ZStack 技术白皮书精选

分布式存储—硬盘容量不均衡导致的缓存盘寿命急速衰减分析

扫一扫二维码，获取更多技术干货吧

版权声明

摘要

大道至简·极速部署，ZStack 致力于产品化私有云和混合云。

ZStack 是新一代创新开源的云计算 IaaS 软件，由英特尔、微软、CloudStack 等世界上最早一批虚拟化工程师创建，拥有 KVM、Xen、Hyper-V 等成熟的技术背景。

ZStack 创新提出了云计算 4S 理念，即 Simple （简单）、Strong （健壮）、Smart （智能）、Scalable （弹性），通过全异步架构，无状态服务架构，无锁架构等核心技术，完美解决云计算执行效率低，系统不稳定，不能支撑高并发等问题，实现 HA 和轻量化管理。

ZStack 发起并维护着国内最大的自主开源 IaaS 社区——zstack.io，吸引了 6000 多名社区用户，对外公开的 API 超过 1000 个。基于这 1000 多个 API，用户可以自由组装出自己的私有云、混合云，甚至利用 ZStack 搭建公有云对外提供服务。

ZStack 拥有充足的知识产权储备，积极申报多项软著和专利，参与业内标准、白皮书的撰写，入选云计算行业方案目录，还通过了工信部云服务能力认证和信通院可信云认证。ZStack 面向企业用户提供基于 IaaS 的私有云和混合云，是业内唯一一家实现产品化，并领先业内首家推出同时打通数据面和控制面无缝混合云的云服务商。选择 ZStack，用户可以官网直接下载、1 台 PC 也可上云、30 分钟完成从裸机的安装部署。

目前已有 1000 多家企业用户选择了 ZStack 云平台。

分布式存储—硬盘容量不均衡导致缓存盘寿命急速衰减分析

Ceph 分布式存储在扩展性、可靠性、性能上具备独特的优势，可以实现快速扩展多台服务器，动态伸缩到 PB 级容量，多副本机制保障数据高可靠，数据均衡分布，并发性能高等场景。目前广泛应用于互联网、科研、教育、制造业、政府等诸多领域。ZStack 云平台目前支持对接 Ceph 分布式存储，使用的是分布式块存储，即使用 librbd 的块设备接口提供给 Qemu 访问，进行云主机、云盘的 IO 读写。

虽然 Ceph 分布式存储具备上述的优势特点，但在实践中，对硬件的选择及配置均存在特别要求，尤其是硬盘、网络上，如果配置不当，存储的可靠性和性能均会受到影响。

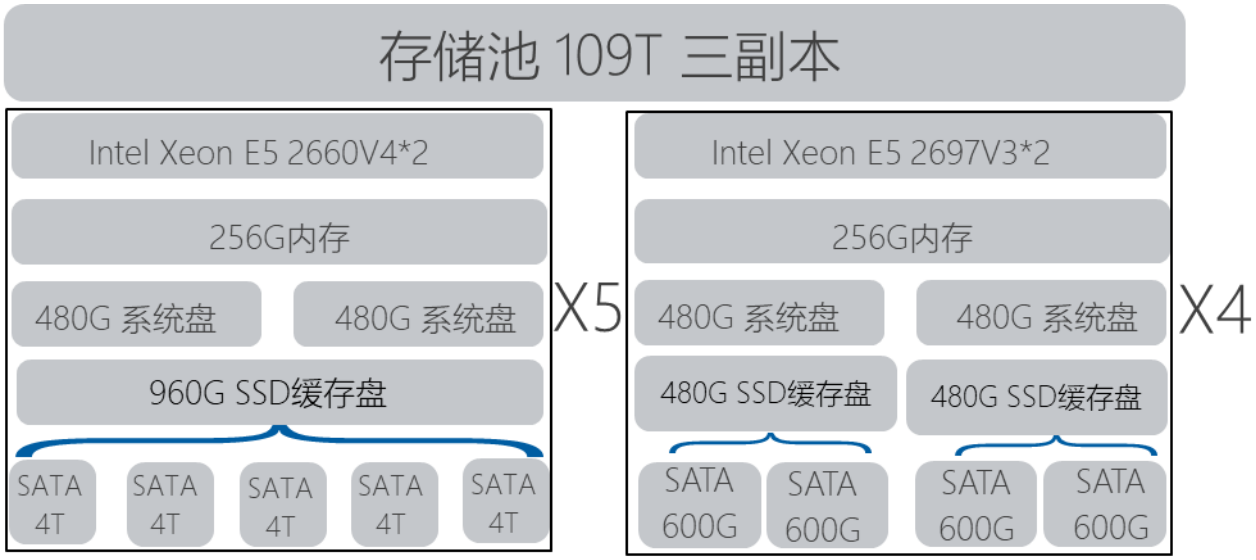
最近在日常巡检一套 ZStack 生产环境的 Ceph 分布式存储时，我们发现客户新购的五台服务器的 SSD 寿命损耗存在异常。具体的现象是使用半年后，服务器带外管理界面看到 SSD 的寿命损耗只剩下 89%，但使用 smartctl 读取介质损耗参数依然显示为 100%。

