漫谈分布式存储方案，GPFS 对话 CEPH

1. 数据中心云存储趋势

过去 20 年，传统的集中式数据存储经历过了它最辉煌的时代，云计算带来的底层存储架构变革，已经慢慢宣告这个王者时代的谢幕，接棒数据中心下一代存储黑科技的架构，是近年关注度更高的关键词，如分布式存储，对象存储，超融合等技术，在这其中，除了传统一线厂商的战略迭代产品，也不乏优秀的民间开源解决方案，经过多年的技术沉淀，而今也结束了野蛮纷争的草莽时代，各家产品划地为王，纷纷找到了它的用武之地，而其中的佼佼者，一个是商业软件界的霸主 GPFS，另一个是开源软件界的王者 CEPH。

2. GPFS 和 CEPH 的初次亮相

GPFS 关于我：

各位好，我叫 GPFS，是一个高性能的共享并行文件系统，自诞生起，就为高性能、数据共享、开放、安全而生。我今年 23 岁，是一名 95 后，两年前，为了更好的融入 IBM 光谱存储大家庭，我有了个更好听的名字——SPECTRUM SCALE，当然对于我来说，这不仅仅是名字的变更，也意味在我身上，增加了关于闪存、容灾、备份、云平台接入等诸多特性，我扮演的角色更加重要，职能定位也愈加明晰了。关于未来，我也有自己的想法，有更大的愿景，希望能和数据中心的其它小朋友们相处愉快，和谐。

CEPH 关于我：

大家好，我叫 ceph，是一个 00 后，今年 12 岁了，法律上还未成年，但我的公众形象已经十分成熟，我的名字来源于宠物章鱼的一个绰号，头像就是一只可爱的软体章鱼，有像章鱼触角一样并发的超能力。我平常主要活跃在云计算领域，经过多年的脱胎换骨，不断迭代，我积攒了良好的口碑，好用，稳定，关键还免费，我可以提供对象，块和文件级存储的接口，几乎可以覆盖所有…哇，说着说着突然感觉自己原来无所不能呢，谢谢大家这么关注我，当然，目前我还在长身体的阶段，很多特性在趋于完善，希望未来我们可以相互促进成长。

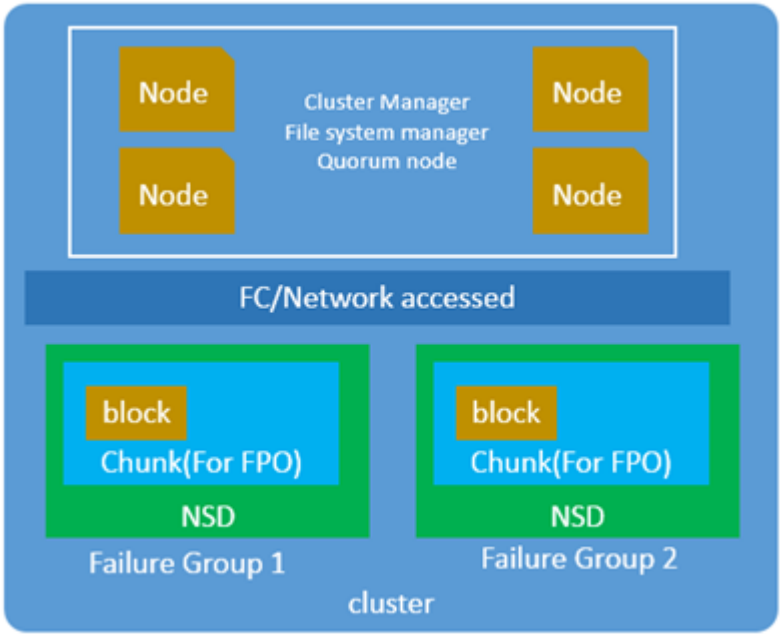
3. GPFS 的前世今生

作为一款成熟的商业产品，GPFS 的发展史早已百转千回了，在揭开 GPFS 的面纱之前，我们还是先来扫扫盲，复习一下在 GPFS 集群架构中涉及到的基本概念和组件。

