联想超融合 AIO H3000

产品白皮书

[版权声明]

目录

1 引言

联想超融合软件定义数据中心解决方案是与虚拟化管理程序(Hypervisor)无关的,具有高度弹性的 IT 基础架构平台，适合各类软件定义的虚拟化应用。联想以软件为中心的超融合解决方案正在革新企业虚拟化和私有云市场，联想超融合 AIO H3000 是企业虚拟化的经典架构和最佳实践，也是企业建设私有云的最佳组件。联想超融合在各个层面完全集成了服务器虚拟化技术，从用户界面到数据管理；同时支持所有可能的虚拟数据中心部署形式，包括虚拟化、私有云和混合云。联想超融合 AIO 通过将标准 X86 服务器超融合化为包含虚拟化，计算、存储和网络的软件定义的 IT 基础设施解决方案，同时利用服务器内闪存和磁盘优化了性能和容量。联想的超融合解决方案提供给客户企业级数据服务和完整的扩展能力并实现存储资源池，通过统一的管理方式,降低了企业 IT 技术门槛,使 IT 部门聚焦于业务推进和变革,而非 IT 技术本身。因此，极大地提升了企业 IT 的效率，并显著降低了成本。

本文将涵盖数据中心的演变，联想超融合术，以解决现代 IT 管理员所面临的挑战。

1.1 软件定义数据中心内面临的存储挑战

软件定义的数据中心（SDDC）是融合了虚拟化、计算、存储和网络的统一数据中心平台。SDDC 利用商用 X86 硬件和智能软件集中资源进行管理，旨在提升 IT 基础设施的使用效率。

尽管服务器虚拟化技术已经通过实现计算资源（CPU，内存）的复用而创造了一个高效的基础设施，然而存储技术在过去十年中只取得了渐进式创新。共享存储通常指 SAN 或 NAS，其技术发展是来源于使用服务器虚拟化诸多好处过程中的需求。然而，SAN 或者 NAS 在 IT 发展过程中也面临很多问题。例如存储构建（例如，LUNs、卷、文件系统）和虚拟化构建（虚拟机或 VM）之间的技术与流程不匹配；存储网络解决方案成本昂贵，由于其存储网络技术的复杂性并增加了存储管理的复杂度。所以，尽管某些新技术（例如闪存）的出现显著提升了存储性能，但这些解决方案未解决整体 IT 基础设施存储部分的成本和复杂度的问题。

近年来，基于 IP 网络的存储方案，诸如虚拟存储设备（VSA）和分布式文件系统应运而生。然而，这些产品和技术在可扩展性和性能方面仍受到限制，同时又未能显著提升存储的可管理性。因此，它们在大多数情况下无法非常完美的替代传统存储方案。虽然其他替代方案如 Hadoop 等也因大数据环境和特定的存储工作负载而产生。但是，这些替代方案也不适合于通用虚拟化环境的存储工作负载。

1.2 超融合的定义

超融合（Hyper Converged Infrastructure）是基于标准 X86 商用硬件，通过软件提供了存储供给和管理能力，以应对虚拟化环境中 IT 管理面临的挑战。超融合软件部署在虚拟化管理程序（Hypervisor）中，并实现各个服务器中物理存储资源集中，形成统一存储资源池进行管理。超融合技术提供存储池来实现以虚拟机为中心的 IT 基础设施管理，从而填补了存储基础设施和虚拟化平台之间的差距。

通过超融合的方式，可以实现一个完整的软件定义的基础设施。超融合将利用业界标准服务器、万兆网络、固态硬盘 SSD 或 PCI 连接的闪存和磁盘驱动器来交付整个 IT 基础设施资源。其通过与虚拟化管理程序集成的软件实体（融合计算、存储和网络），而不是孤立的 IT 硬件资源（存储阵列）来满足业务部门对于 IT 基础设施的需求。软件定义的方法也可以实现快速响应以支持业务快速发展的要求。将来，采用超融合技术的 IT 应用团队可以聚焦于业务，无需聚焦于 IT 技术本身，也无需预测和管理不同的物理资源池。

1.3 联想超融合

1.3.1 技术概述

基于超融合技术的联想超融 AIO H3000 产品，可助力 IT 完全实现软件定义数据中心的愿景。产品具有以下特性，可解决之前讨论的问题：

- 支持 VMware 虚拟化
- 完全集成到虚拟化用户界面
- 按需独立扩展计算资源和存储资源，支持纵向扩展和横向扩展

- 以虚拟机为中心提供存储资源
- 以虚拟机为中心实现企业级数据服务，例如克隆和快照等
- 支持实时迁移，动态负载均衡，高可用性，数据保护和灾难恢复
- 优化闪存性能和硬盘容量

总之，联想超融合 AIO H3000 显著简化了 IT，提高了效率，极大地降低了资本和经营支出。

1.3.2 解决方案

联想超融合提供两种完全不同类型的解决方案：

- AIO H3000 超融合一体机设备
- AIO H3000 软件+联想硬件升级

超融合 AIO 一体机设备提供了软硬件整合在一起的预定义、预验证的解决方案。这种方式消除了互操作性问题和性能问题，并简化了设备上线过程。

而 AIO H3000 软件定义的超融合解决方案则为客户提供更好的灵活性，可定制化的方案满足了客户的需求。通过升级现有的 X86 服务器上硬件配置，满足用户采用超融合 IT 基础架构的诉求。

2 系统架构

2.1 综述

联想超融合 AIO H3000 通过聚合多个服务器节点的本地存储资源，通过把所有存储资源合并成全局命名空间，最终构建一个超融合存储池供虚拟机使用。联想超融合AIO的管理软件LHS的管理实例是安装在虚拟化集群中的每台物理服务器上。该软件利用闪存存储资源如 NVMe 或 SATA 固态硬盘的特性提升存储性能，同时利用 SATA 或 SAS 硬盘驱动器提供存储容量。所有运行 LHS 软件的服务器都可以访问聚合的联想超融合存储池。然而，联想超融合的技术特点可以支持：1，构成超融合集群的服务器单独提供计算资源，即不要求所有服务器都具有存储资源或将存储资源聚集到联想超融合存储池；2，构成超融合集群的服务器支持异构，即不要求服务器提供相同的 CPU，内存资源，或者服务器贡献相同的容量将存储资源聚集到联想超融合存储池中。需要注意的是，所述 LHS 实例之间的通信建议部署 10GbE 的以太网专用网络。

