

联想LeoStor分布式存储系统

Lenovo™



李秀峰

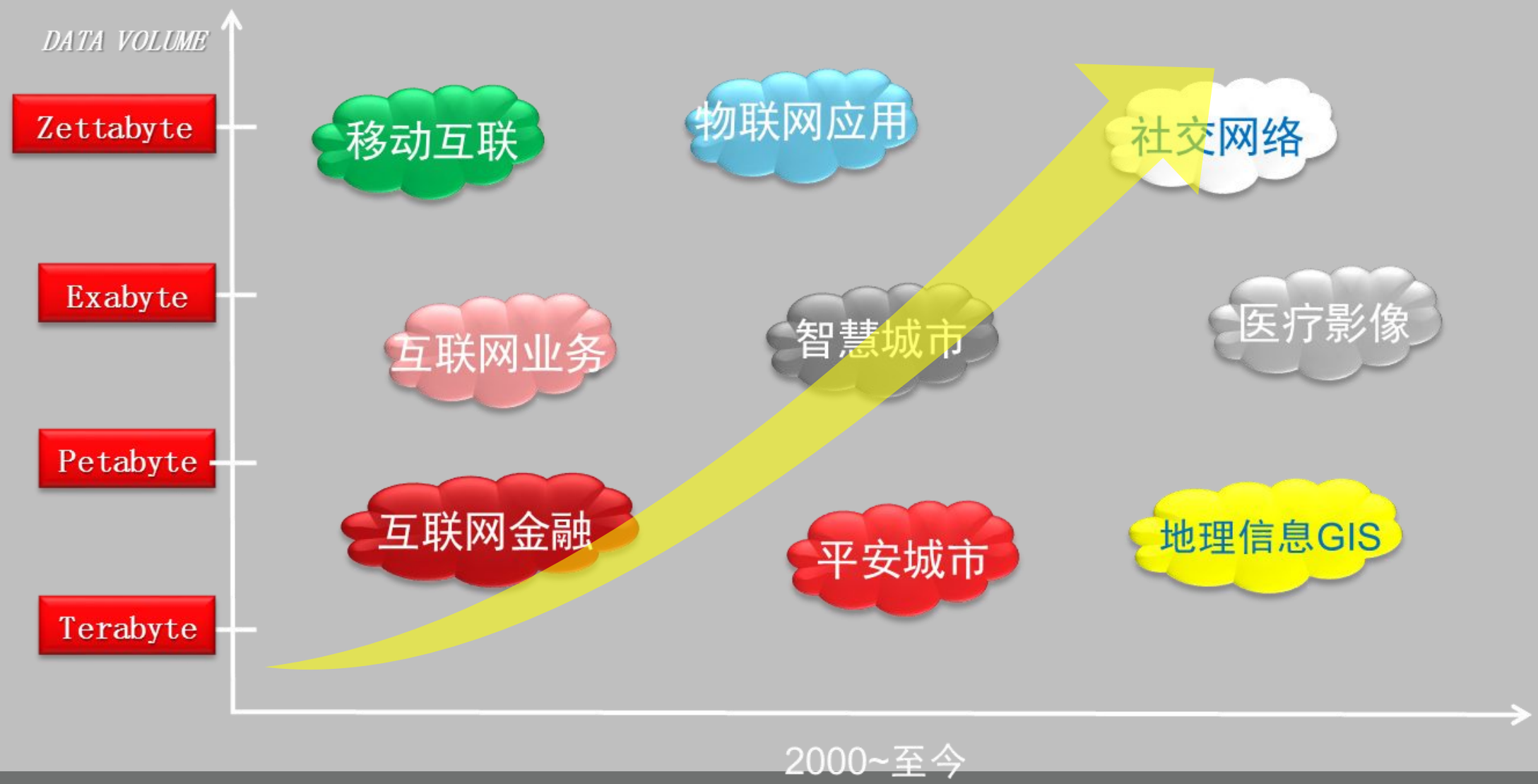
解决方案产品经理

联想大客户企业级产品营销

Mail : lixf12@Lenovo.com



新业务/应用模式带动数据爆炸式增长





大数据时代数据访问的挑战

时长

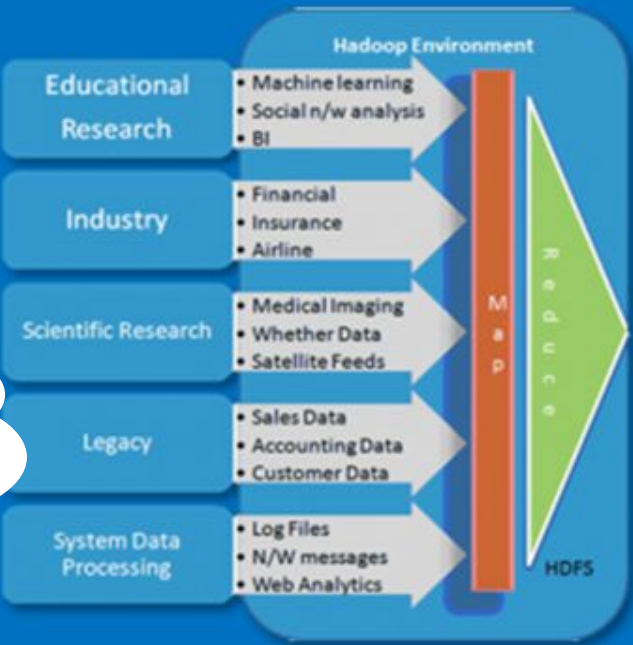
数据分析无法及时准确地进行分析，造成数据价值损失

数量

产生的数据飞速增加，但是可用数据却无法同比增加

种类

数据种类更为复杂，使存储访问、数据分析更为困难





传统存储面对大数据的挑战

开放性不足

- 盘阵存储标准化程度低，组件的更换常常导致系统故障或者硬件的升级，这些组件通常是由制造商专门开发用于此系统，不能轻易的利用市场上通用部件

成本较高

- 盘阵造价昂贵，导致系统建设成本居高不下

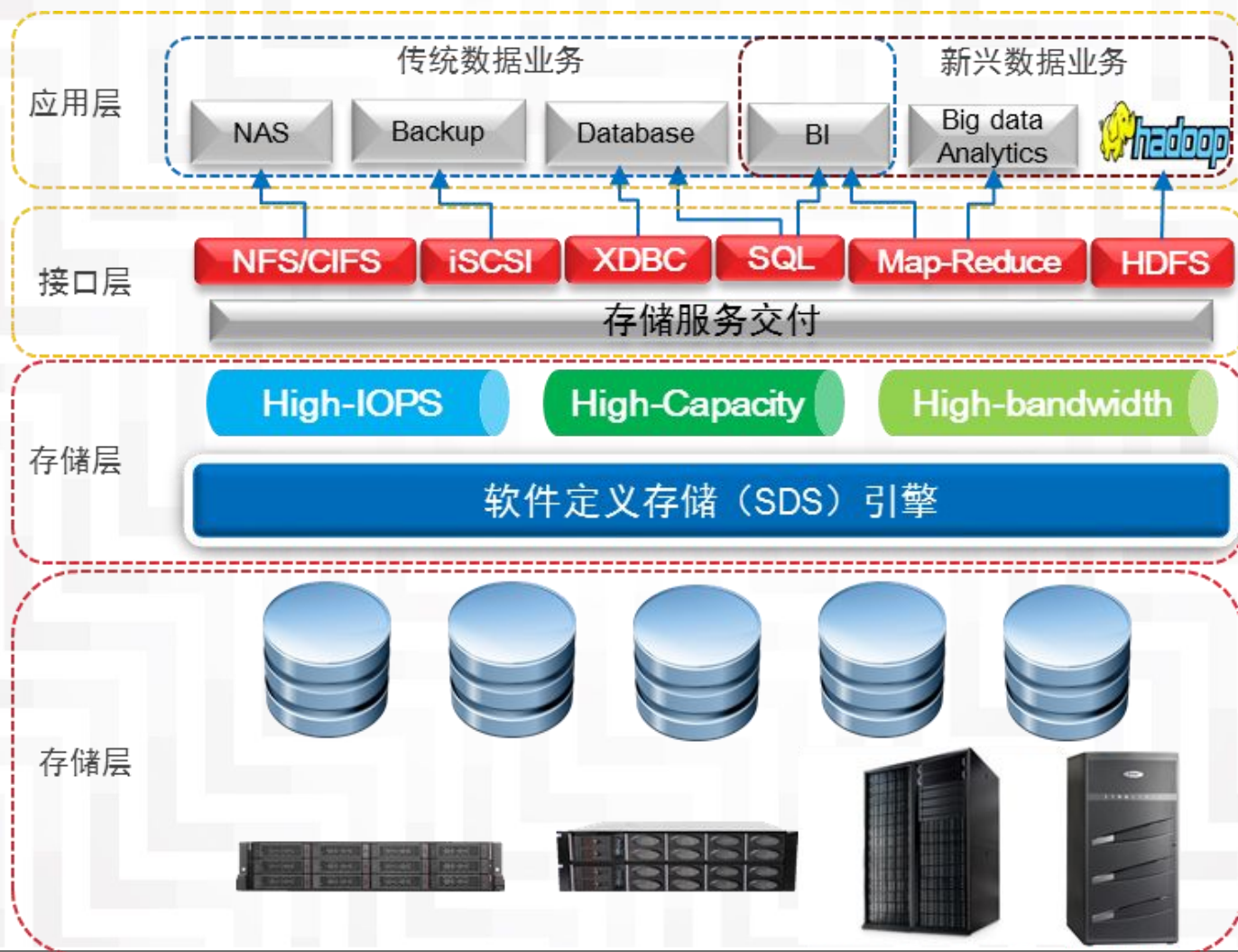
扩展性较差

- 传统盘阵单点扩展存在容量上限和接口带宽等限制，面对PB级的海量存储需求，传统的SAN或NAS在容量和性能的扩展上无法满足



软件定义存储SDS提供灵活、高效的数据解决方案

抛弃专有硬件设备的束缚，采用开放的灵活架构，提供存储级的性能和可管理性



软件定义存储的特质：

- 软件在存储中处于一个核心的地位，而底层的硬件多用通用的**服务器**和**磁盘**；
- 对资源和数据实现高效、整体上的有效管理；
- 具有灵活性和高可扩展性的特质，在对资源和数据进行整体把控的基础上，通过软件实现灵活



