中移苏研中心成果展示(四)

编者按

进入新世纪以来，以互联网为核心的新技术、新应用、新平台蓬勃兴起，深刻影响着全球电信业的发展模式和发展路径。习惯了过去高增长的电信运营企业，现在要适应低增长的“新常态”，并且这个过程可能还会较长。流量和固网宽带是电信运营企业未来收入保障，但是在目前的政策环境和竞争态势下，流量经营收入增长的“天花板”效应日渐显现。与此同时，互联网的业务创新不断激发市场需求，云计算、大数据、物联网、智慧城市等新业态、新模式风起云涌，开拓了越来越广大的发展前景。

目前，全球的电信运营商都在积极进行转型，越来越多地使用软件和云计算提供网络功能和服务，但运营商网络设施“烟囱式”的传统架构，长期依赖厂商提供硬件形式的网络构件，使得实施软件化道阻且艰。同时，以AT&T、中国移动为代表的运营商，正在从内部发起变革，成立软件子公司或者云计算研究院， 培养专业技术团队，以期能在行业冲击中寻找到新增长。

在此背景下，通信世界全媒体平台特邀中国移动（苏州）软件技术有限公司，采用连载方式，刊登中国移动在云计算、大数据、IT支撑、智慧城市等方面的转型发展新动向、创新技术以及实践案例，以便给广大同行业者带来启示。

SSD在分布式文件系统中的应用场景及方案对比

大云容器化操作系统自发布以来，已经在中国移动内部商用推广，目前部署规模已经近两百节点。

中移（苏州）软件技术有限公司 | 郭建楠

在云计算和大数据趋势下，企业数据存储和虚拟化应用需求海量增长。基于业界标准x86服务器，采用全分布式无共享（Share Nothing）架构的分布式存储系统，凭借其自身高可靠、高扩展、低成本的优势，得到了越来越广泛的应用。其中，Ceph（分布式文件系统）因其自身先进的架构，活跃的社区资源成为了Server SAN领域的翘楚。但低成本并不是Server SAN惟一的发展方向，市场上也同样亟需高性能的Server SAN产品。

另一方面，近几年来SSD（固态硬盘）发展迅速，技术不断地成熟，容量及密度越来越大，价格越来越低，但其“容量价格比”相较于HDD仍有一定差距，SSD全面取代HDD仍需时日。

基于以上现状，讨论SSD与Ceph的融合应用，使用高速但少量的SSD来满足关键性业务或IO敏感型业务对性能的需求，达到成本与性能的平衡，对生产实践有着重要的意义。本文总结了Ceph集群中

使用SSD的几种典型场景，作出了比较，并给出了推荐的使用方式。

作为Ceph-osd的日志盘使用

假设磁盘正在执行一个写操作，此时由于发生磁盘错误，或者系统宕机、断电等其他原因，导致只有部分数据写入成功。这种情况就会出现磁盘上的数据有一部分是旧数据，另一部分是新写入的数据，使得磁盘数据不一致。

Ceph引入事务与日志，来实现数据写盘操作的原子性，并解决数据不一致的问题。即所谓的“ceph数据双写”：先把数据全部封装成一个事务，将其整体作为一条日志，写入Ceph-osd journal，然后再把数据定时回刷写入对象文件，将其持久化到ceph-osd filestore中。

基于以上过程，可以将SSD作为ceph-osd journal的底层存储设备，来加速写入性能。同时，由于SSD设备IO性能较高，可以将SSD划分成多个分区，以配比多个HDD设备使用，如图1所示。

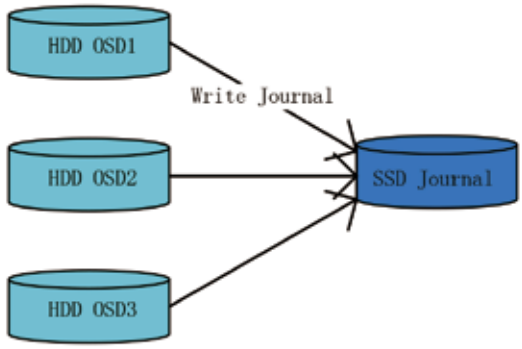

图1 SSD划分成多个分区使用

该方案的优点为使用高速的SSD设备加速ceph-osd journal的写入性能，尤其是对小块数据随机IO的场景，加速效果尤为明显。

但上述方案也存在缺点，即由于ceph-osd journal在实现逻辑上具有循环写入、定期回刷的特性，其对SSD设备容量空间的利用率很低。在典型场景下，SSD设备与HDD设备的配比为1:4，而每块HDD设备一般只使用10GB的SSD设备分区，造成了SSD设备容量空间的浪费。