架构解藕

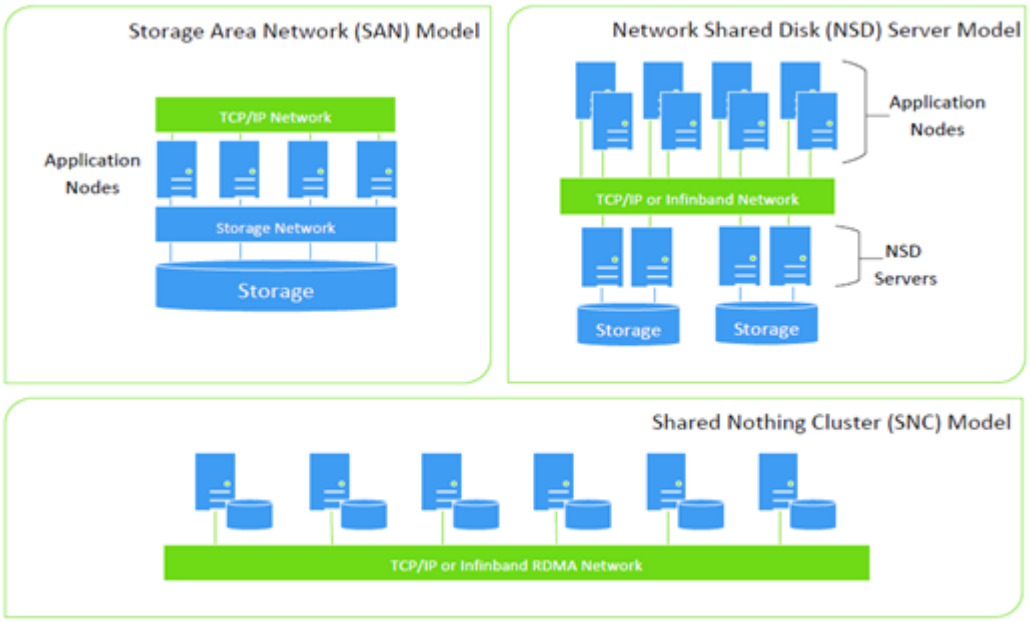
a) Cluster：GPFS 的组成架构，由一系列的节点和 NSD 组成，集群的配置文件通常保存在两台主备的节点上。
b) Node：安装了 GPFS 软件的主机，它可以通过直接或者通过网络访问其它节点的方式来访问存储，每个节点在集群配置中有不同的角色。
c) Cluster manager：负责整个集群配置的正确性和完整性，主要负责监控磁盘租约，检测节点故障和控制节点的故障恢复，共享配置信息，选举文件管理节点等任务。
d) File system manager: 维护文件系统中磁盘的可用性信息，管理磁盘空间，文件系统配置，磁盘配额等。
e) Block: 一个集群中单个 I/O 操作和空间分配的最大单位。
f) NSD: 提供全局数据访问的集群组件，如果节点和磁盘间没有直接连接，则 NSD 最好具有主服务节点和辅服务节点。
g) Chunk: FPO 架构中的概念，它是一组 block 块的集合，看起来像一个大的 block，一般用于大数据环境。
h) Failure Group: 一组共享故障的磁盘组，当其中一块盘失效时，整个组会同时失效。
i) Metadata: 包括集群配置信息和非用户数据。
j) Quorum Nodes: 用于保持集群活动的仲裁节点，一般有两种仲裁方式，节点仲裁和带 Tiebreakerdisk(心跳盘)的仲裁

上述组件如何有机的组合在一起提供存储服务呢，把以上组件拼接起来，就可以得到下图所示的集群大体架构：



使用方案

基本架构了解了，那怎么用呢？先祭出三张架构图，业内人士一看应该懂，不明白没关系，往下针对这几张图稍作解释：



左上：传统的 GPFS(Direct attach)使用场景，图中分为主机，交换机，存储三类角色，前端主机(即图中的 Application Nodes)类型可以是 AIX/Linux/Windows 服务器节点，存储使用集中式的存储架构，存储 lun 通过光纤交换机 Zoning 的方式映射给所有主机，数据传输走光纤网络，在前端主机中能识别到所有 lun 的块设备，并为其创建 NSD，主机与存储间通过 SAN 网络直接通信，不借助通过其它结点。

右上：混合的 GPFS 使用场景，是左上架构的拓展，跟左上不同的是，在该架构中，并非 GPFS 用到的所有 NSD 都映射给了所有的主机，我们给集群中的 NSD 添加了 NSD Server 的属性，这就决定了集群中一部分节点访问存储的方式，不是直接访问存储，而是通过 TCP/IP 网络访问 NSD Server 的方式来访问后端存储。主要适用于以下情形，1. 集群节点跨站点，但两个站点间没有足够稳定的光纤通道可以互通。2. 节省光纤资源。