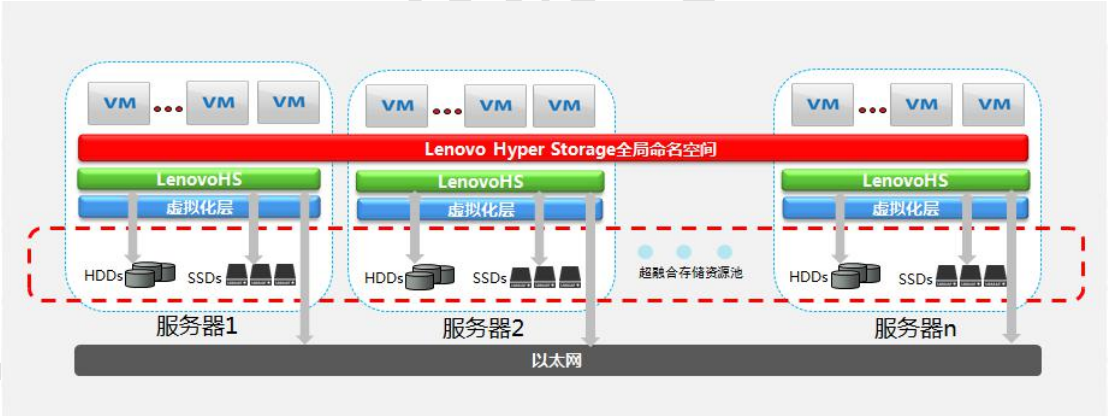

图 1

2.2 无关虚拟化管理程序的超融合基础架构

LHS 是在虚拟化环境中实现与 Hypervisor 无关的企业级超融合基础架构。下图描述了基于联想超融合技术的 IT 基础架构。

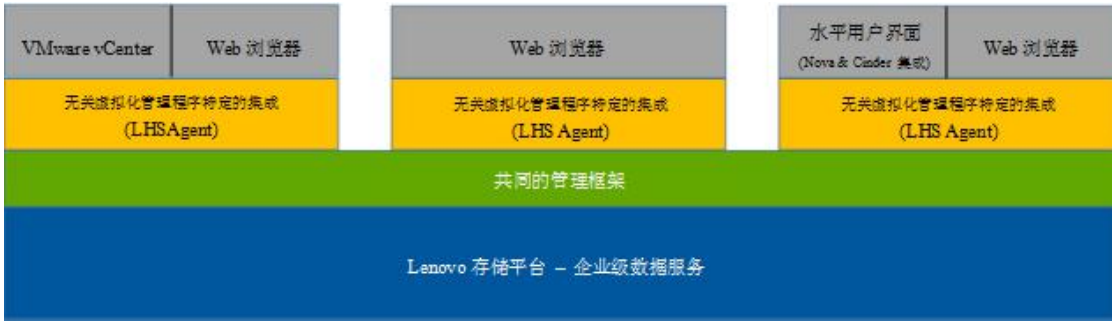

图 2：LHS 无关虚拟化管理程序的超融合架构

联想超融合的框架在所有的虚拟化管理平台上都是通用的。联想超融合 AIO 开发了一个 LHS Agent 并可集成到每一种虚拟化管理平台之中，从而实现在不同的虚拟化平台下适用同样的联想 web 管理界面。此外，也开发了一个可插入 VMware vSphere 客户端的插件，从而实现基于 vCentor 对联想用户界面进行管理。

2.3 VMware 环境中的联想超融合架构

联想超融合 AIO H3000 解决方案有几个方面是专门为 VMware 虚拟化环境而优化的。联想超融合提供 NFS 接口的存储资源池，可在一个超融合集群内的所有 ESXi 主机上共享。LHS 作为一个控制管理虚拟机运行需要消耗 4 个 vCPU 和 8GB RAM，实际部署中可以根据应用的性能要求来调整参数。非常有特色的是，联想超融合用户界面可以作为插件直接集成到 vSphere 客户端中，并且单一管理平台提供了全面的管理能力。联想软件与 VMware vSphere5.5 U2 或更高版本的所有版本兼容。

LHS 组件与服务

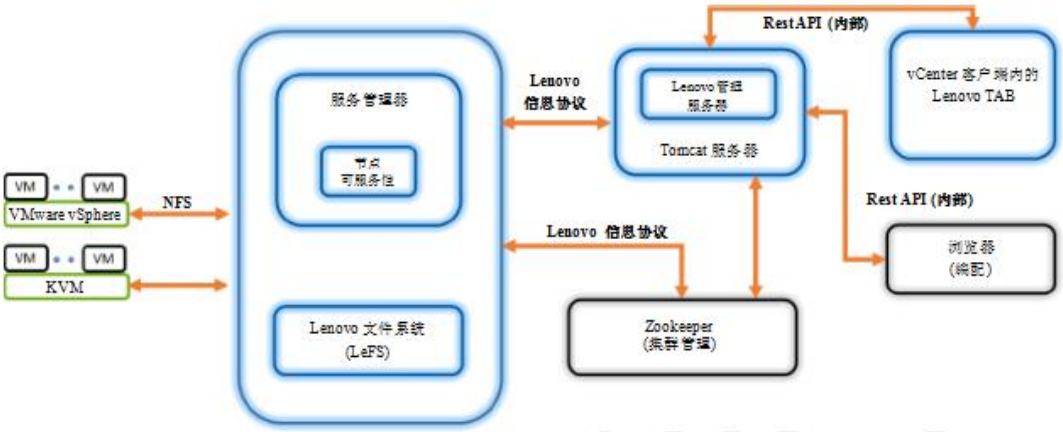

图 3：LHS 组件关系

联想超融合管理模块包括几个关键组件：Object Manager，分布式文件系统，数据库和 ZooKeeper。如上图所示。这些组件之间通过私有通信协议彼此通信。下面是对这些组件的介绍：

- NFS – 在联想超融合集群中所有服务器的物理资源被汇聚成独立的存储资源池通过 NFS 协议对所有集群内的虚拟机提供存储资源。这使得联想超融合存储与其他存储阵列能够共存，无需淘汰并替换现有的存储基础设施。
- Tomcat 运行在联想超融合集群内的某个节点上，以便支持管理服务。
- 内部的 Rest APIs 在 vCenter 客户端和浏览器上用于与联想制表通信。
- 服务管理器通过对现有环境添加磁盘和节点，为联想超融合集群提供向上扩展和横向扩展的能力。

