天翼云
State Cloud

存储资源盘活系统

产品白皮书

天翼云科技有限公司

目 录

1 产品概述

1.1 产品定义

HBlock 是中国电信天翼云自主研发的存储资源盘活系统（Storage Resource Reutilization System，简称 SRRS），是一款轻量级存储集群控制器，实现了全用户态的软件定义存储，将通用服务器及其管理的闲置存储资源转换成高可用的虚拟磁盘，通过标准 iSCSI 协议提供分布式块存储服务，挂载给本地服务器（或其他远程服务器）使用，实现对资源的集约利用。同时，产品拥有良好的异构设备兼容性及场景化适配能力，无惧 IT 架构升级带来的挑战，助力企业用户降本增效和绿色转型。

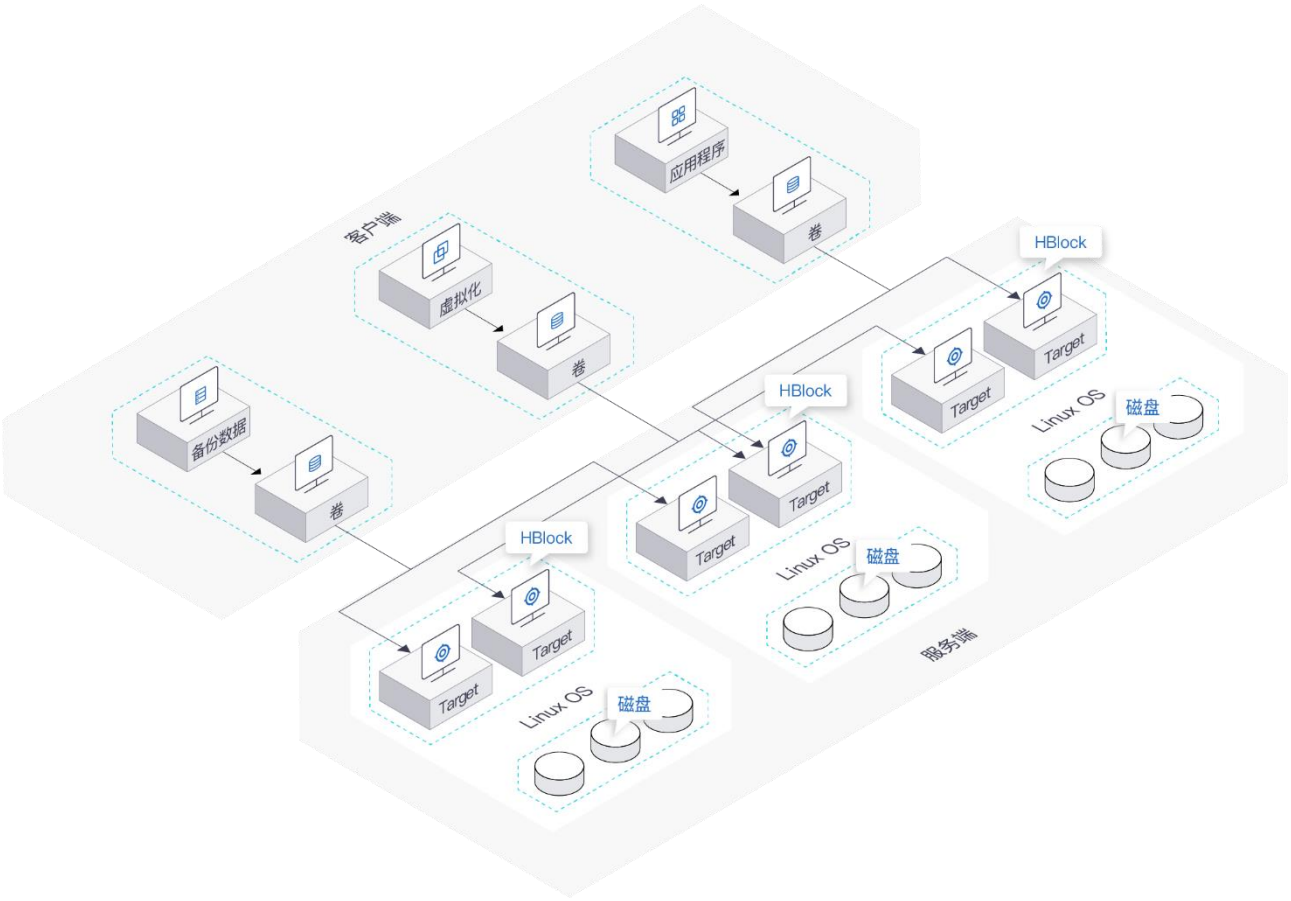

图1.HBlock 架构图

HBlock 可以像普通应用程序那样以非 root 方式安装在 Linux 操作系统中，与服务器中的其他应用混合部署，形成的高性能高可用的虚拟硬盘供业务使用。如此一来，HBlock 可以在不影响用户业务、无需额外采购设备的情况下，直接原地盘活存储资源！

传统的硬件存储阵列可以为每个逻辑卷提供低延迟和高可用性，但存在横向扩展性差、成本高的问题，并且可能形成许多“数据孤岛”，导致存储资源成本高和利用率低。传统的分布式存储虽然具有很强的吸引力，但通常存在部署复杂、性能差、稳定性差等问题。

HBlock 以完全不同的方式提供存储阵列：

- **易安装**：HBlock安装包是一个zip包，可以安装在通用64位x86服务器或者ARM服务器上的主流Linux操作系统上，支持物理服务器、裸金属服务器、虚拟机。HBlock与硬件驱动程序完全解耦，用户可以自由使用市场上最新的硬件，减少供应商锁定。
- **绿色**：HBlock作为一组用户态进程运行，不依赖特定版本的Linux内核或发行版，不依赖、不修改操作系统环境，不独占整个硬盘，不干扰任何其他进程的执行。因此，HBlock可以与其他应用同时运行在同一Linux操作系统实例中。我们称此功能为“绿色”。一方面，它可以帮助用户提高现有硬件资源的利用率，另一方面，它也降低了用户使用HBlock的门槛 — 甚至不需要虚拟机。