Lenovo



联想LeoStor并行存储系统



软件定义存储：由软件引擎和x86服务器组成

→ 软件定义的灵活性，轻松应对复杂应用环境



Scale out存储：横向按需扩展方式，性能线性增长

→ 从100TB到100PB无缝扩展



分布式存储架构：数据保存在多个存储节点上

→ 跨节点实时热备，可靠率高达99.9999%

→ 无传统存储的单一数据挂载点I/O热点问题



优化TCO，运维管理简单

→ 高性价比的通用x86架构，one view管理界面

LeoStor云存储体系架构





LeoStor并行存储系统产品系列

P 系列

- 高性能版
- 每PB可输出100GBps
- Lenovo RD650 集群存储服务器构成

高性能计算：

- 石油勘探、电力仿真、气象预测、医药筛选、汽车航空的风动实验、科学院的天体物理；

大数据分析：

- 公安图片、指纹、汽车影像追逃；
- 金融行业量化实时风控分析；
- 影视同期制作和4K后期渲染；

C 系列

- 大容量和性能平衡版
- 每PB可输出14GBps
- Lenovo RD450X 集群存储服务器构成

高性能计算：

- 高校数学、物理、化学等专业学院；

云计算统一存储：

- 虚拟化后端的存储

海量存储：

- 广电和电视台的媒资库，设计院、图书馆、政府档案、金融电子票据的文档库、医疗PACS系统、高清视频集中存储、智慧城市；



统一存储平台，满足数据中心多种需求

媒
资
库

非
现
编

O
T
T
视
频

日
志

大
数
据
分
析

V
M
虚
拟
化

数
据
库

LeoStor

从100TB到100PB的统一解决方案

Lenovo



LeoStor并行存储系统技术规格

优异的体系架构

控制器节点：

- 可配置2/4/6/8 ... 256个控制器，；
- 单控制器的缓存32G~512GB；
- 所有控制器都同时工作，且支持异构和控制器升级；
- 所管理的存储节点数量不限；

存储节点：

- 单存储节点可接入128块SATA盘,缓存32G~512GB；
- 支持盘体内插和盘体外插两种模式，无微码硬盘；
- 直接响应数据请求，不再经过控制器，减少瓶颈；
- 所有存储节点都同时工作，且支持异构；
- 可利旧原有NAS、IP-SAN和FC-SAN设备；

网络及协议：

- 支持GE、10GE、40GE、IB网络接口；
- 支持LUN块设备输出协议、标准NAS协议和快速NAS协议；
- 支持冗余网络和802.3ad链路冗余和聚合；

优异的性能和可靠性

性能指标：

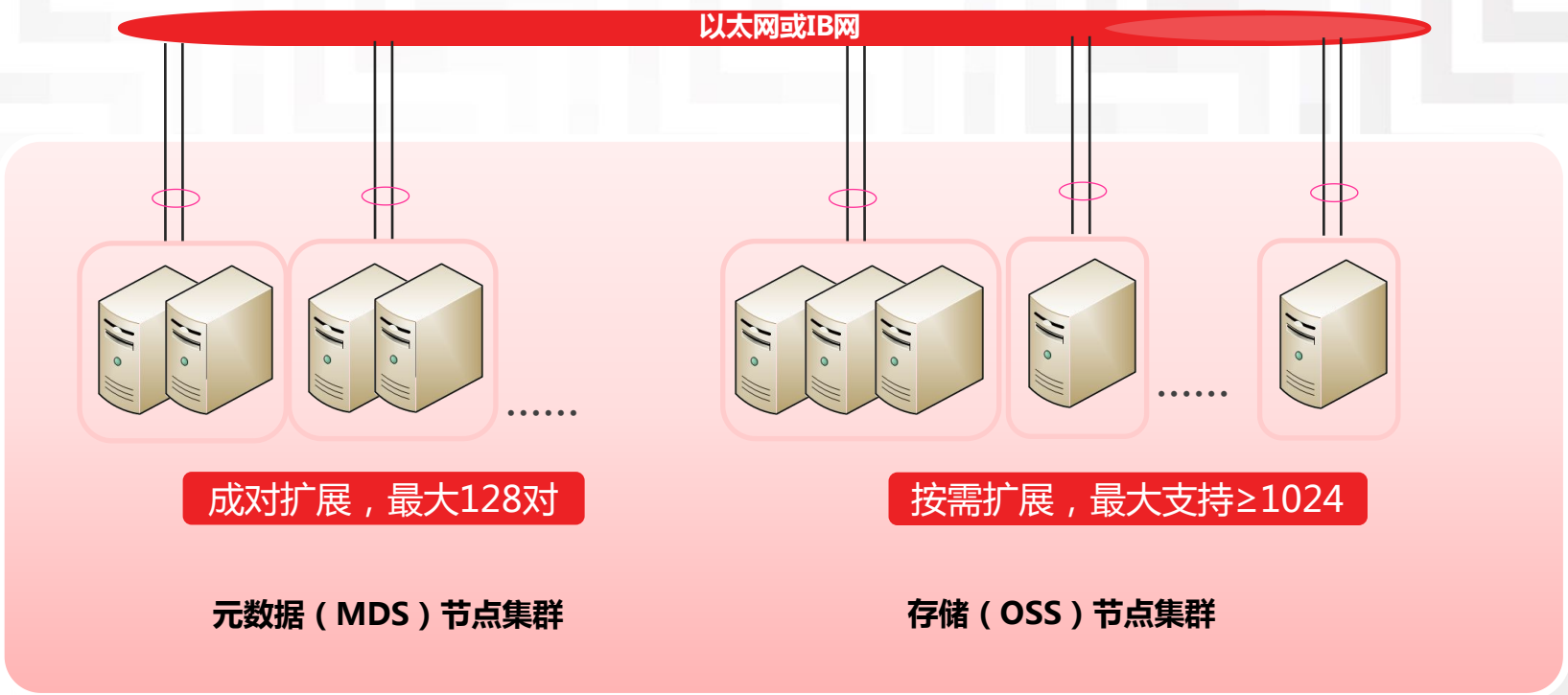
- I/O带宽输出：不少于100GB/秒；
- IOPS并发打开文件；不少于100万次/秒；
- 响应时间：小于1ms；
- 单卷容量：不少于100PB；

数据保护：

- 跨存储节点的副本、N+M纠错码，最大可利用空间80%；
- 无需节点内硬件RAID保护；



LeoStor并行存储系统典型部署架构





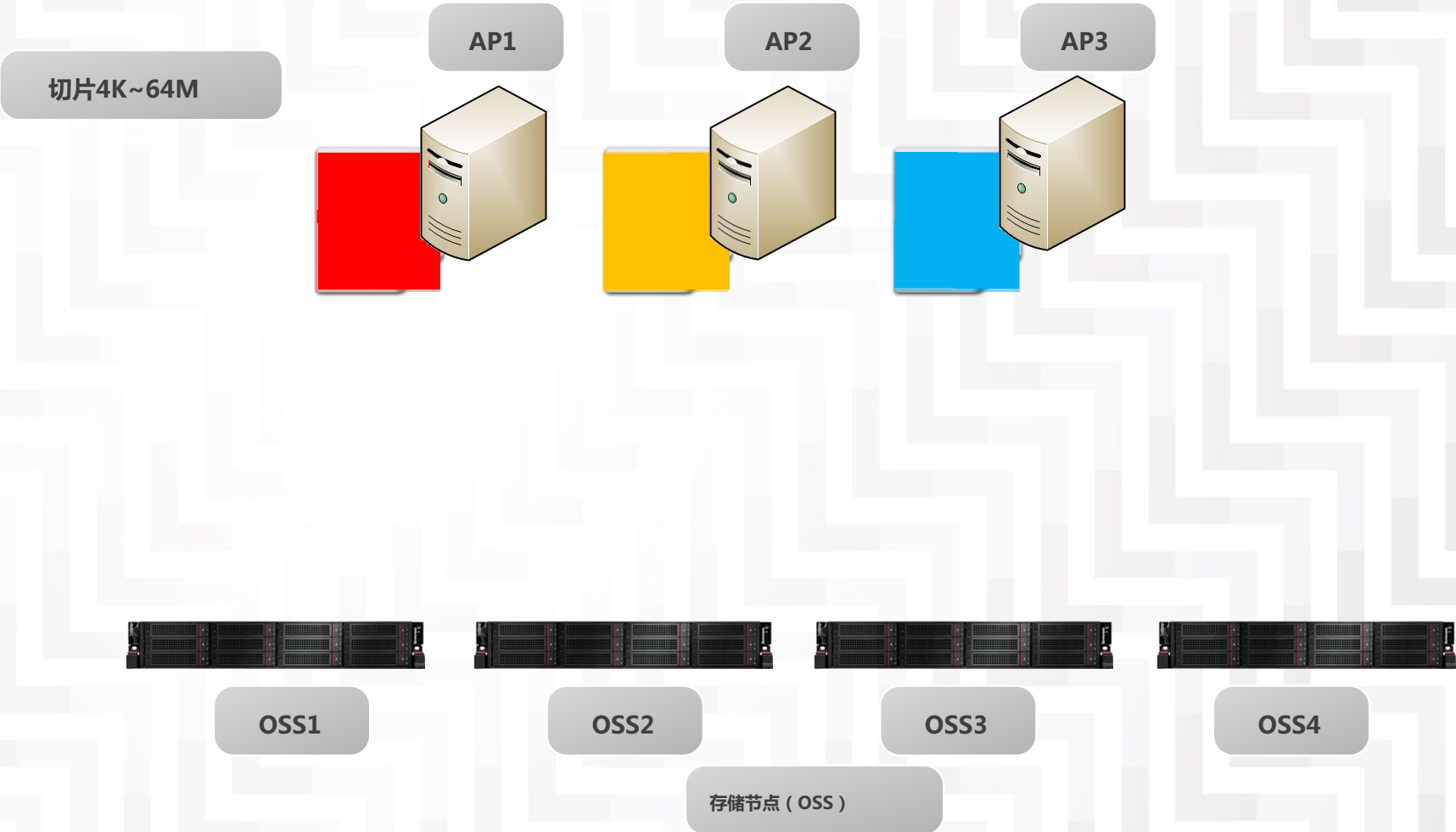
LeoStor提供多种类型接口



*POSIX 表示可移植操作系统接口 (Portable Operating System Interface, 缩写为 POSIX)