本文还将进一步详细介绍其他的组件，运行在 Hypervisor 中的 LHS 的任何实体包括以下服务进程：

- 系统级
  - monit，监控系统状态
  - 针对 NFS v3 服务器的 nfsd
- Lenovo 分布式文件系统(LeFS)

- mfsd，针对 LeFS
- mfsmonit，监控 LeFS

- 对象管理(OM)
  - omNodeWatcher，监控节点的状态
  - omWalker，必要时启动数据同步
- 管理框架
  - Tomcat 提供管理服务
  - ZooKeeper 提供集群服务

2.4 集群与全局命名空间

LHS 在集群级的管理界面中呈现的是共享存储池。全局命名空间服务进程为超融合集群中的所有节点提供同一层次可视化的对象的状态。该服务还为所有的节点提供了修改这些对象的能力。全局命名空间还提供确保数据的全局一致性的能力，例如当两个节点同时要修改同一组数据，全局命名空间确保仅其中一个节点就能完成该操作。在任何时间点，为运行全局命名空间服务至少配置三个节点，同时大多数运行全局命名空间的节点必须在线以保证系统的写一致性。

在联想超融合管理软件 LHS 中，ZooKeeper 服务提供数据同步并维护联想超融合集群中所有节点的写一致性。ZooKeeper 服务和全局命名空间服务部署在联想超融合集群的所有节点上，然后根据集群大小选择某些节点激活这两个服务。ZooKeeper 服务也至少要求三个节点，大多数节点处于在线状态以保持写一致性。LHS 的高可用性服务确保当节点出现故障时数据仍然可用，并可以根据集群大小、副本策略以及通过机架感知等特性来提供更强大的容错能力。

2.5 Lenovo 分布式文件系统

LHS 使用联想超融合自身高度可扩展的、弹性的 Lenovo 分布式文件系统（LeFS）为企业级服务交付共享存储。研发联想分布式文件系统的目的是满足企业超融合虚拟环境中的需求。LeFS 通过同步在整个集群中不同进程之间的操作来消除系统“脑裂”的情况。LeFS 运行在虚拟化集群的每台服务器上，交付高可用的和可扩展的全局命名空间。

2.5.1 I/O 路径

联想超融合已优化其 I/O 路径，在利用由机械硬盘和闪存介质组成的混合磁盘配置的虚拟化环境中提供高性能存储。读写操作落到机械硬盘之前会提前重定向到 SSD，通过发挥 SSD 的物理性能而最大化整体系统的性能。读取/写入操作在虚拟环境中的模式在本质上主要是随机读写，因为被虚拟化服务器上寄宿的多个虚拟机承载的存储工作负载是混合在一起的。因此，虚拟化环境的业务负载对存储产生的工作负载总是随机的，这与单独考虑单个虚拟机的存储工作负载是不同的。LHS 通过使用基于日志的数据分布加速随机读取/写入操作，其通过分布式元数据将数据块映射到其储存位置。

联想超融合充分利用 SSD 的特性来作为读取和写入缓存。超融合系统中的闪存设备的特性可以保证数据在断电的情况下的数据不会丢失。写操作顺序得定向到 SSD 中的 Intent Log，同时将副本顺序写入其他节点上闪存中的 Intent log，副本方式可以保持数据可用性。然后通过把 intent log 中收集几千个小的随机写入整合成大 I/O 数据片。在 HDD 磁盘上按顺序分配空间，用整合后的少量的大 I/O 数据片按顺序写入到机械硬盘。按顺序写入的方式将最大化提升到 SSD 和 HDD 的 I/O 性能和固态硬盘耐力。

写操作 I/O 如下：

1) 联想超融合写操作 I/O 是通过距离虚拟客机最近的本地 LHS 实例完成的。
2) 被写入的数据在会尽量大的被分片并创建复本。数据及其副本通过网络写入到两个不同节点 SSD 上的 intent log 中。
3) 两个节点都告知确认数据已写入闪存介质。
4) 所有数据完成后写操作（包括已写入闪存的副本）后，写 IO 的确认信息返回给应用。
5) 数据最终倒盘到机械硬盘中。

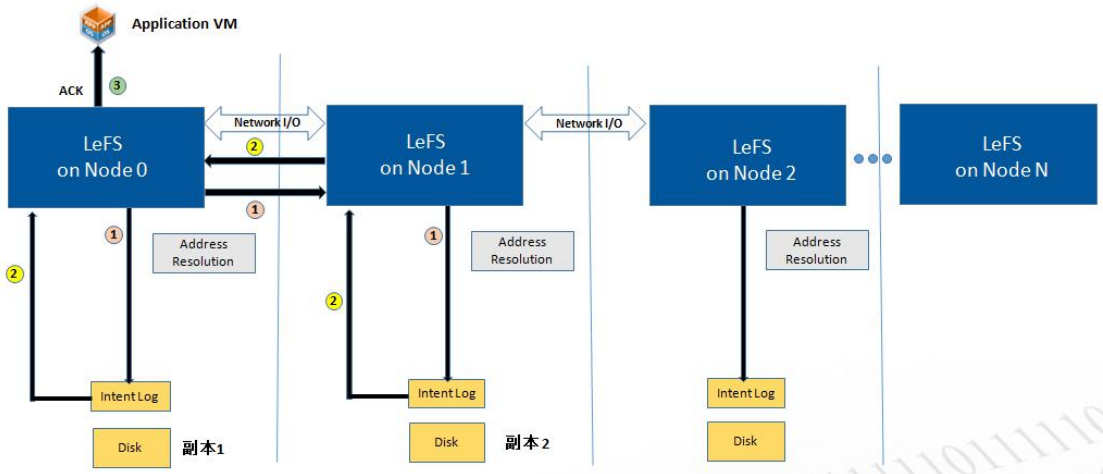

图 4：LHS 写入 I/O 路径

联想超融合也提供多样的数据分布策略，如机架感知功能和大城市集群功能。如果在数据写入时启用这些特性中的任意一个，LHS 可以确保其中一个及其副本遵守指定策略的数据放置于被限制过的位置。

通过在 SSD 缓存上保持元数据和热数据可加速读取操作。LeFS 也从其控制器虚拟机上利用 RAM 作为闪存之前的一个缓存层。然而，闪存基本上比 RAM 更大，因此能够支持更大的工作集。LeFS 进行了优化，可在混合存储配置（硬盘驱动器和闪存/固态硬盘的混合配置）中提供具有成本效益的高性能超融合集群。