- **高利用率**：HBlock支持异构硬件，集群中的每个Linux操作系统实例可以具有不同的硬件配置，例如不同数量的CPU、不同的内存大小、不同容量的本地硬盘等。因此可以提高现有硬件资源的利用率。
- **高性能**：HBlock采用分布式多控架构，提供像传统硬件存储阵列一样的低延迟和高可用性，以及像传统分布式存储一样的高扩展性和高吞吐量。支持在不中断业务的情况下，从 3 台服务器扩展到数千台服务器，并从数千台服务器逐台缩小到3台服务器。
- **高质量**：当集群中同时发生的磁盘故障数不大于逻辑卷冗余模式允许的故障数（对于3副本模式，允许的故障数为2；对于纠删码N+M模式，允许的故障数为M），不影响HBlock的数据持久性。在集群中发生单个服务器、链路或磁盘故障时，HBlock保证服务可用。HBlock是面向混沌（Chaos）环境设计的，可适用于弱网、弱算、弱盘等不确定环境，并在发布之前已经在复杂和大规模的环境中进行了充分的测试。

1.2 HBlock vs 传统硬件存储阵列

	HBlock	传统硬件存储阵列
安装	简单 高度优化的压缩安装包，约 170MB，3 分钟即可完成集群安装。	复杂 涉及数据中心、机架空间、电源和网络等环境准备，至少需要几天的现场部署。
硬件要求	低 进程级软件定义存储控制器，支持通用服务器、裸金属服务器、虚拟机。	高 专用硬件控制器
硬件资源利用率	高 可以与其他应用同时部署在同一 Linux 操作系统实例中。 集群中的每个 Linux 操作系统实例可以具有不同的硬件配置，例如不同数量的 CPU、不同的内存大小、不同容量的本地硬盘等。 2 GB 内存便可满足最低配置要求。	低 需要新的硬件并独占整个硬件资源
水平可扩展性	高 3 到数千个软件定义控制器共享一个容量无限的虚拟存储池。	低 通常不超过 8 个硬件控制器共享容量有限的物理磁盘框。
聚合吐量	高 无单点吞吐瓶颈。	低 控制器和磁盘框之间的有限带宽是吞吐量的瓶颈。
延迟	低 使用 10G 以太网互连+HDD，4KiB 随机写入延迟可达亚毫秒级。	低 使用背板互连+SSD，4KiB 随机写入延迟小于 0.1 毫秒。
可用性	高	高