LeoStor 智能前端切片技术





数据写入模式


Servers


客户端算法计算元数据




Servers




Servers



客户端将元数据信息一步定位写入
元数据返回数据切片以及分布信息


交换机



MDS元数据节点



存储服务器



OSS数据节点

元数据服务器通过负载均衡算法进行负载均衡运算



数据读取模式



Servers



Servers



Servers

客户端算法计算文件元数据信息所在位置

客户端一步步定位读取文件元数据信息

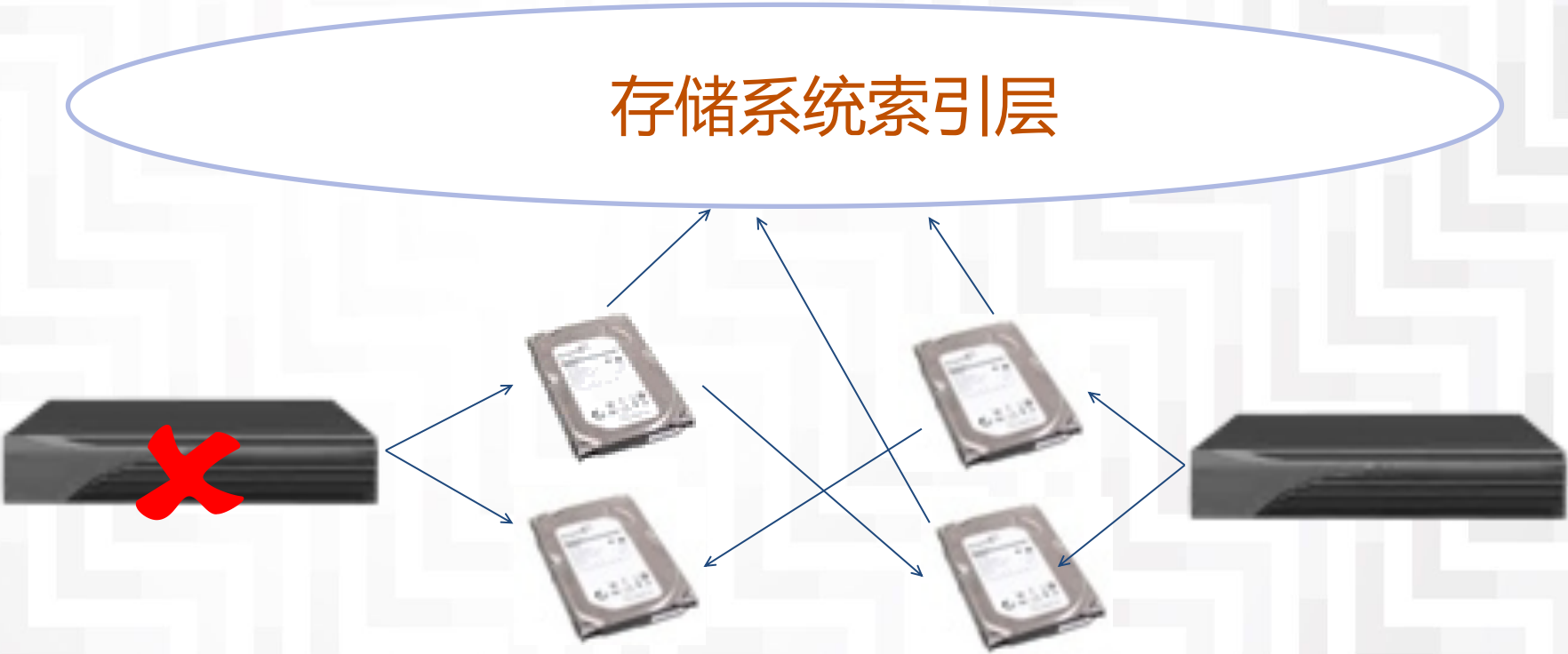
MDS元数据节点



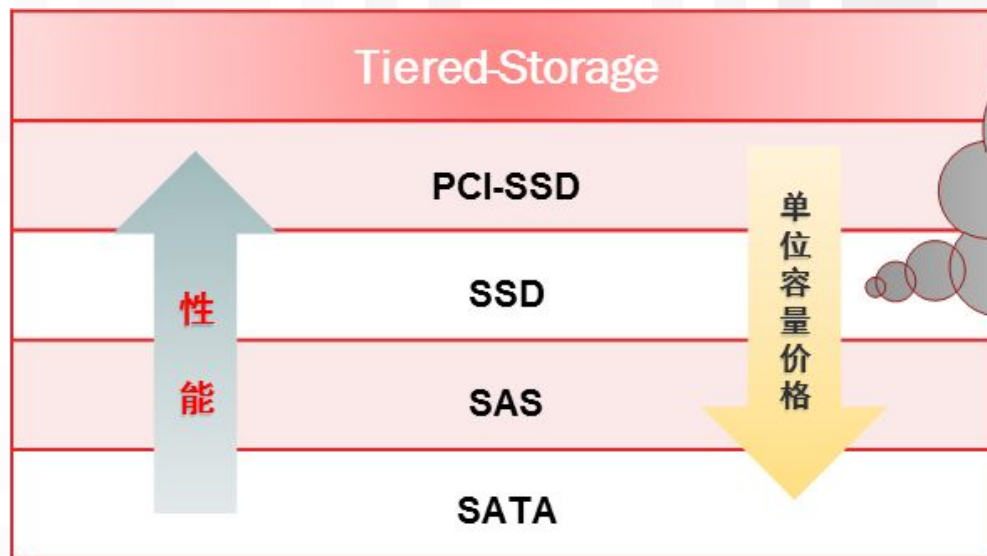
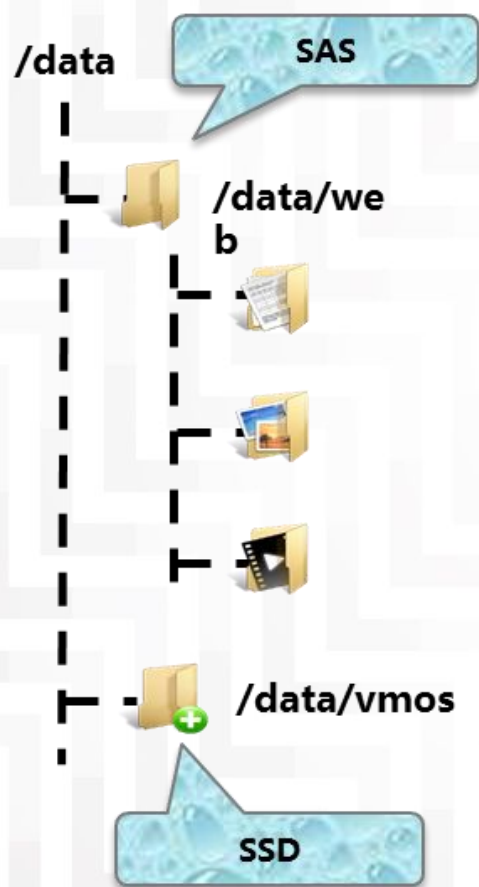
OSS数据节点