下图：这张图是我们今天的主角，也是 GPFS 之所以拿来和 CEPH 对决的核心架构，我们在后面再展开说说。接下来说说应用场景。

应用场景

在传统 DB2 数据库双活方案 GDPC 的使用场景中，为了实现跨站点的双活+容灾，底层存储方案选用 GPFS，双站点架构中，两个站点均配备主机和存储资源，每个站点的存储形成一个 failure group， 远程访问对端存储采用 nsd server 的方式访问，两个 failure group

间完全冗余，任何一个站点出现故障都不影响文件系统的正常使用，并通过第三方站点的一台服务器和 nsd 作为仲裁节点，是真正意义上的双活。

GPFS 可以用来替代 HDFS 作为大数据的底层存储，GPFS FPO+Symphony 作为相对 Mapreduce 更领先的分布式计算框架，可以更灵活和支持和对接企业的 IT 使用场景。在 IBM 的部分企业级云产品中，GPFS FPO 也被用来作为私有云产品的底层存储来使用，用来存储虚机镜像和介质，这一点上使用和 CEPH 也极为相似。

4. CEPH 的发展之路

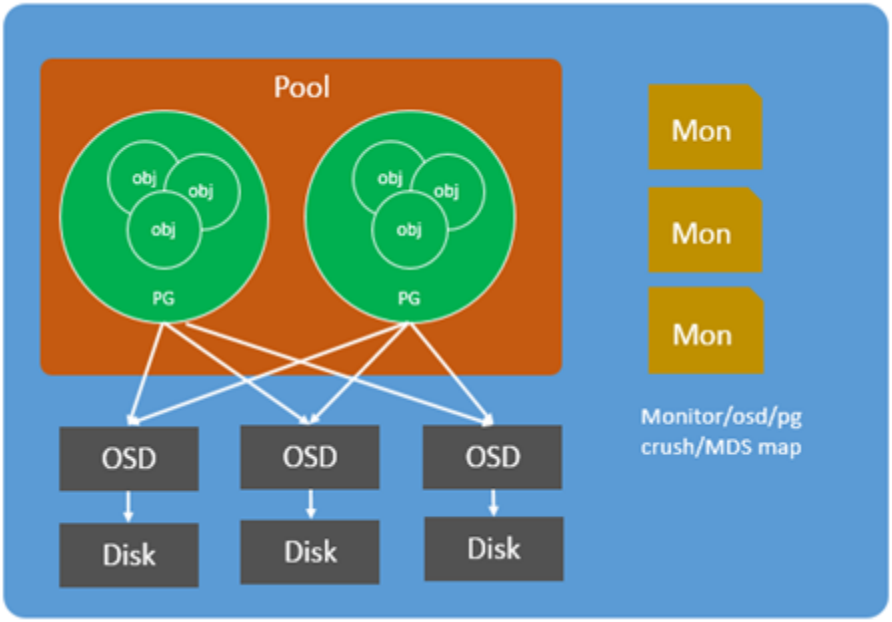
作为云计算的三架马车，网络，存储，管理平台，业界的开源方案里，网络层面 SDN 日渐成熟，管理平台上，Openstack 已经创造了一个时代，而 CEPH，无疑成为存储最犀利的开源解决方案。谈起它的架构之前，我们有必要先来了解以下这些概念，同时为了更加形象化，我们将部分组件对应到 GPFS 的组件上来理解，但请注意实际的功能和结构仍然差别巨大。