进行读操作 I/O 路径如下（注意：如果数据的正本被标为过时（Stale），那么将从其副本中读取数据）：

1) 联想超融合读操作 I/O 是通过距离虚拟客机最近的本地 LHS 实例进行的。
2) 读取缓存的第一层是 LHS 内存。
3) 如果内存缓存没有命中读取的数据，那么从闪存介质中留存的读缓存中读取数据。
4) 如果闪存中未发现数据，LHS 将从硬盘驱动器中读取数据。
5) 读取 I/O 被确认返回到应用虚拟机。

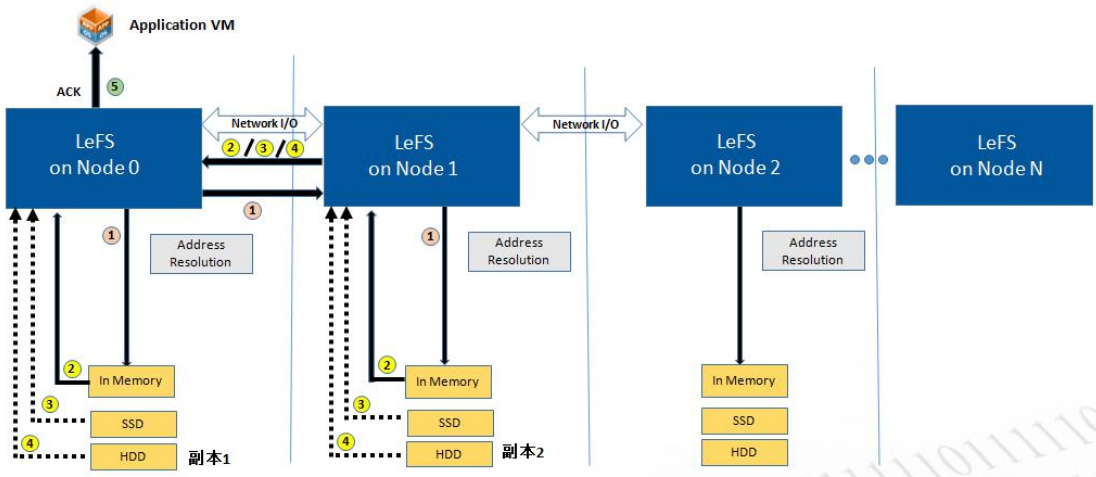

图 5：LHS 读取 I/O 路径

2.5.2 SSD 缓存与 Metadev

联想超融合在 SSD 上实现写缓冲 intent log、读缓存和元数据管理的三项功能，通过利用 SSD 的高速的物理性能实现混合存储架构下超融合解决方案最大的性价比。在 AIO H3000 群集中的每个服务器需要满足的闪存容量包括：4GB 的 intent log，200GB 的读取缓存，以及元数据管理的空间（5%的硬盘驱动器裸容量）。

联想超融合 AIO H3000 的 Metadev 是在超融合文件系统中的组件。Metadev 仅超融合文件系统的元数据，以加速元数据的读取/写入性能。联想超融合 AIO H3000 支持设置不同的存储块大小，但通常使用小尺寸的块（4K 或 8K）来平衡系统容量和性能。太小尺寸的存储块会导致大量的元数据，如果元数据不能完全缓存在内存中，就可能产生随机读取/写入 IO 性能问题。Metadev 是将元数据从数据分离的一种方式，从而允许将元数据存储在高性能设备（如固态硬盘）上。该方法显著加速了元数据的读取和写入操作。Metadev 存储元数据，即文件间接块和文件系统的空间分配记录，为实际数据本身提供了 Key 信息。

硬盘驱动器与 Metadev 的比率主要依赖于 LHS 部署过程中定义的存储页面块大小。使用 4K 页面大小的部署推荐该节点的硬盘驱动器裸容量的 5%的 Metadev 容量，用 8K 页面大小的部署要求该节点容量的 3%。如果硬件配置无法达到最低 Metadev 固态硬盘比率的要求，超融合系统则会自动使用现有的固态硬盘作为读取/写回缓存。

有两种类型的 Metadev：独有模式和共享模式。独有模式配置下 Metadev 占用的 SSD 不提供读取/写缓存。创建独有模式 Metadev，可以减少对 Metadev 固态磁盘的磨损。共享模式配置的 Metadev 可以在具有读/写高速缓存分区的固态硬盘上进行创建。共享模式 Metadev 配置可支持 5 年（一般如此）每天至少 10 次写入的固态硬盘。Metadev 可以是共享模式，也可以是独有模式——但用户不能同时配置二者。在超融合的安装过程中，用户可以选择独有模式或共享模式 Metadev 配置。如果有两个固态硬盘是可用的，并且用户可选择“启用 Metadev”，那么独有模式 Metadev 将配置一个固态硬盘分配给 Metadev，另一个固态硬盘指派给读取/写回缓存。

2.5.3 写操作 intent log

联想超融合的 intent log 存在于 SSD 上，它是超融合文件系统的组件，主要负责数据写入过程。写缓存 Intent log 的作用是吸收写入操作期间可能会发生的延迟峰值。达到这样的效果的方法是通过把小的随机写操作汇集到大数据片写操作，从而生成更少的元数据，最终提升了效率。数据写到闪存设备上的 Intent log 时，不会读取和修改元数据。写缓存 intent log 经过一段时间收集到一些随机写操作后，最终会在后台把数据倒盘到机械硬盘中。上文 I/O 路径小节更加全面地介绍了完整的读写 I/O 路径。

2.5.4 数据服务

联想超融合提供无限次的、高性能和高效的零拷贝快照和克隆。这些快照和克隆在虚拟机层进行配置和管理，而不是在存储层。这使得虚拟机管理员无需具备高深的存储专业知识能够利用先进的数据服务功能。

- 零拷贝快照

超融合利用基于日志的数据分布方法进行数据放置，并利用元数据将数据块映射到存储位置。所以一旦数据发生变化，数据块镜像就不会落在包含之前数据块镜像相同的存储位置上。而是把数据块的镜像被写入一个新的存储位置。同时，

元数据会被更新，以反映数据块的存储位置中的变化。利用这种方法，创建快照和把文件的元数据标记为只读一样简单。

快照只是某个时间点的数据虚拟副本，其特点是只读的，不占用任何容量，不要求任何前期空间预留。几秒即可创建快照，它独立于资源容量，并且快照创建不影响虚拟机的性能。即使当一个虚拟机的副本数据由于一个节点脱机不可用时，也可以创建快照。一旦该节点返回到集群中的可用状态，错过的写入数据会重放回该节点。重建快照，该节点再次完全同步。