元数据工作模式



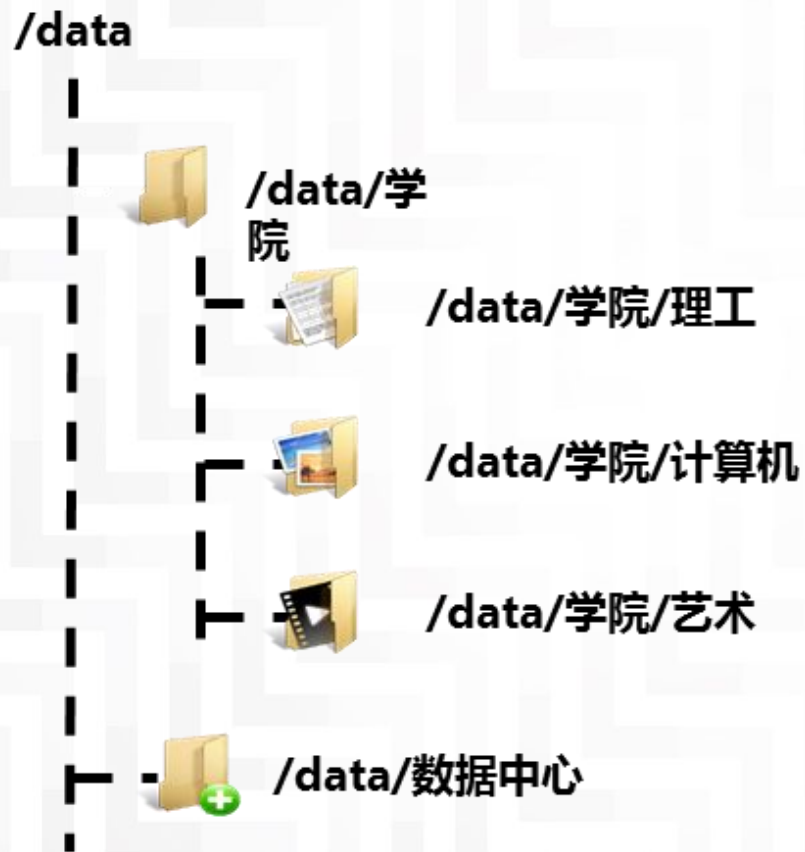
+ Tiered Storage 基于目录的分层存储技术



热点数据和VM
镜像文件保存在
SSD中

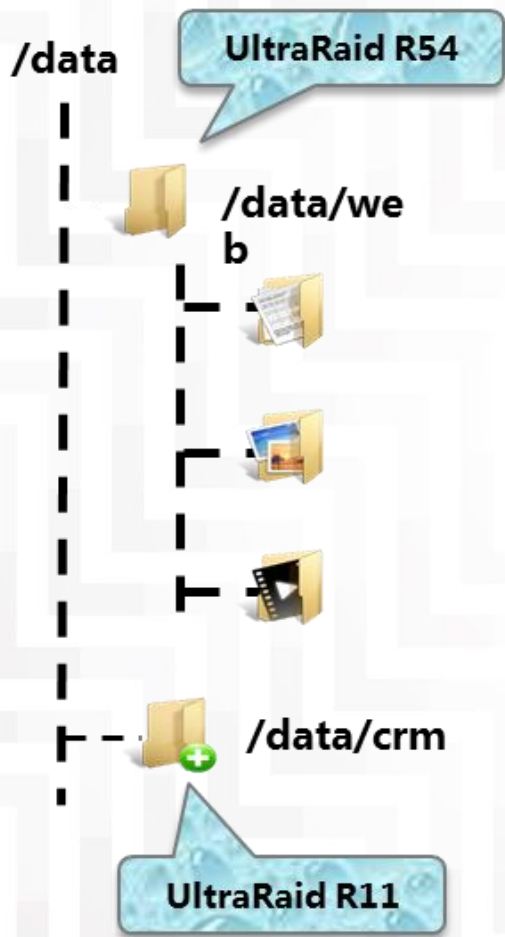


LeoStor基于目录和用户的配额管理



根目录	物理容量	分配容量	实际容量
/data	200TB	250TB	75TB
/data/学院		150TB	45TB
/data/学院/理工		50TB	10TB
/data/学院/计算机		50TB	15TB
/data/学院/艺术		50TB	20TB
/data/数据中心		100TB	30TB

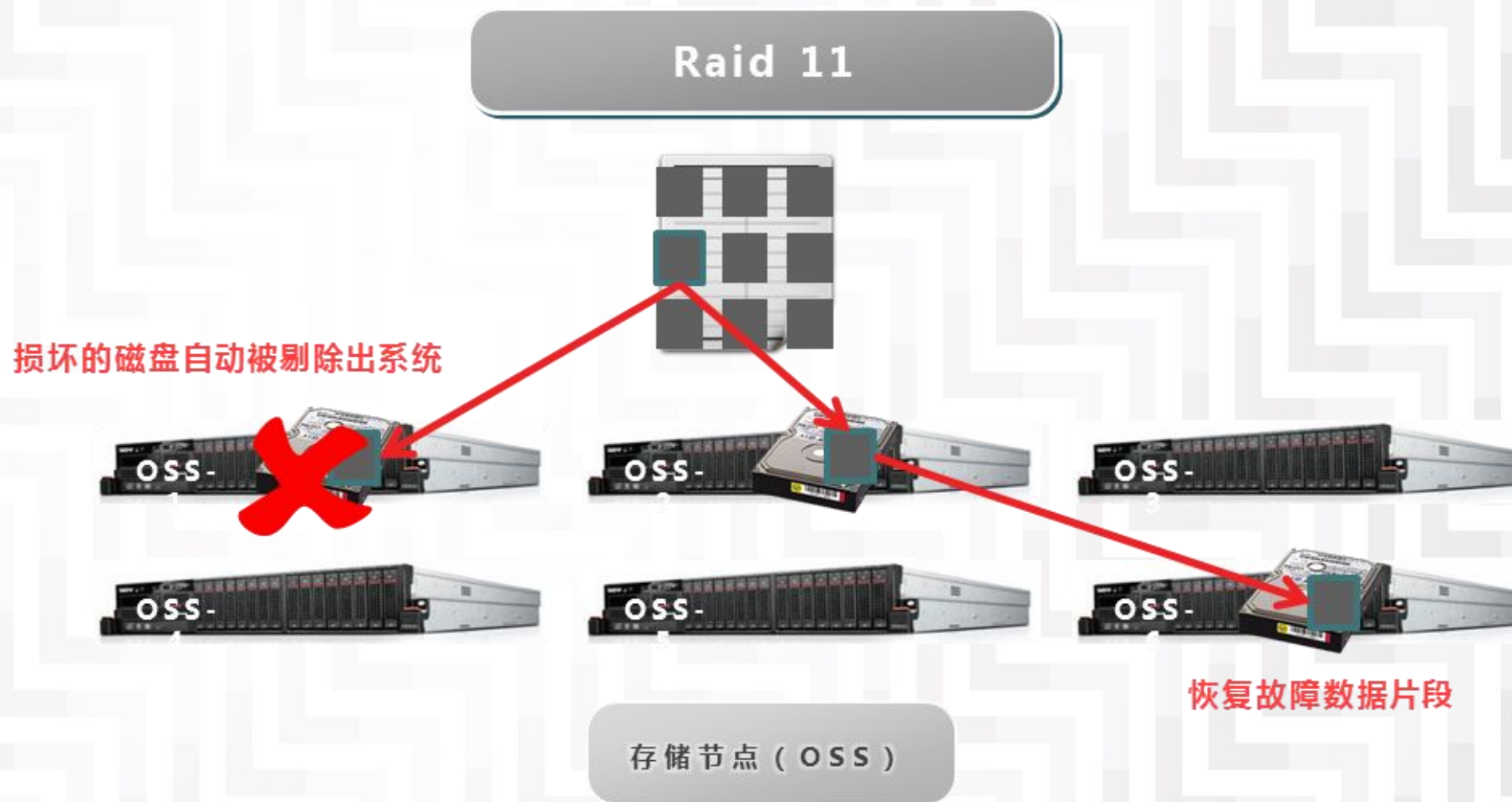
UltraRaid 基于目录的数据保护



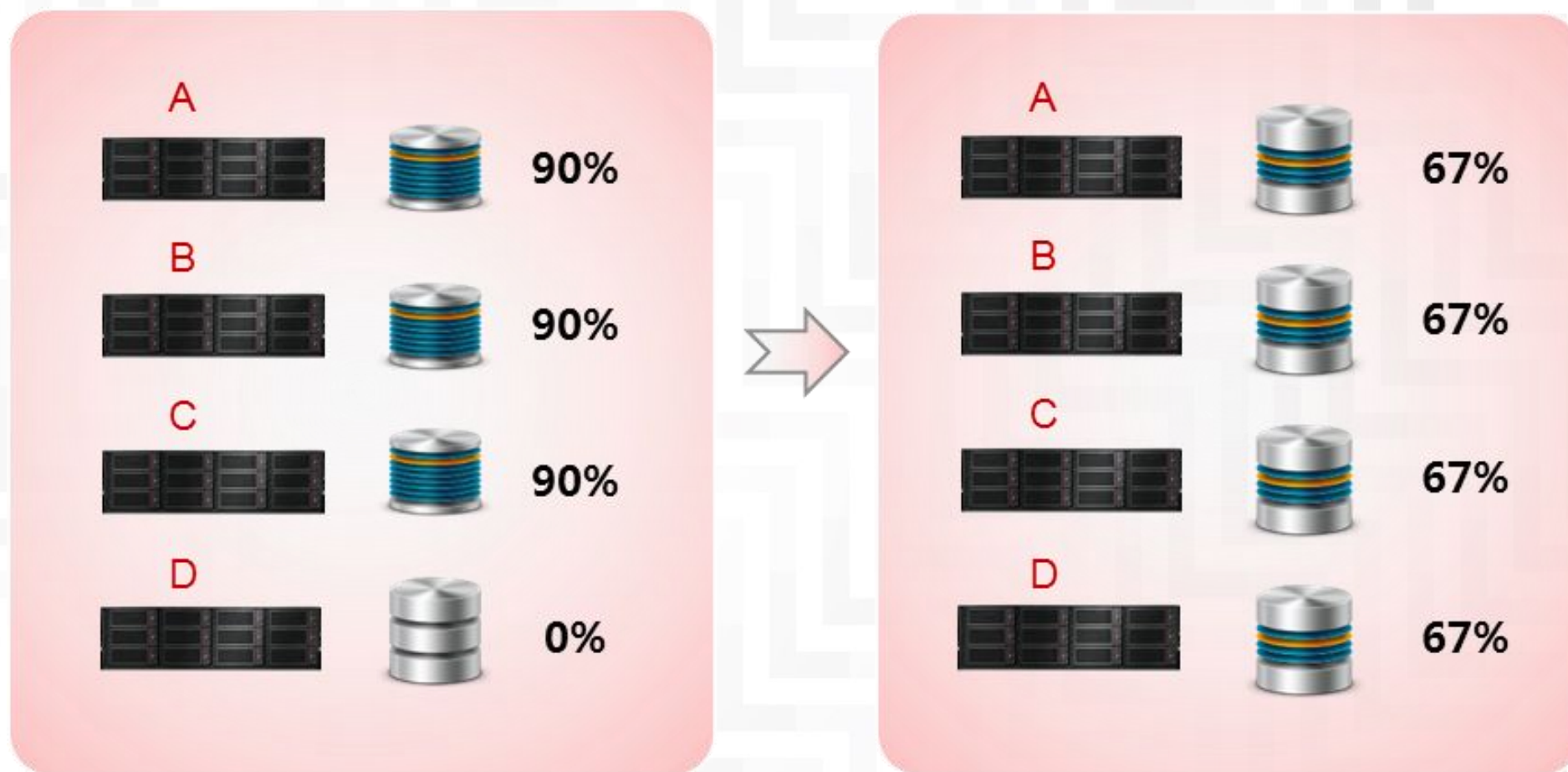
UltraRaid (副本和纠删码)				
类型	数据	备份/校验	备注	空间利用率
UltraRaid R11			一份数据一份拷贝	50%
UltraRaid R12			一份数据二份拷贝	33%
UltraRaid R52			二份数据一个校验值	66%
UltraRaid R54			四份数据一个校验值	80%