架构解藕

a) Ceph monitor——对应 quorum + cluster manager:保存 CEPH 的集群状态映射，维护集群的健康状态。它分别为每个组件维护映射信息，包括 OSD map、MON map、PG map 和 CRUSH map。所有群集节点都向 MON 节点汇报状态信息，并分享它们状态中的任何变化。Ceph monitor 不存储数据，这是 OSD 的任务。
b) OSD——对应 NSD: CEPH 的对象存储设备，只要应用程序向 Ceph 集群发出写操作，数据就会被以对象形式存储在 OSD 中。这是 Ceph 集群中唯一能存储用户数据的组件，同时用户也可以发送读命令来读取数据。通常，一个 OSD 守护进程会被绑定到集群中的一块物理磁盘，一块磁盘启动一个 OSD 进程，可以对应 GPFS 的 NSD 概念。
c) Pool：是存储对象的逻辑分区，它规定了数据冗余的类型和对应的副本分布策略，副本支持两种类型：副本（replicated）和 纠删码（ Erasure Code）
d) PG(placement group)——对应 Chunk：是一个放置策略组，它是对象的集合，该集合里的所有对象都具有相同的放置策略；简单点说就是相同 PG 内的对象都会放到相同的硬盘上； PG 是 ceph 的核心概念， 服务端数据均衡和恢复的最小粒度就是 PG；
e) MDS——对应 Filesystem manager：Ceph 元数据服务器，MDS 只为 CephFS 文件系统跟踪文件的层次结构和存储元数据。Ceph 块设备和 RADOS 并不需要元数据，因此也不需要 Ceph MDS 守护进程。MDS 不直接提供数据给客户端，从而消除了系统中的故障单点。
f) RADOS：RADOS 是 Ceph 存储集群的基础。在 Ceph 中，所有数据都以对象形式存储，并且无论是哪种数据类型，RADOS 对象存储都将负责保存这些对象。RADOS 层可以确保数据始终保持一致。要做到这一点，须执行数据复制、故障检测和恢复，以及数据迁移和在所有集群节点实现再平衡。

g) RBD：RADOS 块设备，提供持久块存储，它是自动精简配置并可调整大小的，而且将数据分散存储在多个 OSD 上。RBD 服务已经被封装成了基于 librados 的一个原生接口。
h) RGW:RADOS 网关接口,RGW 提供对象存储服务。它使用 librgw 和 librados，允许应用程序与 Ceph 对象存储建立连接。RGW 提供了与 Amazon S3 和 OpenStack Swift 兼容的 RESTful API。
i) CephFS——对应 GPFS 文件系统：Ceph 文件系统提供了一个使用 Ceph 存储集群存储用户数据的与 POSIX 兼容的文件系统。和 RBD、RGW 一样，CephFS 服务也基于 librados 封装了原生接口。

同样，如果把上述元素和概念按照逻辑进行拼接，可以得到以下这张 CEPH 的基本架构图，图中反映了各个组件的逻辑关系。



CEPH 提供了一个理论上无限扩展的集群，客户端和 ceph osd 进程通过 crush 算法来计算数据位置，而不必依赖一个中心查找表，我们知道凡是网络设备都有并发连接数据的限制，集中式/单体式的存储系统，对于大规模部署来说，很容易达到物理极限，在 CEPH 的数据访问机制中，客户端和 osd 进程直接通信，提高了性能和系统总容量，消除了单点故障，CEPH 客户端仅在需要时与 osd 进程建立一个会话。

osd 进程加入一个集群，并且报告他们的状态，分为 up 和 down 两种状态，代表是否可以响应 ceph 客户端的需求，如果 osd 进程失败，则无法通知 ceph monitor 它已经 down 掉，ceph 通过周期性的 ping OSD 进程，确保它正在运行，CEPH 授权 OSD 进程，确定授信的 OSD 进程是否已关闭，更新 cluster map，并报告给 CEPH Monitor.

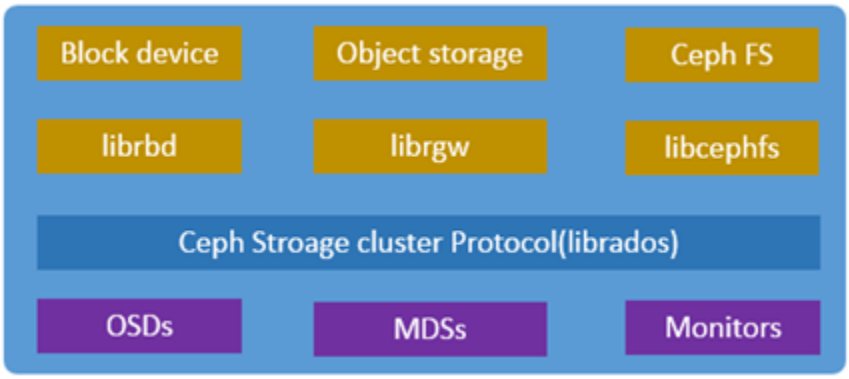
OSD 进程也通过 crush 算法，计算对象的副本应该存放的位置，在一个写场景中，客户端使用 crush 算法计算应该在哪里存放对象，并将对象映射到一个 pool 和 placement group，然后查询 crush map 来定位 placement group 中的主 OSD 进程。