联想超融合 AIO H3000 的快照技术使用写重定向的方法，无数据移动。这样的设计使快照可以被快速、无序删除，因为联想超融合文件系统只删除不再需要引用快照的元数据。此外，联想超融合上的所有快照具有故障一致性，提供第一层数据保护，而不会影响性能或容量。

- *应用一致性快照*

通过适当的应用一致性工具（例如，VSS quiescing），联想超融合支持在环境中提供应用一致性快照，这确保了应用程序从某时间点恢复的虚拟机从一致的状态开始。

可从 H3000 用户界面获取快照，包括选择快照是否应该具有应用一致性的能力。

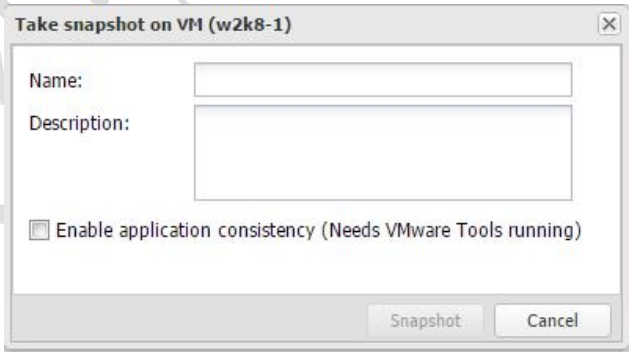

图 6：通过 AIO H3000 用户界面实现应用程序一致性

- 零拷贝克隆

联想超融合 AIO H3000 的零拷贝克隆技术采用的原理与快照技术相同，区别在于克隆是对虚拟数据拷贝的读取/写入。在虚拟机读写或者对虚拟机克隆导致

数据块被更新后，虚拟机或克隆相应的元数据随之更新以反映新镜像的存储位置，但是该虚拟机或克隆的原来所有快照和克隆的元数据保持不变。除非数据被修改，克隆不占用任何容量，也不要求任何前期空间保留，可在一秒钟内完成克隆，同时不影响虚拟机的性能。管理员可以通过克隆虚拟机的快照来构建完整的快照/克隆体系结构，并利用该体系复原并启动虚拟机。

克隆是一种在不同的环境中轻松地复制虚拟机的一种便捷方式；例如，开发和测试环境。可以在查看快照库存时通过 Lenovo 用户界面创建克隆。

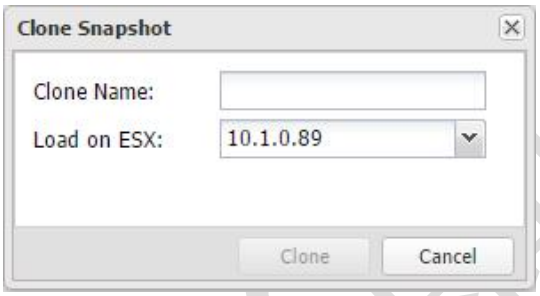

图 7：创建虚拟机的 Lenovo 克隆

- 自定义克隆策略

除了创建单个克隆，联想超融合通过选择添加客户机自定义脚本，支持自动创建多个克隆。管理员利用唯一的域名和其他字符来创建多个克隆，并使用脚本来订制个性化的客户机操作系统。LHS 完全集成 VMware 的客户机自定义规范，从而能通过自定义的向导把这些规范使用到联想超融合的克隆能力。

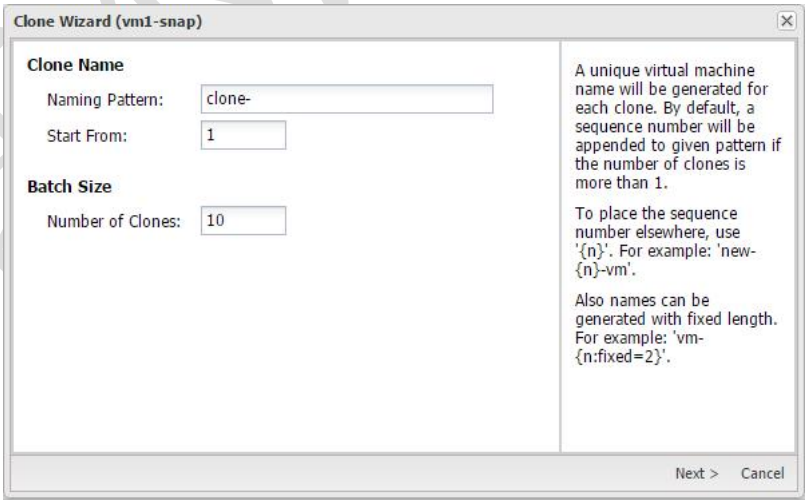

图 8：创建多个克隆

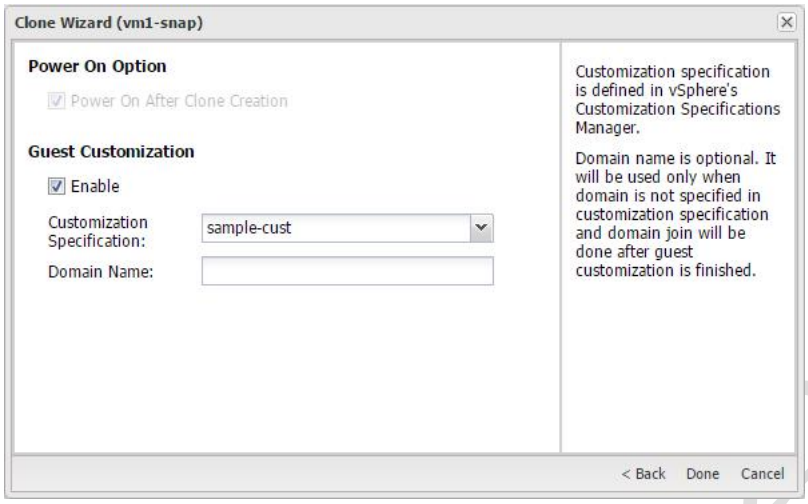

图 9：使用客户自定义规范

2.6 数据布局

联想超融合为虚拟化的工作负载优化了数据布局，从而简化了管理员的工作，消除了客户对成本和性能之间的权衡。虚拟化环境中的读取/写入操作在本质上都是随机的，把宿主服务器上的多个虚拟机的存储工作负载混合在一起生成一个单一的混合存储工作负载。因此，从存储的角度看，存储工作负载总是随机出现，它独立于各个虚拟机所产生的单独存储工作负载——有时被称为“I/O 混合器”现象。LHS 使用基于日志的布局方式对随机写入操作进行加速。该布局与一个日志结构的文件系统的布局是非常相似的。高可用的元数据管理用来便将数据块映射到其存储位置。