此时会很疑惑，到底哪个数据更可靠，如果 SSD 寿命只剩下 89%，那么如何去调整优化 Ceph 分布式存储？

问题回顾

针对这个问题，我们回顾一下这套分布式存储的架构。当时采用了新购+利旧的方案来部署分布式存储。

相应的配置信息如下：



其中，新购的 5 台机器采用了 Intel Xeon E5-2660 v4 的 CPU，内存为 256G，机器整体可插入 8 块 3.5 寸硬盘，采用了两块 480G SSD 硬盘配置 RAID1 安装系统，采用一块 960G SSD 做 Ceph 分布式存储的缓存盘，每个缓存盘对应了 5 个 OSD 数据盘，每个缓存分区约 160G 的容量，每个 OSD 容量 4T。存储采用万兆网络，做链路聚合 LACP Mode 4。

利旧的 4 台机器采用了 Intel Xeon E5-2697 V3 的 CPU，内存为 256G，机器整体可以插入可插入 8 块 2.5 寸硬盘，采用了两块 480G SSD 硬盘配置 RAID1 安装系统，采用两块 480G SSD 做 Ceph 分布式存储的缓存盘，每个缓存盘对应了 2 个 OSD 数据盘，每个缓存分区约 240G 容量，每个 OSD 容量 600G。存储采用万兆网络，做链路聚合 LACP Mode 4。

前五台机器，每台机器配置 5 块 4T 硬盘容量，总存储容量 100T，后 4 台，每台机器 4 块 600G 容量，总量 9.6T。

初期将所有容量规划到同一个存储池中，总裸容量约 109T，配置三副本后，容量约 36T。

环境主要运行了 MySQL，Redis，ELK，Zabbix，Web 服务，App 服务等业务，合计业务类型主要偏向 IOPS 密集型业务。业务运行前两个月，整体系统没有任何问题。

SSD 寿命参数分析诊断

针对 SSD 寿命损耗的不一致性，参考 SSD 的寿命参数，我们进行了以下分析：

Endurance Rating (Lifetime Writes)： 生命周期内总写入容量，客户环境使用的 960G SSD 生命周期内总写入量为 1.86 PBW，即最多可写入 1.86PB 的数据。

DWPD：Device Writes Per Day，硬盘每天写入次数，全盘写入，写满算一次，用于评估硬盘的耐久度。此款 960G SSD 的官网标称耐久度为 1 DWPD，即每天可全盘写入一次。

所以从 SSD 生命周期总写入量的角度来看，服务器带外管理界面看到的寿命损耗更为合理一些。

结合此硬盘的生命周期总写入量和每天可擦写一次，可了解此硬盘在 1.86PB/960G/每天=1860000B/960G=1937 天，约 5 年多的使用时间，与厂商承诺的 5 年质保的时间一致。