基于跨设备硬盘粒度的数据保护技术

LeoStor 数据自我保护



LeoStor的存储容量均衡技术



系统自动将数据均匀分配，可无需人工干预，同时增加系统输出带宽！



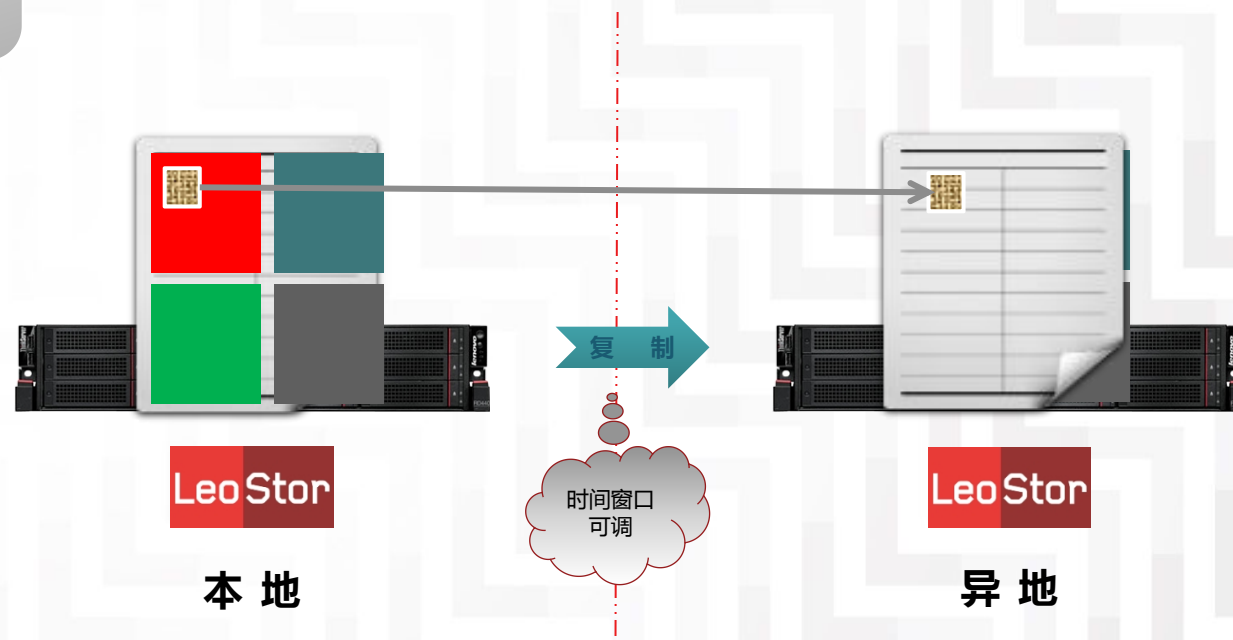
LeoStor智能复制技术支持构筑多活云数据中心



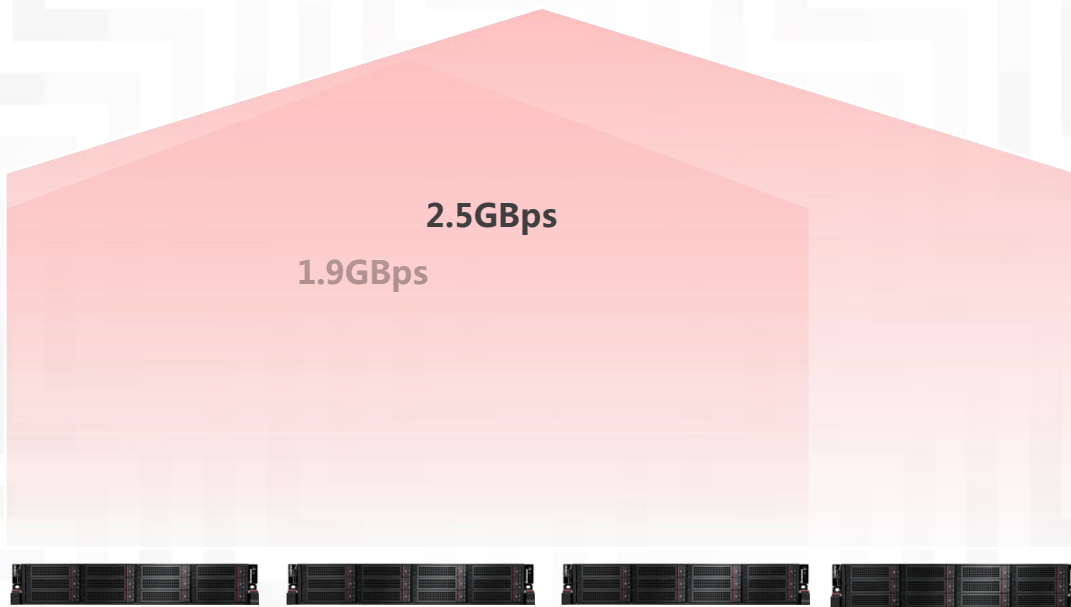


基于智能碎片技术的异步复制方案

切片4K~64M



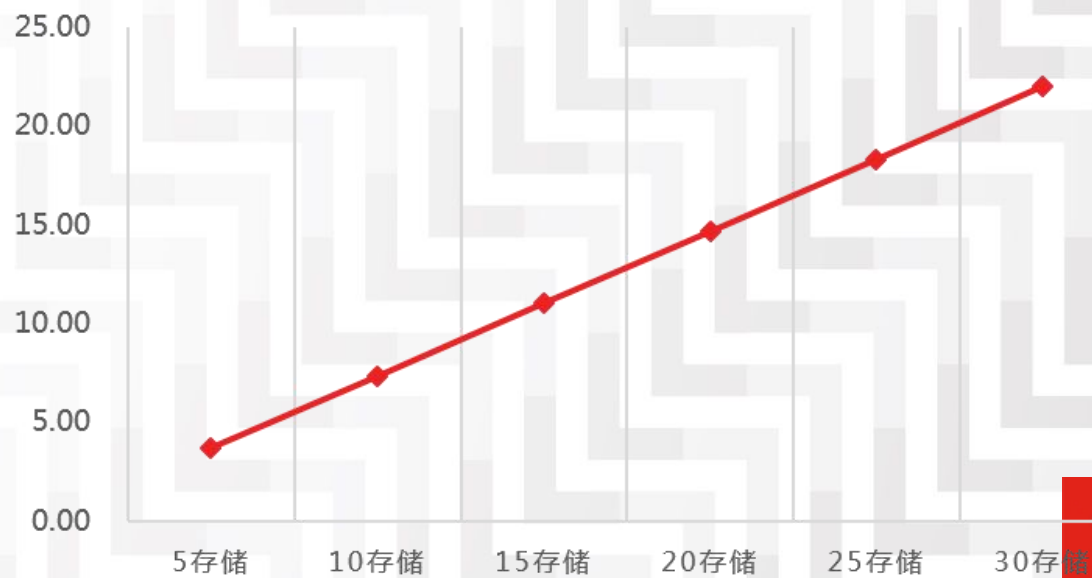
LeoStor带宽聚合技术



存储节点 (OSS)