联想超融合也智能地优化虚拟机的映射以存储资源，提供数据最大的概率在本地的缓存读取。在大多数情况下，为虚拟机提供的存储来自部署该虚拟机的本地服务器的资源。在正常操作下，虚拟机的读取和写入操作都是宿主服务器的操作。

- 容量再平衡

超融合集群的虚拟机常见情形是个别节点在容量和性能两方面被过度利用。这种情况取决于虚拟机在数据分布模型中是如何供给，单个节点可以包含比集群中的其他节点更多的组件。为了解决这种情况，联想超融合提供一个工具，运行该工具可以自动重新平衡集群中所有节点的数据。该工具被称为 MxBalance，它

确保集群的工作负载均匀分布。关于如何操作请参考文档《联想 AIO H3000 产品高级版操作指导书-对集群进行数据均衡》

- 超融合存储池

超融合文件系统创建了包含底层物理存储设备的虚拟存储池。存储池包括多个物理磁盘的总资源。每个存储池具有相关策略来确定这些设备的数据布局和冗余。可以通过添加任意数量的物理磁盘实现存储池增加——可立即实现存储池容量增加。

超融合存储为物理节点上所有硬盘驱动器维护一个存储池，同时为该物理节点上的闪存介质维护一个单独的存储池。硬盘驱动器的存储资源池会执行数据布局策略把数据分片后分布到所有硬盘中。当新的磁盘被添加到存储池时，超融合文件系统将立即利用新增的容量并将新数据写入可用资源。闪存介质的存储池被划分为三个部分：intent log、读取缓存和 Metadev。启用本地副本后，存储池也会使用一个本地镜像以确保节点本身内的冗余。为了给跨集群提供冗余，联想超融合将存储池的数据镜像到不同节点上。

注：为简单起见，下面的图表显示了按比例缩小的环境，在该环境中每个磁盘仅包含来自一个文件中的一个成员。实际部署中在每个磁盘上会有多个文件的多个成员存在。

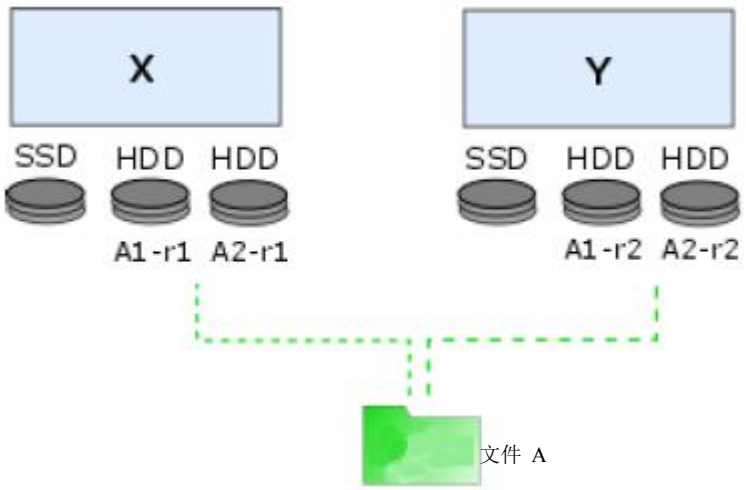

图 10：磁盘上的垂直条带数据

在上面的图中，文件 A 被映射到节点 X 和 Y 内的硬盘驱动器上，其文件被分片后其成员（A1-R1，A2-R1）及其相应的副本（A1-R2，A2-R2）分布在每个节点上。

当启用超融合 Metro Cluster 集群（或机架感知）策略时，还有更多关于数据布局的指南。在安装过程中，数据的一个副本被写入机架（或网站/故障域）内，数据的其它副本（或备份）会被写入到不同的机架。

- 本地副本

联想超融合的集群范围内的副本技术可以确保在硬盘发生故障时不会丢失数据。为了确保硬盘发生故障时零宕机，超融合 AIO H3000 提供了两种配置：硬件 RAID 和本地副本。一般建议客户为硬盘驱动器配置 RAID5 或更好的，为固态硬盘配置 RAID1，以确保在磁盘发生故障时零宕机。作为硬件 RAID 的替代方案，超融合提供本地拷贝的选项，该选项可以在安装过程中进行配置。

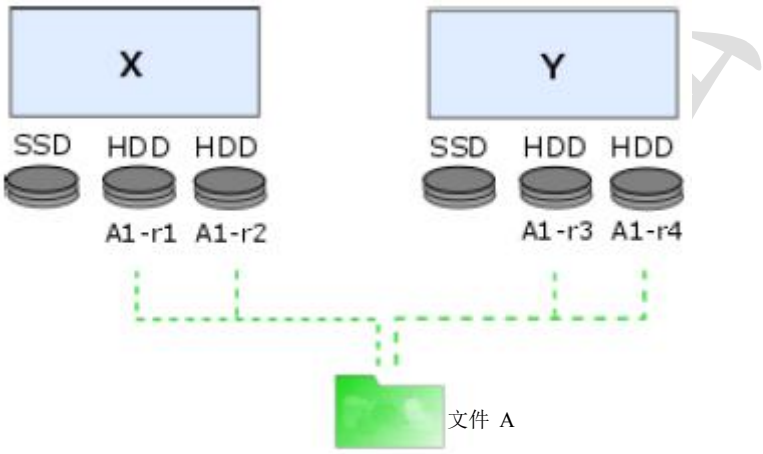

图 11：本地副本的数据布局

当配置联想超融合系统没有设置本地副本的功能时，两个副本数据存在于集群内两个节点上。如果设置了本地副本，除了集群范围内的镜像，还将会创建数据的本节点内镜像。这使得集群内两个节点共存在四个副本。如上图所示，该集群在节点 X 内的单独磁盘上现有 A1-R1 及其副本 A1-r2，还包括节点 Y 上的副本 A1-R3 和 A1-R4。LHS 提供固态硬盘和硬盘驱动器的本地镜像，从而确保任何磁盘故障都是零宕机。当 Lenovo 控制器虚拟机从节点内读取数据时，如果检测到磁盘出现故障，将读起操作重定向到同一个节点中的健康硬盘上。

