全管理。

* **技术优势**

在简单方便的图形化Web管理界面中提供易学易用的基本管理功能。低成本的管理工具。

# 产品特性

* **灵活的多协议架构**

基于软件定义储理念设计，可同时支持FC/FCoE/iSCSI/InfiniBand/CIFS/NFS/HTTP/FTP多协议访问，后端具备SAS 12Gb硬盘扩展接口，可支持SAS硬盘、SATA硬盘及SSD固态硬盘，满足业务弹性发展，简化业务部署，提升存储资源利用率。

* **双控双活，有效保障业务连续性**

MK9000系列存储系统部件采用双冗余设计和A/A工作模式（Active-Active Mode）。在正常情况下，2个部件同时工作，处理存储业务。当其中1个部件出现故障或离线时，另外1个部件就会及时接管其工作，不影响现有任务。提供99.999%的高可用性，有效地保障业务连续性，提升企业的核心竞争力。

* **SSD自动数据分层，助力企业应用加速**

SSD与大容量磁盘的完美融合，数据自动分层存储，既能利用SSD的高性能优势，同时又能兼顾大容量磁盘的容量优势，二者合一将存储的性能与容量发挥到极致，满足用户在性能和成本上的最优均衡。

* **领先一代的存储硬件平台，全面提升存储性能**

MK9000系列存储产品采用Intel全新多核处理器，支持16Gbps FC/10Gbps FCoE/56Gbps InfiniBand等主机接口；内部采用新一代PCI-E 3.0总线，并采用12Gbps SAS 3.0高速硬盘接口，可提供高达10GB/s的系统带宽，满足视频、大文件等高带宽应用场景所需；提供百万级IOPS，性能及规格全面领先。

* **高可靠性**

采用创新的RAID PX技术，既提供三重数据校验技术，同一RAID组三块磁盘同时损坏的情况下数据不丢失，且实现了硬盘自动负载均衡，当存储系统某一硬盘发生故障时，硬盘域内的所有正常硬盘参与数据重构，而且仅重构业务数据，数据重构速度相对传统RAID提升20倍，极大降低了多盘失效概率。

* **丰富的软件功能，帮助企业轻松管理海量数据**

MK9000系列具有丰富的软件功能，包括灵活卷技术、高效数据快照技术、一键快照恢复技术、数据克隆技术、在线据据重删技术、在线数据压缩技术、法规遵从WORM技术等功能，借助这些功能让企业轻松应对数据爆炸式增长所带来的烦恼，管理海量数据变得如此简单，帮助企业更专注于自身的业务发展。

# MK9000系列产品技术规格

|  |
| --- |
| **系统硬件规格** |
| 　 | **MK9080** | **MK9060** | **MK9040** | **MK9020** |
| **控制器架构** | 多控器架构， | 多控器架构， | 多控器架构， | 多控器架构， |
| 最多24个控制器 | 最多24个控制器 | 最多24个控制器 | 最多8个控制器 |
| **控制器外形规格** | 双机箱 HA； | 双机箱 HA； | 双机箱 HA； | 双机箱 HA； |
| 两个4U机箱中有2个控制器，共8U | 两个2U机箱中有2个控制器，共4U | 两个2U机箱中有2个控制器，共4U | 两个2U机箱中有2个控制器，共4U |
| **最大盘位数 (双控制器)** | 4320 | 2880 | 1920 | 960 |
| **系统缓存 (随控制器数量扩展）** | 768GB-9216GB | 512GB-6144GB | 256GB-3072GB | 128GB-1536GB |
| **最大前端主机接口数 (双控制器)** | 80 | 56 | 56 | 56 |
| **最大I/O模块数 (双控制器)** | 20 | 14 | 14 | 14 |
| **端口类型** | 1Gbps/10Gbps Ethernet、10Gbps FCoE、8Gbps/16Gbps FC、56Gbps InfiniBand、SAS3.0（后端接口，单端口4\*12Gbps） |
| [**支持的磁盘扩展柜**](http://www.netapp.com/cn/products/storage-systems/disk-shelves-and-storage-media/index.aspx) | ED4224A（2U；24 个驱动器，2.5 英寸 SFF） |
| ED4324A（4U；24 个驱动器，3.5 英寸 LFF） |
| ED4360A（4U；60 个驱动器，3.5 英寸 LFF） |
| **系统软件规格** |
| **系统平台** | OmniCOS v2存储专用操作系统 |
| **文件系统** | OFS事务型文件系统,无单卷容量限制,在线动态伸缩卷尺寸,用户和目录级的配额管理 |
| **RAID保护** | RAID0,RAID1,RAID10,RAID P1(RAID5),RAID P2(RAID6),RAID P3(三重校验) |
| **支持的存储网络** | FCP、IB、FCoE、iSCSI、NFS、CIFS/SMB、HTTP、FTP |
| **高可用性特性** | 主动-主动控制器,带有控制器容错和多路径 HA 存储功能 |
| **访问权限控制** | ADS,NIS,LDAP,Kerberos,Home Directory  |
| **网络协议支持** | NTP,Telnet,SSH,SMTP,802.3AD,LACP,NDMP v2/v3 |
| **系统监控** | 在线系统日志查看,系统故障诊断,Email自动故障报警,SNMP v2/v3协议 |
| **管理界面** | B/S图形化操作界面(InnoView)标准中文和英文,命令行管理（CLI) |
| **数据管理** | 灵活卷(OmniVol),重复数据删除(OmniDedup),数据压缩(OmniCompress),卷克隆(OmniClone) |
| **数据保护** | 数据快照（OmniShot),一键快照恢复（OmniRestore) |
| **许可软件** | 法规遵从WORM(OmniLock),远程数据镜像(OmniMirror),智能分层(OmniCache),存储虚拟化(OmniGateway),企级网盘(OmniDrive),阵列双活(OmniMetro) |
| **最大LUN数量** | 8192 |
| **支持的SAN主机数量** | 最多可支持 4096个主机 |
| **支持的操作系统** | Windows 2000、Windows Server 2003、Windows Server 2008、Windows Server 2012、Windows XP、Linux®、Oracle Solaris、AIX、HP-UX、Mac® OS、VMware® ESXi、Citrix®XenServer |
| **系统环境规格** |
|  | MK9080 | MK9060 | MK9040 | MK9020 |
| **额定功率** | 1500W | 750W | 750W | 750W |
| **高度** | 14英寸(35.8厘米),适合8U空间 | 7英寸(17.9厘米),适合4U空间 | 7英寸(17.9厘米),适合4U空间 | 7英寸(17.9厘米),适合4U空间 |
| **宽度** | 可安装在 19 英寸 IEC 兼容机柜上（17.6 英寸，44.7 厘米） |
| **工作温度、海拔高度和相对湿度** | 10°C 至 40°C（50°F 至 104°F）；海拔高度不超过 3,000 米（10,000 英尺）；相对湿度 20% 至 80%，非凝结 |
| **存放温度和相对湿度** | -40°C 至 70°C（-40°F 至 158°F）；在无空调的飞机货舱中，一般不高于 12,192 米（40,000 英尺），相对湿度 8% 至 80%，非凝结（在原包装中） |