	无需业务中断连接，实现秒级故障切换。	无需业务中断连接，实现秒级故障切换。
数据冗余保护	高 软件 RAID。 支持多副本和纠删码冗余模式，副本存储在不同的服务器。	中 仅支持硬件 RAID。 副本存储在同一磁盘框中。
总成本	低	高
故障处理	可以远程诊断	供应商现场诊断
应用场景	小型分布式系统存储高可用、存量资源利旧、新建自主管控。	企业核心业务。

1.3 HBlock vs 传统分布式存储

	HBlock	传统分布式存储
安装	简单 高度优化的压缩安装包，约 170MB，3 分钟即可完成集群安装。	复杂 需要做大量的准备工作，如配置网络连接、NTP 服务器、服务器访问身份认证等。
硬件兼容性	高 兼容支持 Linux 操作系统的 x86、ARM 硬件。没有设备驱动程序兼容性问题。	中 取决于具体硬件兼容性列表（HCL），驱动程序兼容性问题经常发生，硬件和软件供应商常常相互推责。
硬件资源利用率	高 可以与其他应用同时部署在同一 Linux 操作系统实例中。 集群中的每个 Linux 操作系统实例可以具有不同的硬件配置，例如不同数量的 CPU、不同的内存大小、不同容量的本地硬盘等。 2 GB 内存便可满足最低配置。	低 至少需要专用的虚拟机。 集群中每个实例都需要相同或者相似的硬件配置。
可用性	高 不中断业务，实现秒级故障切换。	中 通常需要配置虚拟 IP，节点和虚拟 IP 存在对应关系，因此要求有权限变更节点的网络配置，存储系统与网络系统之间紧耦合，不利于系统扩展和更新。
延迟	低 使用 10G 以太网互连+HDD，4KiB 随机写入延迟可达亚毫秒级。	高 使用 10G 以太网互连+HDD，4KiB 随机写入延迟大于 1 毫秒。

数据冗余保护	高 软件 RAID。 支持多副本和纠删码冗余模式，副本存储在不同的服务器。	低 仅支持多副本，磁盘利用率低于纠删码。
水平可扩展性	高 3 到数千个软件定义控制器 共享一个容量无限的虚拟存储池。	高 支持数台服务器到数千台服务器。
聚合吞吐	高 无吞吐量瓶颈。	高 无吞吐量瓶颈。

2 产品优势

2.1 易于安装

安装包可直接官网下载，只需 3 个命令即可将 HBlock 安装在 Linux 操作系统上，从安装包解压到集群初始化不超过 3 分钟，即可享受本地磁盘的读写体验与无限可扩展存储空间。只有一台服务器的情况下，只需满足单核 CPU、2GB 内存、10G 剩余硬盘空间即可进行单机版本的安装，同时再加上一个至少 5GB 大小的数据盘即可实现数据存储。

2.1.1 安装包小

安装包为 zip 类型，经过了高度优化，只有大约 170MB，安装部署非常方便。

2.1.2 硬件驱动程序解耦

HBlock 与硬件驱动程序完全解耦，可以安装在物理服务器、裸金属服务器、虚拟机的 Linux 操作系统上。因此，用户可以自由使用市场上最新的硬件，减少供应商锁定。只要服务器之间网络互通，就可以搭建集群。

2.2 高利用率

2.2.1 混合部署

HBlock 是一个用户态进程级软件定义存储控制器。与其他系统级、软硬件集成或云服务级软件定义存储解决方案相比，HBlock 不依赖、不修改操作系统环境，不独占整个硬盘，也不干扰其他进程的执行。因此，它可以与其他应用同时运行在同一 Linux 操作系统实例中，帮助用户提高现有硬件资源的利用率，同时也降低了用户使用 HBlock 的门槛。此外，HBlock 支持对数据目录设置配额，限制对磁盘空间的占用，在与其他业务混合部署的场景中，可避免对磁盘空间的争抢。

2.2.2 异构硬件部署

集群可以由不同架构的服务器组成，每个 Linux 操作系统实例可以具有不同的硬件配置，例如不同数量的 CPU、不同的内存大小、不同容量的本地硬盘等。因此，可以提高硬件资源利用率。