作为Ceph数据多副本中的主副本存储介质使用

Ceph使用多副本机制来保证数据的安全性。

针对于写操作，在多份副本之间，ceph使用强一致性写策略，来保证数据的一致性。Ceph的PG（Placement Group）依据Crush伪随机算法，选择出副本数个数的ceph-osd存放数据，其中PG中的第一个osd为主osd，其他osd为从osd。Ceph先在主osd写入数据主副本，再由主osd将数据同时分发至其他多个从osd，进行数据从副本的写入。

针对于读操作，Ceph的读取请求只在主副本所在osd进行。

基于以上过程，可以将SSD作为数据多份副本中的主副本底层存储介质使用，来加速IO读写性能，如图2所示。

该方案的优点为利用SSD设备高性能优势，对于写操作，该方案缩短了数据第1份副本的写入执行时间；对于读操作，全部请求操作SSD设备，读请求能够快速处理并返回。

该方案的缺点为由于集群副本数的限制，SSD设备的容量空间必须与HDD设备容量空间有严格的配比关系。否则，由于木桶原理，设备中的一方先被写满，另外一方剩余的存储空间就无法再被ceph使用，从而出现存储容量空间的浪费。

构建小规模全SSD逻辑池单独使用

与以上方案思路类似，但采用更加彻底的方式，将SSD设备与HDD设备彻底分开，修改Ceph的数据存储规则，将高性能的SSD设备单独组成一个逻辑存储池，将正常性能的HDD设备单独组成一个逻辑存储池，对外提供两种不同性能规格的存储服务，如图3所示。

该方案的优点为充分利用SSD设备高性能的优势，将关键性业务或者IO敏感型业务全部放入高性能存储池，为客户提供性能优越的存储服务。

该方案的缺点为受到成本限制，高性能存储池存储容量规模会比较小，只能针对性地对少数关键性业务或者IO敏感型业务进行服务质量保障，且业务一经规划部署至某个存储池后，不能在两个存储池之间进行自动转换，后期进行调整开销较大。