LeoStor

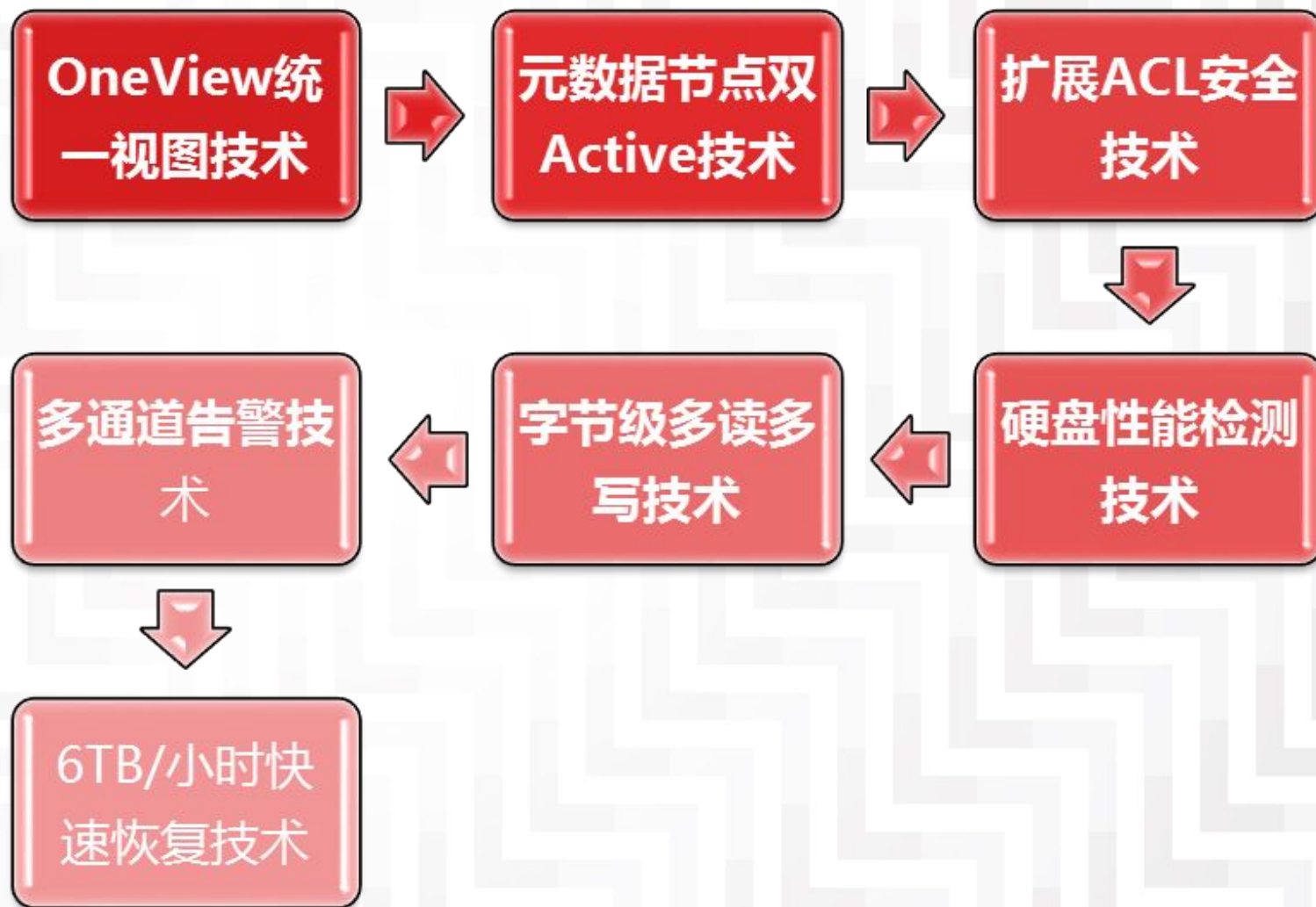
带宽 GB/S



Lenovo



LeoStor还有更多的技术特点





测试配置概述

	联想	H厂商	L厂商
节点配置	2元数据+3存储节点	4个节点	2元数据+3存储节点
数据磁盘个数 (7.2K)	54块	140	36
网络配置(每节点)	2*10Gb	2*10Gb	2*10Gb
备注			

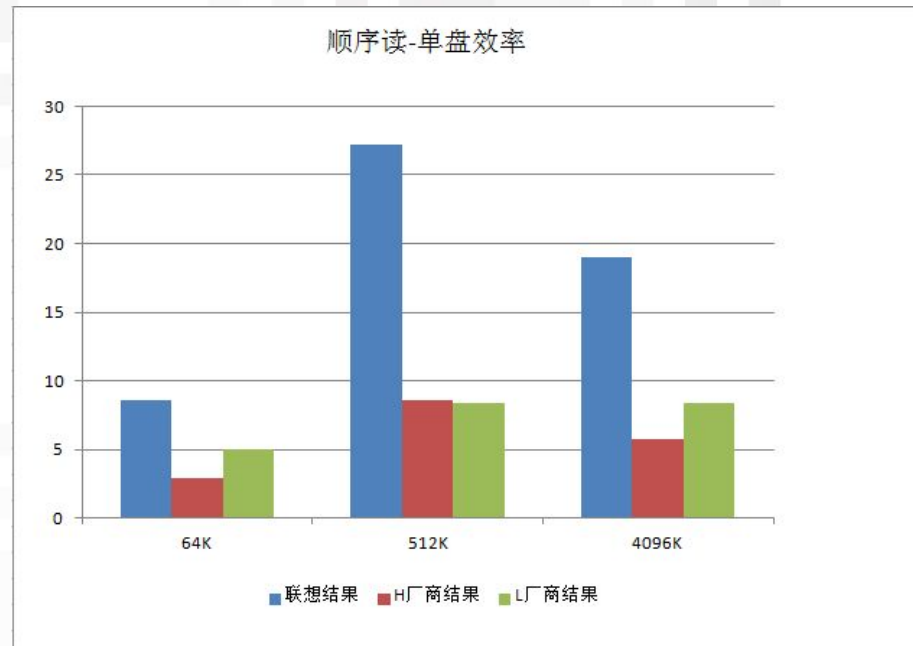
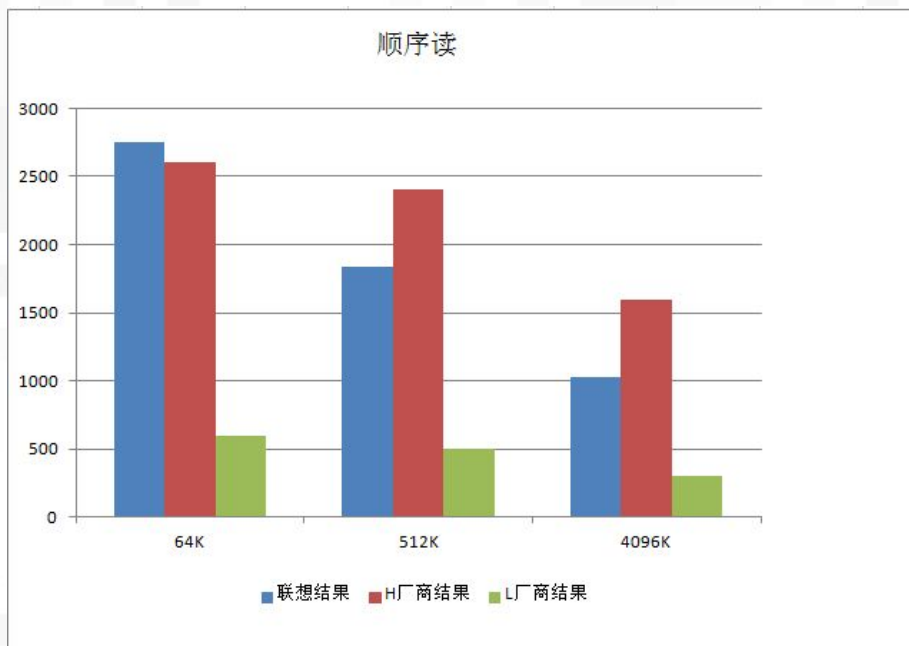




测试环境概述

	名称	个数	备注
测试节点	Window2008 server	5	均为万兆测试环境
性能测试工具	AutoTest	5	专业测试工具
延时测试工具	Dmwdrp	1	专业测试工具

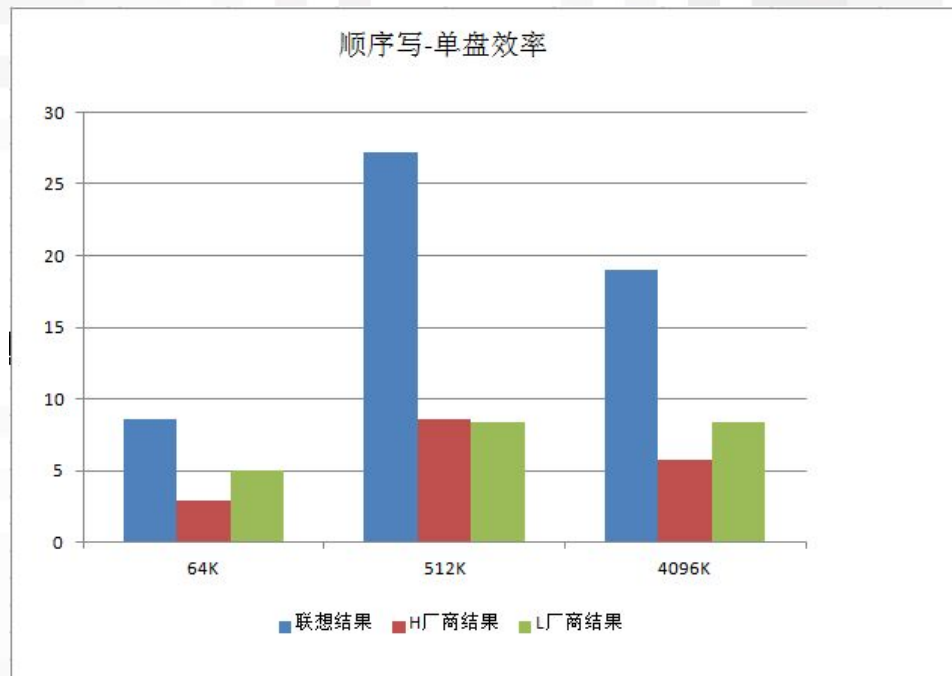
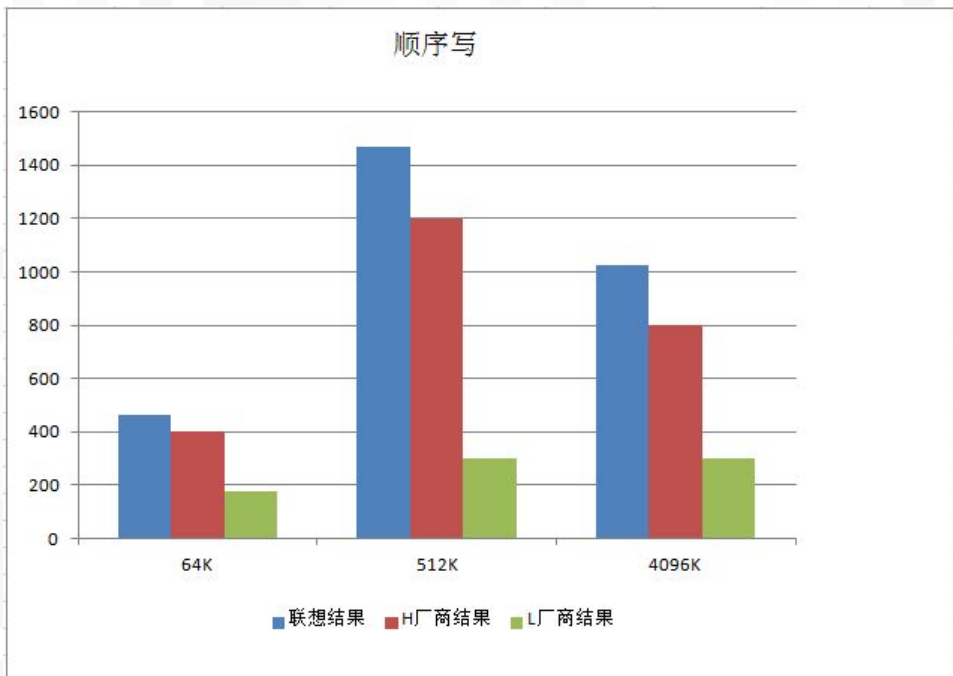
测试 results 分析-顺序读