2.2.3 自动精简配置

精简配置为应用程序提供了比实际物理存储设备上更多可用的虚拟存储空间。在数据写入逻辑卷之前，HBlock 即可以为上层应用提供存储设备，而不占用任何物理存储空间。HBlock 的卷默认自动支持精简配置，提高了存储空间的有效利用。

2.3 兼容性强

HBlock 与通用 64 位 x86 服务器、ARM 服务器上的主流 Linux 操作系统兼容。支持部署在物理服务器、裸金属服务器、虚拟机的 Linux 操作系统上，可以将这些服务器整合成高性能的虚拟磁盘。HBlock 支持同一集群中使用异构的硬件和不同的操作系统，这使 HBlock 比其他软件定义的存储控制器更通用。

2.3.1 硬件兼容性

硬件兼容性包括:

- **CPU 架构**：通用 x86 服务器、ARM 服务器。
- **存储介质**：NVMe SSD、SAS SSD、SATA SSD、SAS HDD、NL-SAS HDD、SATA HDD。

2.3.2 软件兼容性

软件兼容性包括：

- **操作系统**：HBlock 可以部署在 Linux 操作系统上，不依赖特定版本的 Linux 内核或发行版。客户端支持 Windows 和 Linux 操作系统。
- **虚拟化平台**：支持与 KVM 和 VMware 的虚拟化平台整合。

- **数据库**：支持多种数据库应用程序，如 Oracle、MySQL、SQL Server、PostgreSQL、MongoDB、DB2 等。
- **应用**：支持各种企业 IT 应用、行业应用和 web 应用。
- 云、容器平台：提供 OpenStack Cinder 和 Kubernetes CSI 驱动，即插即用。

2.4 高可用

2.4.1 秒级故障切换

在集群模式下，一个逻辑卷对应至少两个 Target：Active Target 和 Standby Target。当卷对应的 Active Target 所在服务器故障时，HBlock 将在几秒内自动切换到 Standby Target，而不会导致业务中断。此外，HBlock 还支持一个逻辑卷有多个冷备 Target，当多个节点连续发生故障时，仍可有效保障数据访问不中断，提高了存储服务的可用性。

传统的虚拟 IP 模式采用“节点和虚拟 IP 对应”的设计方式，要求有权限变更节点的网络配置，存储系统与网络系统之间紧耦合，不利于系统扩展和更新。HBlock 采用先进的多控架构，只需要确保客户端能连接 Active Target、Standby Target 以及冷备 Target 所在服务器的 IP，就可以通过标准的 MPIO 技术实现秒级故障切换，不需要增加额外的虚拟 IP、代理 IP 等，不需要改变网络结构，从而可以方便地与其他业务系统混合部署。

2.4.2 无单点故障

HBlock 采用多控架构，并且集群中的服务器都采用冗余模式部署。在集群中，当单个服务器、单链路或单个磁盘发生故障时，或者在弱网、弱算、弱盘等不确定环境下，HBlock 可确保高可用。任何单点故障，都不会影响服务的可用性。

2.4.3 智能调速器

HBlock 监控数据读/写过程中磁盘空间、内存和其他资源的使用情况。当资源不足时，速度调节器将自动调整数据写入速度，以确保磁盘始终可写、服务始终可用。而其他存储产品在资源不足时，还会持续写入数据，最终把磁盘写满，导致服务突然中断。

2.5 高可靠性

2.5.1 故障域

支持数据目录级别和服务器级别的故障域，源数据的副本或者分片会分布在不同的故障域内，以确保磁盘故障或服务器故障的情况下，业务不中断，数据不丢失。磁盘以数据目录的方式加入到 HBlock 集群的管理范围内，针对拥有少量服务器但是大量磁盘的用户环境，可以选择数据目录级别的故障域，将数据存储到不同的磁盘。用户可以根据实际的资源情况，选择适合的故障域级别。

2.5.2 支持纠删码

HBlock 支持纠删码（Erasure Code，EC），提高数据冗余性和可靠性。纠删码是一种数据冗余保护机制，广泛应用于分布式存储领域。数据写入 HBlock 后，EC 模式会将源数据分割成 N 个片段，然后从 N 个片段中生成 M 个校验数据，得到 N+M 个数据片段，并将这些数据存放到 N+M 个不同的故障域内。对于纠删码 EC N+M 冗余模式，最多允许 M 个数据所在的故障域设备损毁，这 M 片数据可以通过另外 N 片的数据进行恢复。相对于三副本模式，EC 模式提供了相同或者更高的可靠性，显著提高了磁盘的利用率，节省了总成本。