SSD作为Cache层使用

可以将SSD作为Cache层，自动地对冷、热数据进行分层存储，来达到对业务智能服务的目的。

在数据的不同生命周期里，其访问的频度截然不同，即使是在同一生命周期的不同类型的数据，其访问的频率也会不同。因此，在“信息生命周期管理”的基础

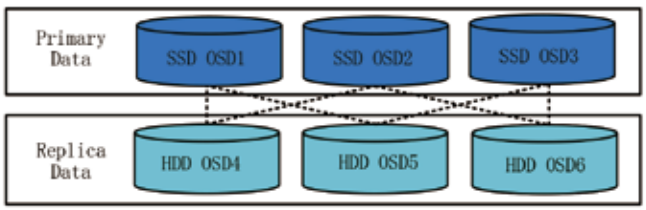

图2 加速IO读写性能

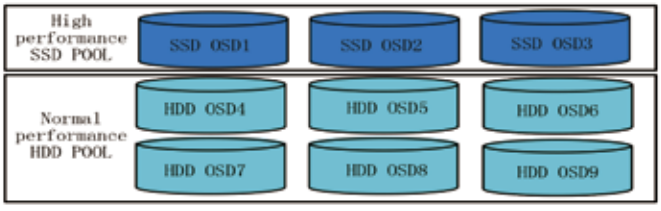

图3 两种不同方案对比

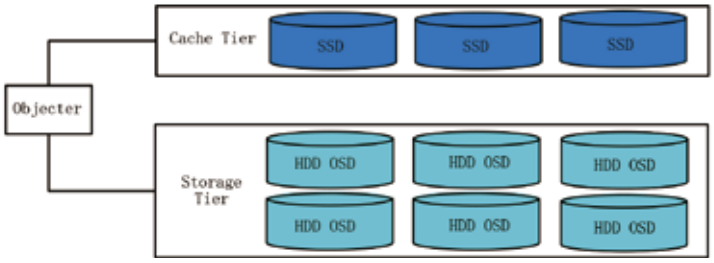

图4　Cache Tiering模块自动迁移

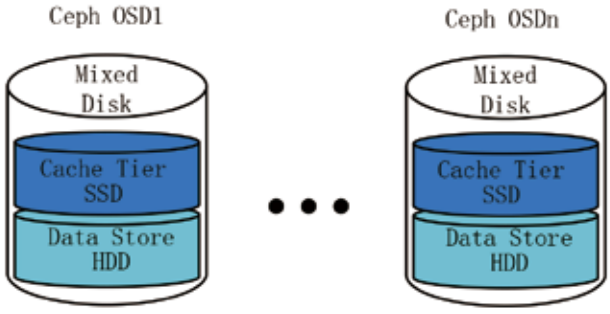

图5　集群层面数据自动分层存储

上对数据进一步进行分层存储十分必要。

自动分层存储技术的目标在于把用户访问频率高的数据放置在高性能、小容量的存储介质中，把大量低频访问的数据放置在大容量、性能相对较低的存储介质中。在提供热点数据存储性能的同时，降低存储成本：首先，冷数据可以自由安全地迁移到更低层的存储介质中，这样可以节约存储成本；其次，热点数据可以自动地从低层存储层迁移到高层存储层，提高访问热点数据的性能。

在Ceph中，有以下两种实现方案：

1)作为Ceph Cache Tiering技术中的Cache层

在Ceph里，Cache Tiering模块在逻辑存储池层面进行设置。可以将一个逻辑存储池设置为另一个逻辑存储池的cache层。据此，可以用SSD设备组建热数据存储池作为缓存层，用HDD设备组建冷数据存储池作为存储层，来达到冷、热数据分离的目的，Cache Tiering模块处理缓存层和存储层之间数据的自动迁移，对应用而言，迁移操作透明、无感知，如图4所示。

该方案的优点为可以充分地利用SSD设备高性能以及HDD设备大容量的优势，智能地对冷、热数据进行分层存放。

该方案的缺点为Ceph在cache tiering模块的逻辑实现尚在验证阶段，暂时不建议生产环境使用。

2)与HDD设备绑定作为混合盘的Cache层

在通用块层将SSD设备与HDD设备进行绑定，提供一个逻辑上的device-mapper层块设备，该逻辑设备的数据IO首先发生在SSD设备上，再定期回刷至HDD设备中。逻辑设备内部维护冷、热数据在缓存层与存储层之间的自动迁移，且对应用透明、无感知。使用这些逻辑上的混合盘作为ceph-osd的底层存储设备，构建存储集群，同样可以达到集群层面数据自动分层存储的效果，如图5所示。

该方案的优点同样为可以充分地利用SSD设备高性能以及HDD设备大容量的优势，智能地对冷、热数据进行分层存放，且不涉及ceph代码逻辑的修改，底层逻辑设备的组建及数据处理过程对Ceph而言完全透明。

该方案的缺点为架构稍显复杂，需要更多的管理、维护开销。

推荐使用方式

综上，当前阶段较为推荐的CEPH使用方式为结合方案1与方案4，即将SSD设备同时用作日志缓存与数据缓存使用。

以典型的SATA3.0的SSD设备为例，先将SSD设备进行多个分区，配比多块HDD设备，其中SSD的部分分区先与HDD设备进行混合存储逻辑块设备的构建，生成的混合盘再与SSD的剩余分区进行配对，共同构成ceph-osd的file store以及file journal，如图6所示。

这种使用方式，可以最大化地利用SSD设备的性能与容量，为Ceph集群的全部业务应用提供普适性、智能化的存储服务，达到性能与成本的平衡。

编辑 | 王熙 wangxi@bjxintong.com.cn

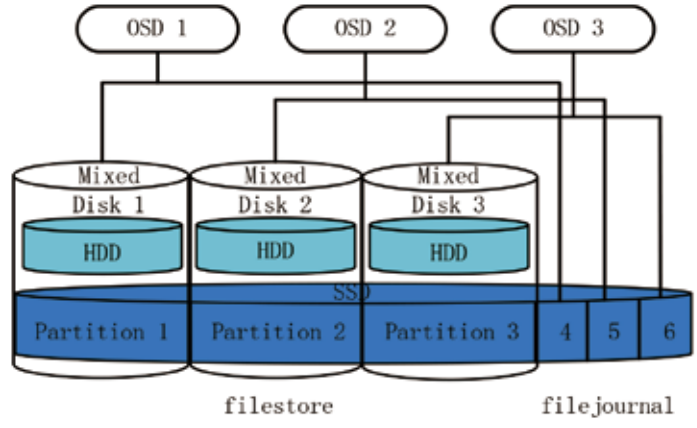

图6　混合存储逻辑块设备构建