结果分析：

- 总带宽上，联想产品在配置较少的情况下，与H友商总性能接近，表现优异；
- 单盘效率上，Leostor有较为明显的优势。

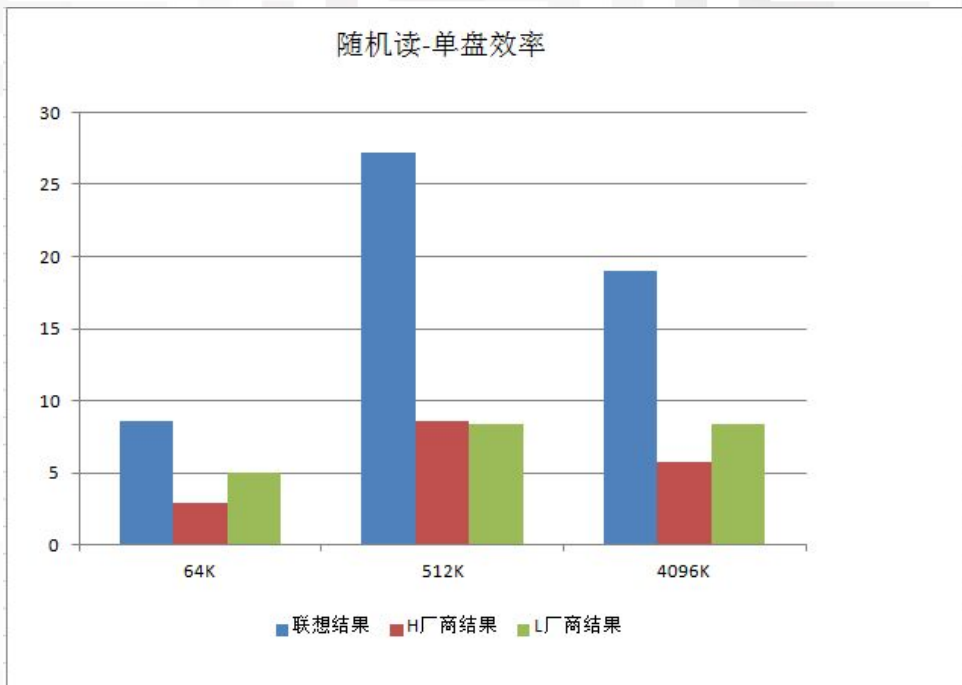
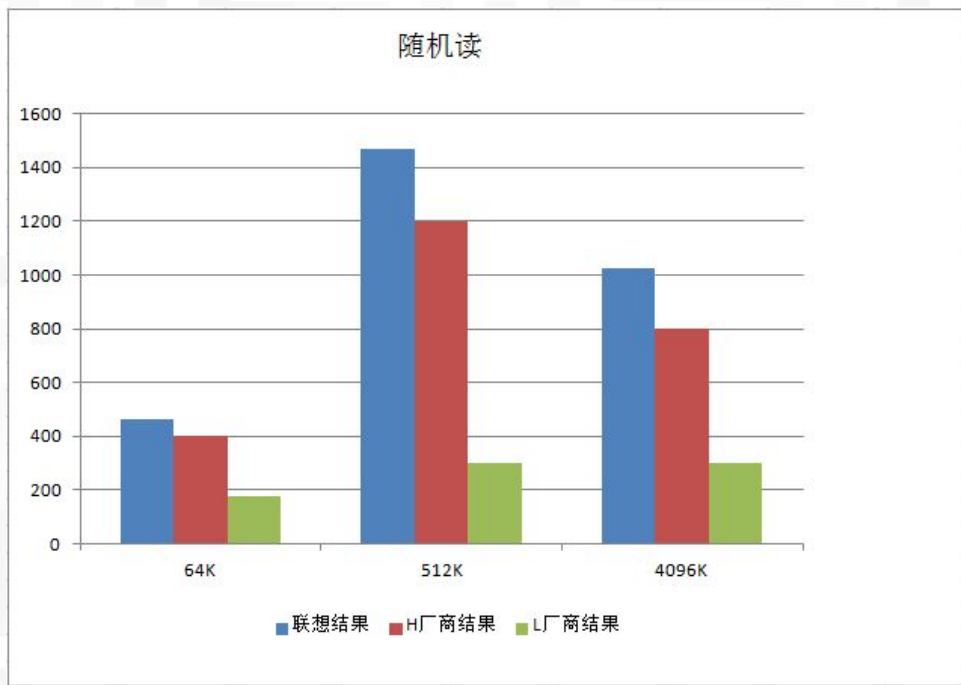
测试结果分析-顺序写