在使用 HDD 和小块读写数据的场景中，HBlock 也支持纠删码模式，并且可以实现低延迟。

2.5.3 数据零丢失

HBlock 支持多副本和纠删码数据冗余保护机制。数据被存储在不同的故障域内，当单个故障域、单条链路或者单个磁盘发生故障时，HBlock 使用存储在其他故障域内的数据片，在后台开始重建/重新平衡数据，以便重新分配数据。所有数据不会出现丢失或者暂时不可用的情况。

HBlock 运行过程中可能遇到各种亚健康状态。例如，服务器中其他应用引起的高负载，CPU/内存使用率高，网络异常（如数据包丢失或高延迟）以及其他异常情况。这些情况统称为亚健康状态。在亚健康状态下，HBlock 仍可以确保数据不丢失，服务不停止。

另外，HBlock 有其独特的内部时钟检查机制，可以确定每台服务器的运行时间。HBlock 的每台服务器不需要配置 NTP 服务，服务器时钟可以任意设置。HBlock 对不同服务器的时钟

偏差没有要求，不会因为时钟不同步而导致数据丢失或者服务不可用。但是对于传统分布式存储，服务器必须配置 NTP，否则会引起数据丢失。

2.5.4 数据一致性

为了确保数据一致性，HBlock 支持多种数据校验，包括客户端-服务器端校验、集群内部全流程数据校验、多台服务器间数据校验等。

- 多台服务器间的数据一致性：集群中多台服务器上的数据以版本号为比较标准，最新数据是带有最新版本号的数据。这确保了数据的严格一致性。当发现异常副本后，HBlock 将自动修复异常副本。
- 内存数据和持久数据的一致性：HBlock 定期扫描内存和磁盘数据。当磁盘数据不可访问或者校验失败时，HBlock 会自动启动数据恢复进程来重建数据。

2.6 高性能

2.6.1 低延迟

HBlock 具备极低的读写延迟，使用普通 10G 以太网互联+HDD 的集群，随机写入 4KiB 块的延迟可达亚毫秒级，充分发挥硬件最大速度。

2.6.2 聚合吞吐

HBlock 中的 iSCSI Target 可以被创建在集群中的任何服务器上。创建 LUN 时，为了使系统负载均衡，HBlock 会选择集群中负载比较低的服务器作为 Target 服务器。因此，HBlock 最大化网络带宽和磁盘吞吐能力，没有单点的吞吐瓶颈。但是，对于传统硬件存储阵列，控制器和磁盘框之间有限带宽会成为吞吐量的瓶颈。

2.6.3 数据重建避免性能瓶颈

HBlock 支持多副本和纠删码数据冗余保护，数据片段存储在不同的服务器上。当一个磁盘或者服务器出现故障、集群中添加新服务器或移出服务器时，HBlock 将在后台自动启动数据重建/重新平衡，来重新分配数据。由于数据片段分布在多个不同的服务器上，因此将在

多个服务器上进行数据重建/重新平衡，从而有效避免了因单个服务器上大量数据重建/重新平衡造成的性能瓶颈，对业务的影响降到最低。

2.7 弹性扩展

HBlock 架构不仅支持纵向扩展（通过增加单服务器的处理器、内存、网络和磁盘进行扩展），还支持横向扩展（通过添加服务器进行扩展）。这使得 HBlock 可以基于 IOPS、存储空间和带宽进行独立扩展。

HBlock 支持灵活的扩展方法：通过添加新磁盘扩展现有服务器容量，或者通过添加新服务器来扩展容量。扩容后，无需重新定位大量数据，系统便可自动实现负载均衡。

使用 HBlock，用户不需要进行大量的前期投入。可以在使用过程中，随时按需添加服务器或磁盘，这些硬件可以是价格低廉易用的通用硬件，添加过程中不会中断业务。

2.8 安全认证

HBlock 支持质询握手认证协议（Challenge-Handshake Authentication Protocol，CHAP）。CHAP 是一种对等身份认证协议，允许 iSCSI 客户端和 Target 端基于密码进行安全身份认证。CHAP 包括单向认证和双向认证。对于单向 CHAP，Target 在连接时对客户端 initiator 进行身份认证。对于双向 CHAP，客户端和 Target 端基于各自的密码进行认证。HBlock 支持单向 CHAP，后续版本会支持双向 CHAP。