2.7 优化性能的配置

提升 LHS 性能的方式有很多种。本节将介绍在安装和使用联想超融合 AIO H3000 过程中的优化信息，

- 针对应用类型对数据块大小设置

块的大小或内存页的大小决定超融合数据存储上虚拟机的最小空间分配单位。联想超融合集权在数据存储上默认页面大小为 4K。安装过程中提供一个选项，以配置超融合集群数据存储的页面大小和部署到数据存储的虚拟机磁盘（VMDK）的默认页面大小。例如，如果页面大小选择为 8K，为所有客户虚拟机的最小空间分配单元将是 8K。

- Metadev 与固态硬盘

Metadev 是联想超融合文件系统中的设备的一类。Metadev 将元数据从数据中分离出来，并储在高性能设备中，如固态硬盘。启用 Metadev 可显著加速元数据读取和写入操作。SSD 缓存与 Metadev 小节中将更详细讨论 Metadev。

- 提高超融合管理虚拟机资源

根据联想超融合存储上运行的工作负载，增加分配给管理虚拟机存储器和 CPU 资源的量会带来性能的改善。如果结果表明超融合控制器虚拟机的默认资源预留太小，而导致无法支持某些繁重的工作负载。通过管理程序的管理界面增加 CPU 和/或内存预留可轻松解决资源不足的这些问题。LHS 无缝地执行数据重建操作，以确保配置更改过程中保持集群正常的运行时间。

- 硬件组件

由联想超融合 AIO H3000 搭建的底层硬件会对集群的整体性能产生一定的影响。考虑联想超融合解决方案的关键组件是存储、网络和计算。提升的计算资源可以利用较少的 CPU 数量实现更多的虚拟机密度。从网络的角度来看，联想超融合系统建议在 10GbE 网络上部署超融合存储网络，要么使用一个专用的虚拟交换机，或者具有专用 VLAN 的共享虚拟交换机，或者使用专用的端口组分布式虚拟

交换机。最后，对于存储设备，无论是硬盘数量和驱动器类型都影响整体集群性能。当每个服务器有四个或更多个硬盘驱动器时在联想超融合集群中可实现更高的性能。

- 虚拟机配置

由于 LHS 的条带化策略，与虚拟机相关的虚拟磁盘的配置会带来性能提升。每个虚拟机配置多个虚拟磁盘将提高联想超融合集群中每个物理磁盘的利用率，从而提高总集群性能。此外，联想超融合在 VMware 环境中进行了优化，以实现与半虚拟 SCSI 控制器连接的数据虚拟磁盘。

3 联想超融合 AIO H3000 产品优势

3.1 业务应用场景

联想超融合解决方案可以直接为企业提供软件定义的数据中心解决方案，也可以为托管服务供应商（MSP）提供超融合的解决方案。超融合系统可以支持被集成到现有的用户访问控制和多租户的框架内，从而满足 MSP 根据需要在其整个环境中按需提供资源的要求。

3.1.1 关键业务应用环境

联想超融合存储平台的灵活性使其能够支持数据中心内的各种用例。超融合可以作为关键业务应用的存储解决方案。联想超融合提供的完整的企业级数据服务符合业界标准，可以支持采用通用存储器上运行的各种工作负载。

3.1.2 远程办公与分支机构的环境

LHS 也非常适合于远程办公室和分支机构（ROBO）。ROBO 环境需要与专用存储相同的企业级数据服务，但在更有限的预算范围内运作，所以面临着更多的基础设施的制约。联想超融合专注于以虚拟机为中心的管理，克服了 ROBO 常见的缺乏专门技能的问题。此外，将计算和存储一起集成到超融合节点内，大大减少了空间和功耗要求。

3.1.3 测试与开发环境

联想超融合 AIO H3000 的零拷贝快照和克隆技术对于测试和开发团队的场景有很好的优势，开发只需要从最新生产数据中某一个时间点的快照拷贝数据即可，而不是拷贝整个 LUN，（这是传统存储阵列模式）。在虚拟机级别获取超融合存储资源池内数据的快照和克隆。该快照和克隆在时间、性能和容量上都是高效的，这意味着可以瞬时创建和删除它们，而对性能没有影响，而且不占用创建上的任何空间。

3.1.4 虚拟桌面基础设施

联想超融合 AIO H3000 的快照和克隆对虚拟桌面应用（VDI）部署也是非常有益的，其中管理员可以在几分钟内部署数千个虚拟桌面。这些环境必须以相较于传统的独立 PC 有更低的成本同时保持更好的终端用户体验。LHS 能够通过利用固态硬盘作为读写缓存来满足 VDI 的性能要求，同时提供具有成本效益的解决方案。

3.1.5 灾难恢复

联想超融合存储可以作为灾难恢复（DR）环境中的端点。LHS 利用现有的复制软件（第三方商业软件等）提供虚拟机级别的可恢复性。灾备的站点也提供 LHS 所有的企业级数据服务功能，包括通过快照保持短期备份的能力。

3.2 系统灵活性

联想超融合为数据中心各个层面提供了灵活性和选择，从而实现超融合愿景。LHS 无关虚拟机管理程序的架构允许管理员选择他们所选的虚拟化管理软件。当涉及到服务器和磁盘时，该软件定义的方法提供了灵活性，使客户能够在他们所选择的硬件上部署联想的超融合解决方案。利用安装在商用 X86 服务器和闪存的存储介质，实现性能最大化，同时最小化成本。此外，联想超融合支持通过一次性的前期永久许可证或基于订重复阅的服务提供购买软件的灵活性。

3.3 简化 IT 管理

通过消除配置专用存储或管理卷、LUN、文件系统、RAID 及 ACL 等的需求，LHS 从根本上简化了存储管理。不再需要设计虚拟机和存储结构之间的映射，如 LUN 和卷之间，最小化管理过程中产生错误的可能性。在标准服务器上融合计算和存储资源，从而消除了存储网络的工作负荷，如分区等。通过将所有存储功能集成到虚拟化用户界面内，LHS 为虚拟机和存储管理提供了单一的管理窗口。一旦安装 LHS 后，管理员在创建虚拟机过程中只需从虚拟化用户界面就可以指向联想超融合存储池，LHS 将采取一切措施，为新虚拟机优化配置和分配资源。