客户端将对象写入主 osd 的 placement group 中，然后主 osd 使用它自己的 crush map 来找到第二、三个 OSD，并且将对象副本写入第二、第三 OSD 的 placement group 中，主 OSD 在确认对象存储成功后会给客户端一个回应。OSD 进程完成数据的复制，不需要 ceph 客户端参与，保证了数据的高可用性和数据安全。

CephFS 从数据中分离出元数据并保存在 MDS 中，而文件数据保存在 CEPH 存储集群的 objects 中，ceph-mds 作为一个进程单独运行，也可以分布在多个物理主机上，达到高可用和扩展性。

使用方案

了解了架构和原理，该怎么使用呢？Ceph 主要用于完全分布式操作，没有单点故障，可扩展到 exabyte 级别，完全免费使用。 其采用的位置感知算法和数据复制机制使其具有容错能力，并且不需要特定的硬件支持，也成为他天生骄傲的资本，大大降低了使用门槛，在贫瘠的物理介质上就可以野蛮生长。一般来说，CEPH 主要提供三种使用场景，rbd(block device),对象存储和 CephFS 文件系统方式，如下图所示：



CEPH 客户端使用原生协议与 CEPH 存储集群进行交互，CEPH 将这些功能打包成 librados 库，因此你可以创建自己的 CEPH 客户端，CEPH 作为分布式存储，对外提供各类型的标准存储服务。

CEPH block device 的快照功能对于虚拟化和云计算来讲很有吸引力，在虚拟机场景中，极具典型的是在 Qemu/KVM 使用 rbd 网络存储驱动部署 CEPH block device，宿主机使用 librbd 向客户机提供块设备服务。而在 K8S 管理的容器平台中，Ceph 也可以提供标准 rbd 设备的动态供给和共享存储空间。

5. CEPH 与 GPFS 的正面交锋

为了更深入透彻的了解 CEPH 和 GPFS 的优劣，我们将从以下这些方面逐一对比 CEPH 和 GPFS 的特性，期望可以提供更科学客观的参考。

管理功能

GPFS——GPFS 提供了一系列完美的商业化产品功能，基于策略的数据生命周期管理，高速扫描引擎，在线数据迁移，闪存加速，这些特性都大幅提升了它的用户体验，在复杂的 IT 环境中有了更多施展拳脚的空间。

CEPH——CEPH 产品相对年轻，周边功能和生态目前尚不完善，延展功能上来说不及 GPFS 丰富，但已经具备管理的基本功能，它的 VSM 功能，即 Ceph 的 web 管理界面，目前也已完善。

平台兼容性

GPFS: GPFS 一个很大的亮点是支持跨平台部署和文件共享，同一集群中可以包括 Windows/ Linux/AIX 等异构平台，良好的异构兼容性尤其对于传统企业复杂的异构 IT 环境有着天然的亲赖。

CEPH: 目前 CEPH 所提供的 rbd 是基于 Linux 内核的，CEPH 仅支持部署在 Linux 平台上，rbd 的块设备不能直接映射给非 linux 的客户端使用(如果要使用可以通过导出为 iscsi 设备的方式)。

服务方式

GPFS——是一个高性能并行集群文件系统，，可支持多种存储设备，包括 Flash、磁盘等块存储、对象存储、文件存储、甚至可以管理磁带。支持多云部署以及 POSIX、NFS/CIFS、HDFS/Hadoop 、Swift/S3 等多种接口。

CEPH:——可同时支持对象存储，块存储和文件型存储，且鉴于当前基于 POSIX 的文件系统方案尚不完善，CephFS 功能正努力完善中。支持 Switft/S3 等云存储环境。

存储性能

GPFS——广泛应用于世界领先的 HPC 超级计算环境。在加速并行访问方面的显著优势有：改善了小文件的 IO 性能，支持超过 4600 个计算节点的高速并发访问，实现 16GB/s 单节点顺序读写带宽，以及每秒可创建 260 万个小文件。作为一个并行文件系统，它将智能融入客户端，并由客户端在集群中的所有存储节点之间分配负载，即使对于单个文件也是如此。

CEPH——CEPH 的算 CRUSH 法和 PG 存放机制，使它可以充分利用多块磁盘的 IO 队列，但最开始基于 HDD 设计，对于 SSD 和 NVRAM 等使用场景没有没有特别的性能优