2.9 易操作和维护

2.9.1 多种操作方式

支持 RESTful API、命令行、中英文 Web 控制台三种操作方式，满足不同用户的操作需求，方便用户根据业务实际情况进行选择。

2.9.2 支持故障报警

HBlock 监控系统中所有的资源，提供一页式概览及详情查看功能，方便用户实时了解资源使用情况和紧急情况。当系统中的组件或资源出现异常时，HBlock 会自动发送邮件通知用户。

以慢速磁盘检测为例：磁盘长时间工作后，可能会出现组件老化等问题，导致 I/O 响应时间变慢，最终导致服务不可用。HBlock 会定期执行磁盘检测、监控、分析和诊断磁盘的读写请求，以评估磁盘是否为慢盘，在发现慢盘后及时通知用户。

2.9.3 支持 NAT 访问

通常，iSCSI initiator 通过内网访问 HBlock 的 Target。如果内网路由器上配置了网络地址转换（Network Address Translation，NAT），iSCSI initiator 可以通过 NAT 的外网 IP 连接到 Target 所在服务器，从而通过 HBlock 将 iSCSI 作为云服务进行远程提供。而其他存储产品不支持用户设置 Target 的 portal IP，因此无法支持 NAT 访问。

2.9.4 业务场景适配

HBlock 支持 3 种写策略，用户可以根据自己的业务场景特点，设置卷级别的数据写入方式。3 种方式如下：

- 回写：数据写入到内存后，立刻返回给客户端写成功，之后再异步写入磁盘。适用于对性能要求较高，稳定性要求不高的场景。
- 透写：数据同时写入内存和磁盘，并在两处都写成功后，再返回客户端写成功。适用于稳定性要求较高，写性能要求不高，且最近写入的数据会较快被读取的场景。
- 绕写：数据写入磁盘后即释放相应内存，写入磁盘成功后，立刻返回客户端写成功。适用于稳定性要求较高，性能要求不高，且写多读少的场景。

3 应用场景

3.1 存量资源利旧

利用 HBlock 的广泛兼容性，纳管各类服务器中的空闲存储空间，整合成存储池，并通过 iSCSI 协议向其他主机提供高可用高性能的虚拟盘。面对业务快速增长带来的存储容量需求，及各类型服务器资源闲置带来的资源浪费问题，HBlock 提供的快速部署和扩容的解决方案，实现了无需额外成本投入，即可提升存储资源的利用率，并支持业务的滚动更新，满足未来业务在容量和性能上不断变化的需求。

3.2 小型分布系统存储高可用

利用 HBlock 纳管应用节点的物理盘，并把所形成的虚拟盘再挂回到应用节点本地，使得应用程序访问的是高可用的虚拟盘。一方面让应用程序更容易实现高可用，另一方面盘活应用节点的存储资源，无需额外购买存储硬件，降低用户 TCO。

3.3 新建自主管控

通过 HBlock 纳管新建的存储资源池，用户可以保持对存储服务器的实际管控权，意味着用户不仅能够使用 HBlock 进行存储管理，还可以在硬件上部署其他应用，以充分发挥硬件的价值。传统的软硬一体式存储产品，或者分布式存储方案，要求独占设备，用户只能通过管理界面进行有限的操作，失去了对设备的管理权。使用 HBlock 管理的存储集群，实现了用户对资源池的全面管控，提升了更多的操作自由度。后期对资源池进行升级扩容，可以任意选择任意规格和型号的服务器，无供应商锁定问题，客户结合自身的业务需求和预算灵活选择合适的硬件，从而保护了客户的投资。