在使用 ZStack 云平台的 IO 监控工具及 smartctl 工具去排查分析 960G SSD 硬盘的每天写入量，发现每天硬盘的写入量在 2.5T 以上，接近 SSD 硬盘容量 960G 的三倍。

同时分析后 4 台服务器的 SSD 缓存盘的硬盘写入量很少，相应的硬盘总寿命未受过多影响。

测试发现，前五台服务器的 SSD，IOPS 95%都在 3000 以上，读写比在 15:85，平均读 IO 块大小为 16K 左右，写 IO 块大小为 18K 左右。而针对前五台服务器的 OSD 数据盘，IOPS 95%在 30 左右，读写比为 86:14，平均读 IO 块大小为 30K 左右，写 IO 块大小为 180K 左右。

所以前五台物理机的 SSD 缓存盘每天写入量接近官网标称值的三倍，按照生命周期总写入量的损耗预估，前五台服务器的 SSD 缓存盘寿命可能不到两年。

但后面 4 台服务器 SSD 的使用率为何没有提上去，对前五台服务器的 SSD 进行均衡使用呢。

我们再来了解一下 Ceph 数据分布的基本原理。Ceph 的 CRUSH MAP 算法，可以实现数据能够均匀地分布在不同容量硬盘的存储节点，Ceph 会根据 OSD 数据盘容量进行权重的计算，并基于存储集群的映射和数据分布策略的 placement rules 进行哈希计算。同一存

储池中，OSD 数据盘容量大的，IO 请求多，OSD 数据盘容量小的，IO 请求少。IO 请求经由数据的哈希到 PG 的映射过程，再由 PG 根据副本数映射到不同的 OSD 中。如果 OSD 硬盘不同，那么容量大的硬盘可以处理更多的 PG。相应的 IO 处理就更多。根据相应的 IO 均衡策略，如果存储池内总容量 109T，使用 30%的容量，则会在所有的数据盘均平均存储 30%的容量，相对于前五台节点采用的 4T 的数据盘，每个盘存储约 1.2T 的数据，后四台采用的 600G 的数据盘，每个盘存储约 180G 的数据。

所以基于这种硬盘容量的不均衡，导致相应的 IO 请求也会不均衡，在业务压力大时，后 4 台机器无法均衡处理整体的 IO 请求，在分布式规划时，需配置各机器的硬盘配置、网络配置一致。

分布式存储优化方案

针对以上情况，考虑进行以下调整：

检查当前业务使用情况，调整业务的使用方式，将部分非重要业务关闭，降低 IO 的使用方式，调整后，再监控相应 IO 的使用情况，发现 960G SSD 的每天写入量已降低至 1.8T，此时业务已无法持续调整。

在业务无法调整的情况下，只能考虑扩容及硬盘调整，在考虑扩容的过程中，还需要考虑后续业务量的增长。

因为当前的存储容量，可以提供当前业务的存储量，但在缓存盘性能上，不足以支撑相应业务的需求，此款 960G SSD 的每天硬盘写入次数 DWPD 为 1，只能全盘刷入一遍。考虑到硬盘的每天写入量，建议新缓存盘采用新款的 960GB SSD，官网的标称值其在生命周期的总写入量在为 5.26PBW，每天硬盘的写入量为 3DWPD，即每天可擦写三遍。