结果分析：

- 总带宽上，联想产品在配置较少的情况下，优于两个友商，表现优异；
- 单盘效率上，Leostor有较为明显的优势。

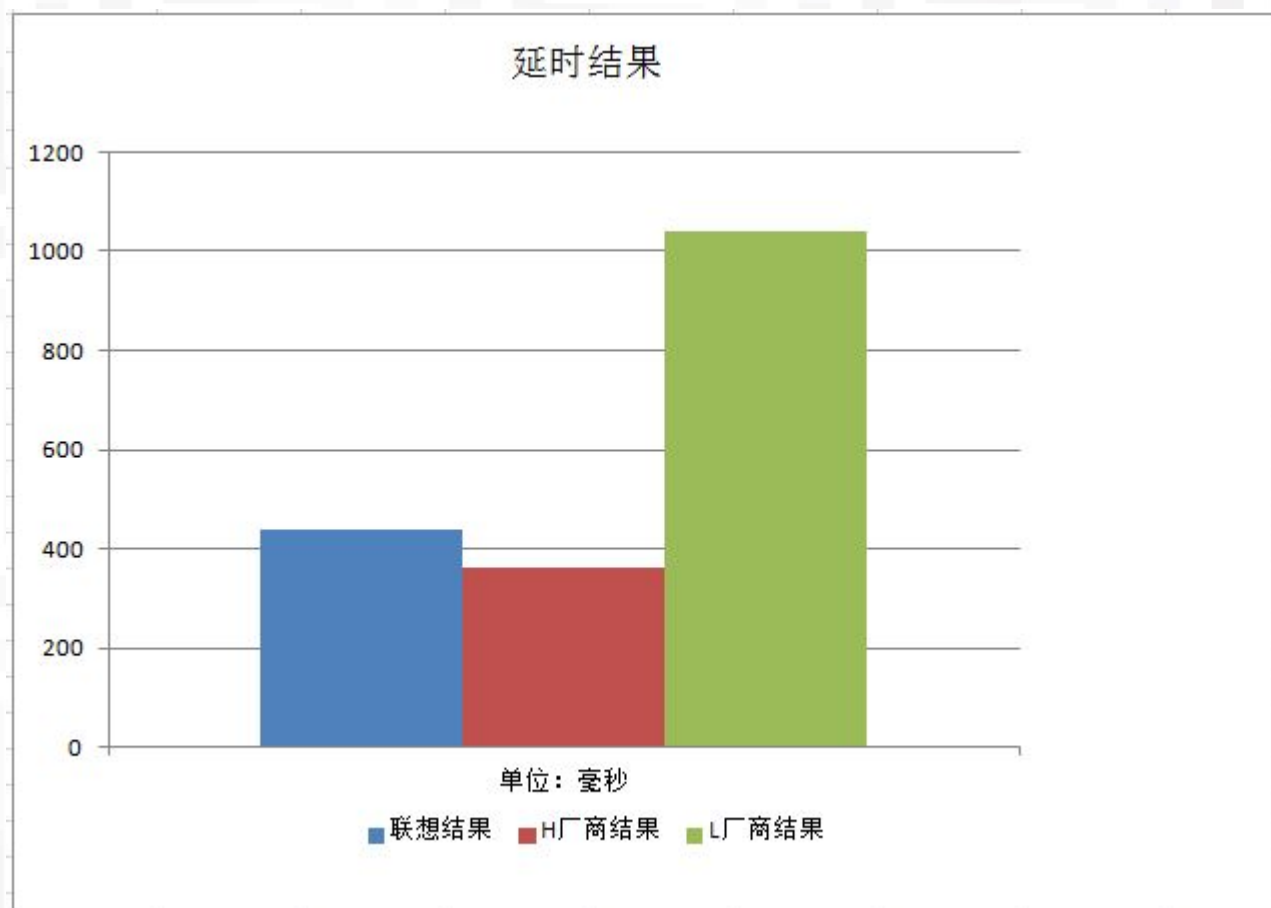
测试结果分析-随机读



结果分析：

- 总带宽上，联想产品在配置较少的情况下，优于两个友商，表现优异；
- 单盘效率上，Leostor有较为明显的优势。

测试结果分析-延时结果



结果分析：

- 联想产品在延时测试中，也取得了优异的成绩

竞争分析——VS NAS

架构参数	NAS	LoongStore
带宽	NAS设备控制器存在数量限制，即数据出口少,容易成为性能瓶颈	基于高效的并行存储技术，聚合带宽随着存储集群规模扩大而线性增长
可靠性	硬件HA备份设计；数据Raid5数据保护；Raid组内两块硬盘损坏，或者单台存储设备失效，会造成数据丢失	硬件系统集群化；数据冗余度可调；可以允许多块磁盘同时损坏，甚至单台存储设备失效也不会影响数据的完整性
并发	所谓双Active集群文件服务器,一个数据对象同时只有一台NFS服务器能够访问	一单个数据对象可以同时被多个数据接收端接收数据
扩展性	不支持在线扩展，扩展能力有限，扩展到一定程度后会造成性能的降低.容量越大，性能越低	可无限的扩展,动态无缝的扩展，扩展之后性能得到提高.容量越大，性能越高
文件规模	单目录十万级文件，超过十万目录无法打开，文件规模庞大时，进行数据更改延迟时间长，需要进行控制器数据同步	支持2000亿所以上的文件，快速检索，单目录千万以上
故障恢复时间	系统重构需要中断业务，Raid重建影响性能，时间很长	在线系统快速重构和数据恢复，速度是RAID的数倍，对性能几乎没有影响
维护升级费用	厂家专用磁盘，价格十分昂贵；技术支持由代理商完成，不能提供用户环境调优和定制开发	系统自恢复，接近零维护；维护升级成本低；可进行用户定制开发

竞争分析——VS FC SAN

架构参数	FC SAN	LoongStore
共享性能	FC SAN设备本身不支持共享，需要共享文件系统支持，如StorNext，或EMC自己的MPFS，但这次文件系统都不是集群设计，性能普遍很低，支持客户端数量也很有限。硬件上，单个LUN被共享访问时性能会急剧下降	基于高效的并行存储技术，系统聚合带宽随着存储规模扩大而近线性增长，带宽可扩展至数百GB/s；单个数据对象可以被多个客户端访问保持性能不变
扩展性	仅有Symmtrix DMX3/4支持在线扩展Raid阵列，其余产品扩容需要停机，重新规划空间分配；容量越大，性能越低	动态无缝的扩展，扩展之后性能得到提高，容量越大，性能越高
可靠性	硬件HA备份设计；数据Raid5数据保护；Raid组内两块硬盘损坏，或者单台存储设备失效，会造成数据丢失	硬件系统集群化；数据冗余度可调；可以允许多块磁盘同时损坏，甚至单台存储设备失效也不会影响数据的完整性
故障恢复时间	系统重构需要中断业务，时间很长	在线系统快速重构和数据恢复，速度是SAN的数倍，对性能几乎没有影响
完备性	SAN设备本身只提供基本存储空间，共享、镜像、安全都需要另购EMC或第三方软件，要实现一个完整的存储解决方案，需要一个“系统集成”的过程	系统提供存储、共享、自修复、负载均衡、安全、镜像等完整的功能，无需另购软件
维护升级费用	1、厂家私有光纤协议，需要专人学习、负责维护2、厂家专用磁盘，价格十分昂贵；3、增加客户端需要收费昂贵	通用的TCP/IP协议，任何网络维护人员都能承担维护任务；系统自恢复，接近零维护；维护升级成本低

竞争分析——VS Hadoop FS

架构参数	HDFS	LoongStore
开发语言	Java 应用程序级别	C 内核级模块
定位	处理型应用	通用大规模文件共享存储
应用接口	Java api	Posix标准文件系统接口
	不支持文件修改 write-once-read-many	
	只能有一位写用户	
	不支持软、硬链接	
文件类型优化	大文件	大、小文件
文件数量支持	百万级（白皮书公布）	超过300亿（案例值）
扩容后数据迁移	不支持	支持
名字节点（MDS）	单台，存在单点故障	集群（1~128台）
兼容性	跨平台	跨平台

竞争分析——VS Lustre

架构参数	Luster	LoongStore
元数据服务器	Active-Standby 2台服务器+SAN (为达到高可用效果需要共享存储 , 如SAN)	Active-Active , 最大支持128台服务器+内置硬盘 (任何类型的磁盘)
可靠性方法	磁盘 : 依靠硬件RAID保护 服务 : 后端硬件用SAN , IO服务器之间failover ; 任何存储节点宕机 会影响服务连续性	磁盘与服务连续性 : 通过使用文件副本策略 , 消除磁盘、存储节点、盘阵 的单点故障
聚合带宽	上百GB/s	上百GB/s
文件数量	1亿	2000亿 , 实际案例300亿文件
可靠性	屏蔽了oss宕机造成影响(Failover) , 如果控制器损坏或者RAID里面多 块磁盘损坏会导致数据丢失	屏蔽了存储服务器、网络、硬盘、控 制器损坏等各类异常 (通过文件副本 策略 , 实现数据的全局恢复)
扩展支持	难以支持在线扩展 , 但是没有数据 迁移功能 , 扩容后数据分配不均	支持在线动态扩展 , 扩展后可以数据 迁移 , 重新平均数据分布 (不停应用)
支持平台	Linux	Linux、Windows、Mac



竞争分析——VS Panasas

架构参数	Panasas	LoongStore
架构重点	存储系统只针对大文件读写做了优化，无小文件优化策略	分别针对大文件和小文件读写做了双重优化
硬件	刀片服务器，每台刀片带2块存储磁盘，硬件要求非常高，硬件要求较高	通用机架式服务器，每台服务器可带1-36块磁盘，硬件要求较低
单卷容量	PB级（官方网站给出）	300PB
数据保护	分层校验，通过动态复制数据段进行数据块损坏时数据保护，单点故障与整个磁盘损坏则完全通过RAID机制加以实现	文件系统可自动监测到丢失或损坏的数据块，并根据数据块进行恢复；可根据应用需求对存储系统中的目录分别部署不同的数据保护机制
数据丢失重建	500G数据重建时间在90分钟左右，且需要借助所有刀片同时参与数据恢复，影响整体性能。	750G磁盘数据恢复时间为15分钟且数据恢复时对前端没有丝毫影响
成本	高	低
支持平台	Linux	Linux、Windows、MAC



竞争分析——VS StorNext SNFS

架构参数	StorNext SNFS	LoongStore
性能	SNFS后端基于的是SAN环境，读写性能完全依赖于盘阵的存储控制器，存储控制器很容易成为性能瓶颈。在实际的测试中也发现即使是中高端的磁盘阵列，随着应用并发数的提高性能会急剧下降，这也导致SNFS文件共享访问效率下降；	LoongStore使用了完全分布式的数据分布访问的方法来消除对于单个存储控制器性能的依赖。可以充分利用集群环境里面硬件的并发处理能力，提供远远高于传统存储方案的聚合吞吐量；
可靠性	SNFS方案的可靠性完全是依赖于底层硬件RAID以及相应的通道冗余实现。而RAID的保护只能局限于单个盘阵内，一旦单个盘阵失效将会导致数据无法访问甚至数据丢失等严重故障；	LoongStore的可靠性机制不会完全依赖于高端、高可靠性的硬件，而是在系统内部提供了一套完整的故障自动探测与恢复机制来消除硬件损坏可能带来的损失，在无需人员参与的情况下保证数据的安全与服务的连续性。
文件管理效率	SNFS的单个文件系统只有一个活跃的MDC提供对应的服务，当应用环境里面存在大量的文件时，前端的工作站的文件检索请求很容易会给MDC造成巨大的访问压力，从而降低了整体的文件管理效率；	LoongStore的单个共享文件系统可以由多个元数据服务器构建的集群来支撑，即使面向存在海量文件的应用环境（数亿甚至数千万文件），LoongStore的元数据服务器集群也能够轻松支撑住庞大的文件检索请求；

竞争分析——VS Isilon OneFS

架构参数	Isilon OneFS	LoongStore
架构	昂贵的Infiniband网络，复杂的相互连接，数据存储服务器最大支持144台	通用服务器和网络，存储服务器可支持超过4096台
可扩展性	10PB（白皮书）	300PB
存储特点	大文件	大小文件
数据保护	跨节点的Raid 5 奇偶校验任何磁盘或机箱故障会导致昂贵的再分配和再平衡的数据在所有参与的节点	多副本机制或者跨节点校验模式，可根据需求动态设置快速恢复时间，1T磁盘半小时内恢复完成
负载均衡	基于IP地址池的后端负载均衡	透明的客户端和存储服务器之间的负载均衡请求
安全性	数据在存储到数据服务器上之后进行切片分发，如果此时服务器出现损坏则会出现数据丢失	数据分发与切片全部在前端完成，只要数据存储到后端就能保证数据的安全性
元数据机制	元数据信息存放在每台数据服务器上，且每台数据服务器要存储所有的元数据信息，在通过后台进行复杂的数据交换，保证数据的一致性	分布式存放在多台元数据服务器中，每台元数据服务器的元数据管理不重复，通过独有元数据保护机制和校验算法保证数据的一致性



THANK YOU

DAKUJEM DANK BEDANKT MERCI TAKK 谢谢
ありがとう СПАСИБО GRACIAS DZIĘKUJĘ DANKE
OBRIGADO БЛАГОДАРЯ GRAZIE תודה GRACIAS



Lenovo™