4 部署方式

- **独立部署**

HBlock 可以部署在物理服务器、裸金属服务器、虚拟机中，既具有传统硬件存储阵列的低延迟、高可用，又具有传统分布式存储的高扩展性和高吞吐量。

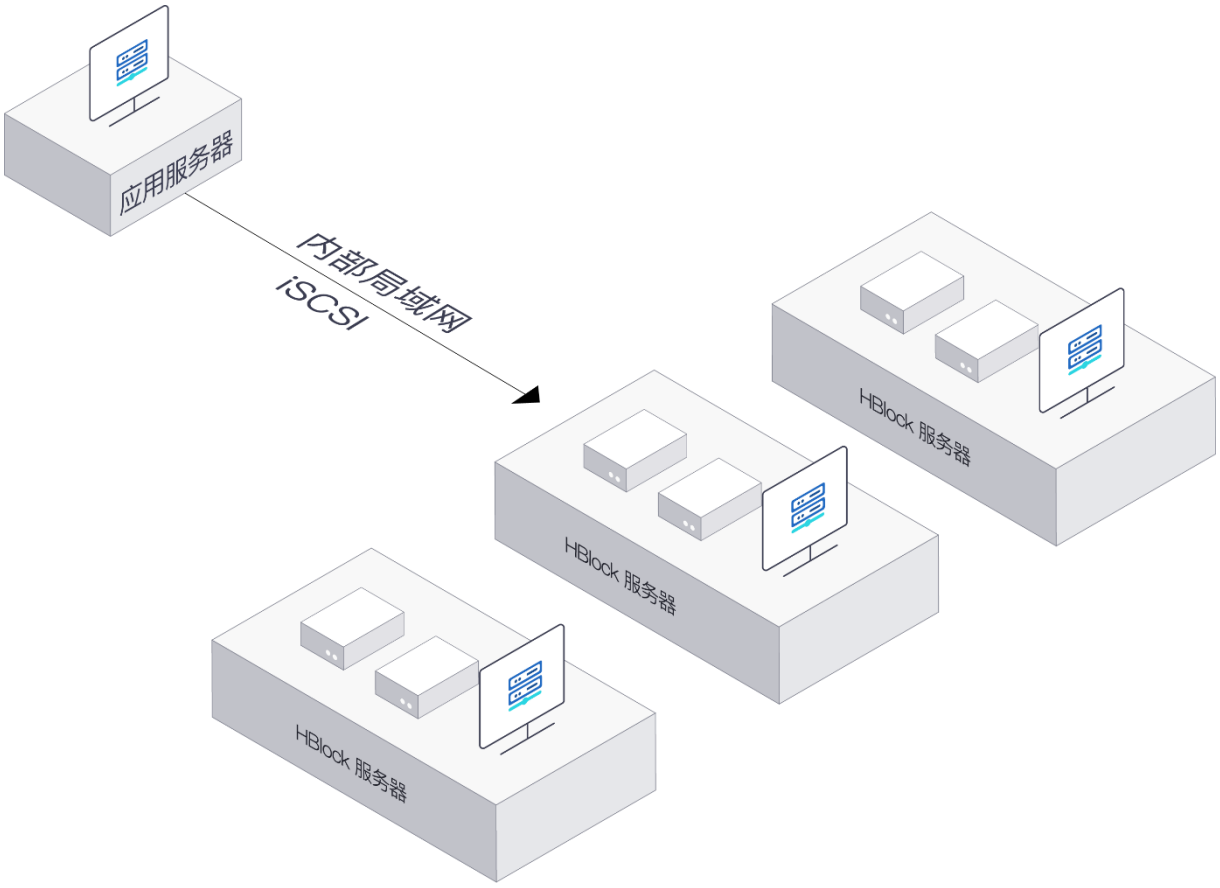

图2.HBlock 独立部署

5 规格

架构	分布式多控架构
支持协议	标准 iSCSI 协议
服务器数	单机版，集群版：3-数千节点
LUN 的个数	不受限
CPU 体系结构	x86、ARM
操作系统	CentOS 7、8、9，CTyunOS 3。64 位操作系统 ARM 架构的硬件环境下，推荐使用 PageSize 为 4k 的操作系统
混合部署	支持 可以与其他应用同时运行在同一 Linux 操作系统实例中
异构硬件部署	支持 允许集群中每个 Linux 操作系统实例有不同的硬件配置
高可用	支持，支持 MPIO
高性能	低时延，高吞吐量
数据冗余保护	多副本：3 副本 纠删码：N+M 冗余模式
存储介质	NVMe SSD、SAS SSD、SATA SSD、SAS HDD、NL-SAS HDD、SATA HDD
网络介质	TCP/IP