化策略，可能导致这些硬件的物理性能在 CEPH 中发挥受限，延迟和 IOPS 在高速硬件环境下得不到显著提升。

技术架构

GPFS——具有集群管理者的概念，节点间采用仲裁机制，在灾备环境下需要引入第三方站点，参与集群仲裁。

CEPH——没有绝对的中心结点，可以完全排除单点故障，无中心化的设计思想，使集群具有理论上无限扩张的可能性。

适用场景

GPFS——适用当下流行的生产环境，其中 FPO 架构可通过多个 block 组成 Chunk 的方式，很好的适应大数据环境，并且可以与 IBM Symphony 分析工作配合使用。同时 FPO 架构也可用于 IAAS 平台的底层存储，用于存储虚拟机镜像，用于 PAAS 容器云环境，用来对容器提供数据存储的接口服务。另外，也可以搭建集群环境提供 NAS 的功能用于文件和影像的共享。

CEPH——更多用来提供对象存储和块存储的服务，不适用于大数据环境，同样可用来 IAAS 和 PAAS 架构的云环境提供存储服务，或者为单一架构的 IT 环境提供块存储服务，作为分布式的优秀解决方案，天生有对接云生态的基因，CEPH 不仅在 OpenStack 时代可以大有作为，同样在容器云时代也可以大放异彩。

数据分层：

GPFS——GPFS 具有很好的数据分层实现机制，cache 机制，将日志卷部署在 SSD 上，在某些场景下可以带来显著的性能提供。

CEPH——Crushmap 可以用来做分级存储，例如根据底层不同硬盘，例如 HDD 或 SSD 等来分为不同的 pool，Ceph 的 Cache tier 技术可以实现 hot data 和 cold data 分离，把热数据放到 Cache 层，过段时间同步到 cold date 层等等。

安全机制

GPFS——该环境中，某一节点的硬盘连接丢失，不会影响到其他的节点，GPFS 使用 RSCT 的功能持续的监控不同文件模块的健康状态，当任一错误被检测到时，相应的恢复动作将自动执行。GPFS 还提供了额外的日志和恢复功能，可以维持元数据的一致性。最大三副本，可支持节点的自动 Failover。除此之外，GPFS 还支持原生文件加密，数据传输加密。授权安全管理，安全对接云存储，文件审计。

CEPH——rados 采用强一致性设计，可容忍网络中断、掉电、服务器宕机、硬盘故障等，并进行自动修复，保证数据的可靠性和系统可用性。也是同样的三副本设计，支持节点的自动 Failover。Monitors 是 Ceph 的管家，维护着 Ceph 的全局状态。Monitors 的功能和 zookeeper 类似，它们使用 Quorum 和 Paxos 算法去建立全局状态的共识。其 OSDs 可以进行自动修复，而且是并行修复。

冗余机制

GPFS——数据冗余可以通过 failure group 机制实现，以文件系统作为复制单元，数据在物理上存储两份或三份，节点冗余上，重要角色如集群管理者，会分配主备两个节点，其它角色会在集群节点间飘移。

CEPH——数据冗余上，底层文件对象默认存储 3 个副本，节点冗余上，多 mointor 机制可以有效防止单点故障，在文件存储上，额外的 ceph-mds 实例可以备用以取代任何失效的 ceph-mds,由 ceph-mon 自动完成，也可以启动多个 ceph-mds 实例，将目录树分离为子目录树，这样能够在多个启动的实例中有效的平衡负载。

6. 分布式存储未来畅想

未来的 IT 架构是生态之争，赢生态者得天下，就像开放的安卓赢得了众多开发者的亲赖，繁荣的产品生态也成就了安卓。运维自动化和智能化运维建设，要求底层 IT 环境实现高度整合，自主可控更是对开放性的要求，开放是一个产品的亲和力，意味着可以更灵活的融入当前 IT 环境，当前云计算的存储标准接口仍然有开放席位，静待新的有生力量入驻。

不管是存储，还是网络等基础架构，都在试图屏蔽底层物理硬件的差异，实现硬件的标准化管理，用软件定义一切，分布式存储就是在这样的趋势下，赢得了蓬勃发展的契机，开放的产品接口，丰富的插件，与当前环境的兼容耦合性，都将成为分布式存储领域制胜的关键，未来分布式存储在安全性、产品化建设、兼容性、可管理性、稳定性上的不懈努力，将是引领分布式存储占领数据中心存储江山的重要砝码。