# 名词解释

* **O****mniHA**

OMNIHA为关键业务环境提供可靠的高可用数据服务。它通过将不可用控制器的数据服务转移到集群中的另一个控制器上来确保数据可用性。通常，数据服务的转移对最终用户和应用程序是透明的，数据服务将快速恢复，不会出现明显的业务运转中断。

* **OmniCache**

OMNICache技术能够自动地复制、存储和服务的数据。OMNICache卷采用缓存模式保存最新访问的源控制器的数据。为保证缓存数据的最新性，OMNICache会清除那些近期没有被访问过的数据。同时，如果源数据卷的数据被修改了，OMNICache会在用户访问这些数据时获取修改后的数据，以保证数据的一致性。因此，OMNICache能够在源数据卷不经常发生变化的环境中优化运行。

* **OmniMirror**

OMNIMirror技术将数据镜像到一或多个网络控制器上。OMNIMirror不断地更新镜像数据，以确保数据是最新的，并且能够用于进行灾难恢复、减少磁带备份、发布只读数据、在非生产性上进行测试、执行联机数据迁移等等。如果您的企业分布在不同的地点，而所有地点都需要访问相同的数据集，OMNIMirror可以将同一数据发布到所有地点。通过自动更新这些数据，并支持对镜像数据的本地访问方式，OMNIMirror可以大大提高员工的工作效率。

* **OmniVol**

可以在不破坏底层物理存储结构的前提下被独立的设置、调整大小，管理和移动。

* **OmniClone**

OMNIClone是一个OmniVol 卷或另一个OMNIClone卷在某一时间点的克隆，该克隆是可写的。创建 OMNIClone卷仅仅需要几秒钟的时间，而且创建过程不会中断用户对被克隆的 OmniVol 卷的正常访问。

* **OmniDedup**

重复数据删除(data deduplication)是一种数据缩减技术，通常用于基于磁盘的备份系统，旨在减少存储系统中使用的存储容量。它的工作方式是在某个时间周期内查找不同文件中不同位置的重复可变大小数据块。

* **OmniCompress**

OMNICompress数据压缩技术可以在主存储、二级存储和归档存储上压缩灵活卷 (OmniVol) 中的数据，从而减少在存储系统上存储数据所需的物理容量。它可以压缩常规文件和 LUN。

* **OmniSplit**

可以基于快照来克隆出一个卷出来，可以实现完全与原卷分离（即完全复制快照内的数据块，据此创建一个可读写访问的空间）和保留与相同的数据块，克隆卷之后只保存与原快照有差异的部分。

* **OmniShot**

OMNIShot是COS文件系统“任意位置写入”功能带来的一项突出优势。 一份 SnapShot 是文件系统的在线只读拷贝。创建文件系统的一份 OMNISnap仅仅需要几秒种的时间，并且除非原始文件被删除或者更改，数据快照并不占用额外的磁盘空间。这种只有当数据块发生改动时才进行数据块复制的技术被称作“copy-on-write”，只有修改活动文件系统中的数据块并写入磁盘中新的位置时，OMNISnap才会占用额外的磁盘空间。

* **OmniRestore**

O**mni**Restore 可以用存储的快照副本快速恢复丢失的文件。

* **NDMP**

网络数据管理协议(或简称为 NDMP)是一种开放式标准，用于集中控制企业级数据管理。NDMP 体系结构可使备份应用程序厂商和其他网络连接服务器控制本机备份和恢复设备。