LHS 的安装和配置过程只需要几分钟的时间，使系统或者虚拟机管理员能够安装和管理存储。在安装过程中，LHS 将所有服务器中的存储资源聚合到超融合集群中，并借助为虚拟机配置存储的全局命名空间将其作为超融合存储资源池。LHS 不会将任何以前分配的存储设备添加到联想超融合存储池，也支持有选择地从超融合存储池中卸载任何存储设备。安装程序会自动安装和配置所需的所有必备软件，并确定安装 LHS 所需的最低硬件和软件要求。安装程序将引导用户通过解决安装过程中出现的问题，以防未满足最低要求。

联想超融合存储池与所有数据服务，如复制、快照和零拷贝克隆按照虚拟机粒度从 VMware vSphere 客户端进行配置和管理。这种简化消除了存储管理的日常任务，使管理员能够集中精力管理应用程序和虚拟机。

3.4 线性扩展

联想超融合的系统架构提供了独立按需扩展容量和性能的能力，而无需过度提供资源。在不会影响性能的前提下，通过向超融合集群中添加新服务器进行横向线性扩展。下面的例子强调了该平台的线性可扩展性在 VDI 用例中的情况。即使增加部署虚拟机的数量以及添加节点到集群之后，访问虚拟桌面的平均延迟时间几乎保持不变。联想超融合支持 VDI 应用时，从 100 到 456 的桌面且不影响性能的线性扩展如下图所示：

桌面数量	100	300	456
节点数量（ESXi 主机）	2+1*	3	4
卡住或无反应的桌面（会话）数量	0	0	0
控制器虚拟机配置（CVM）	4vCPU/8GB 内存	4vCPU/8GB 内存	4vCPU/8GB 内存
性能指标			
最低响应时间（VSIbase）	731ms	797ms	843ms
平均响应时间（VSImax 平均）	823ms	929ms	995ms
最大响应时间（VSImax 阈值）	1731ms	1798ms	1844ms
*2 个超融合计算/存储节点+1 个单纯计算节点			

图 12：LHS 在 VDI 部署中的扩展线性

3.5 最大限度节省

通过利用服务器端的固态硬盘、内部驱动器或连接到标准服务器的 JBOD，LHS 在软件中可以提供存储阵列的能力。利用商品组件将计算和存储资源整合到标准服务器上，LHS 实现显著的 CAPEX 节省，而不会影响性能或可扩展性。

超融合消除了所有的存储配置活动，极大地简化了日常的数据管理活动，如从快照中恢复虚拟机以及使用克隆迅速创建新虚拟机。此外，LHS 消除了存储管理任务，如更改 RAID 参数或存储层管理。将计算和存储资源整合到标准商用服务器上，并消除了对存储阵列需求，因此，LHS 在电力、冷却和占地空间方面实现了大幅减少。

- 存储效率

联想超融合交付容量优化功能，如自动精简配置和在线压缩，实现了存储效率最大化。自动精简配置相较于物理存储资源能够提供数倍的容量。这实现了采购按需增量的需求。LHS 根据每个页面分配空间——如果设置虚拟机的页面大小为 4K，那么 LHS 将只分配 4K 块的数据。这样使得任何可能的过度配置的可能性最小化，并最大限度地提高了容量节省。自动精简配置是默认启用的，并且不需要容量的前期预约。

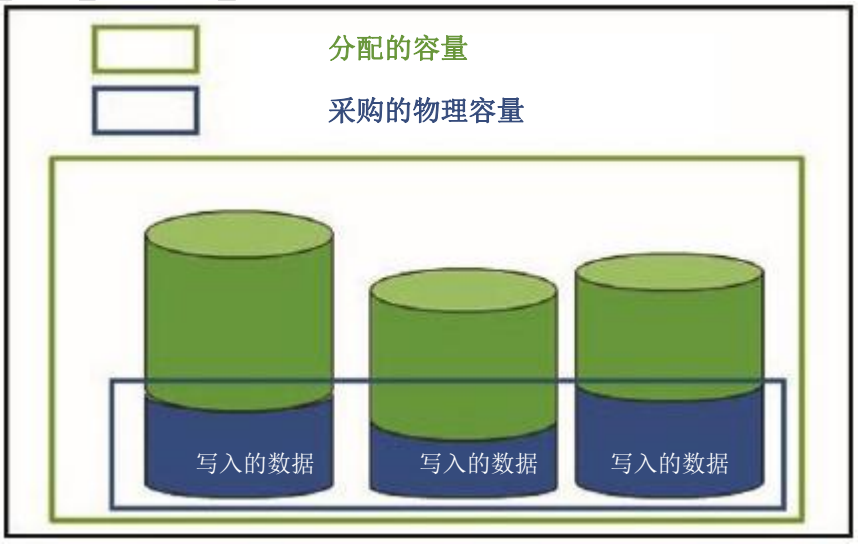

图 13：LHS 最大限度的容量节省

压缩也是默认启用。LHS 对数据和元数据进行在线压缩，而无性能损失。将闪存倒盘至机械硬盘过程中会对数据和元数据进行压缩，但在 intent log 和固态硬盘中的读缓存不进行压缩。LHS 利用 LZJB 压缩算法提供更高效的存储利用率。压缩和自动精简配置的组合为超融合存储部署提供 2-4 倍的容量节省。

- 软件敏捷性

传统存储的模式是将硬件与软件整合在一起。每四到五年更新存储硬件时，就必须重新购买软硬件。在非转让软件模型方面，超融合一体机设备没有什么不同。事实上，这种方式很难让用户适应计算技术创新的速度比存储技术创新更快的现状。IT 的最终用户不得不采用：要么放缓更新计算的速度（导致硬件过时与合作伙伴的效率减缓），要么更快地更新存储（导致更高的存储采购成本）。

图 14：软件灵敏性

联想超融合利用完全转让的软件许可使软件从硬件中解耦。这使终端用户在需要时能够购买他们所需要的硬件，而不必回购存储软件——联想超融合是数据中心内的最后一个存储软件采购。

客服电话：400-819-2223
邮箱:HCI@lenovocloud.com
官网：http://www.lenovocloud.com

联想超融合构建了根植于中国、拥有全球视野的创新三角，在虚拟化和分布式存储领域有超过十年的技术储备，积累了全球 70 多项专利，其中包括冷热数据检测、动态资源分配、任务调度等超融合领域最重要的核心专利。提升企业效率，快速满足不断变化的业务需求。