基于可靠与经济的基本原则，我们考虑以下硬件扩容方案进行扩容：

1. 再新增 3 台服务器，采用总写入量更高的 960GB SSD，480G SSD 系统盘，其他配置与原本前五台配置相同；

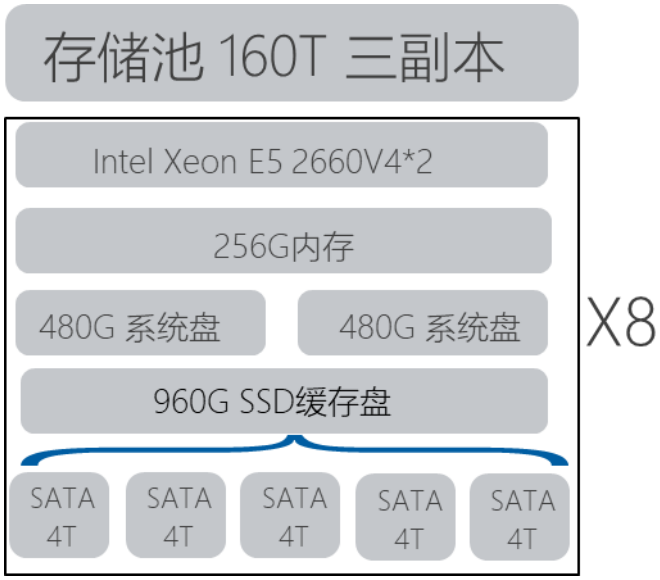
2. 前五台服务器，也采用总写入量更高的 960GB SSD 替换原本的 960GB SSD，将前五台机器扩容成 8 台相同配置的机器；

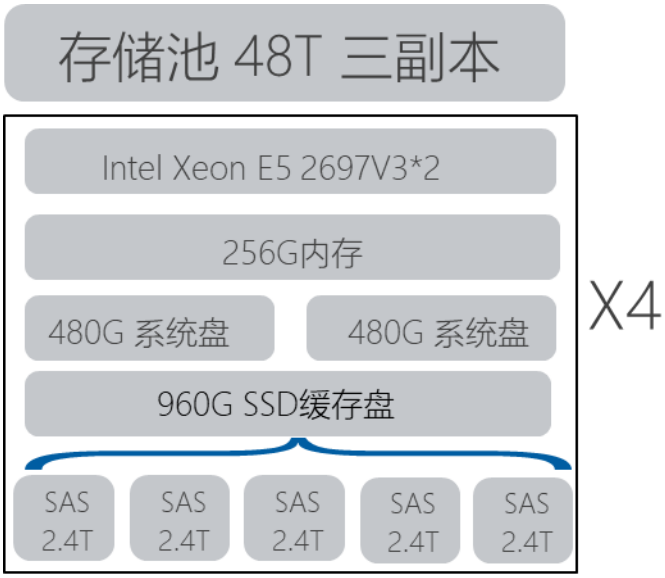
3. 后 4 台服务器，将缓存盘替换成步骤二移除下来的 960GB SSD，此时每台机器可以插入 5 块数据盘；

4. 后 4 台服务器，将原本的 2.5 寸 600G SAS 硬盘，变更为 2.4T 企业版 SAS 硬盘，目前 2.5 寸企业级硬盘最大容量受限于 2.4T；

5. 存储规划，8 台 E5-2660 的服务器提供 5x4Tx8 的存储容量，约 160T。后 4 台服务器提供 5X2.4Tx4 的存储容量，约 48T。

6. 前 8 台单独一个存储池，后 4 台单独一个存储池，均配置三副本。





具体的调整方案步骤，可参考以下步骤：

1. 从存储池，移除后 4 台服务器的硬盘，并关闭这 4 台机器；

2. 在新购入的三台服务器上，安装部署 Ceph 存储节点，加入到分布式存储集群的存储池中；

3. 将原本的前五台机器的一台服务器，移除硬盘，移除服务器，等待 Ceph 存储数据平衡恢复；

4. Ceph 平衡完毕后，关闭此服务器，将其上的 960G SSD 变更为耐久度更高的 960G SSD；

5. 重复步骤 3-4，完成前五台机器的变更；

6. 变更后 4 台服务器的硬件，将前五台机器中原本的 960G SSD 各分配一块到后 4 台服务器，将每台机器上的 600G SAS 硬盘更换成 5 块 2.4T 的 SATA 硬盘，添加到 Ceph 存储，针对这些 2.4T 硬盘，单独规划一个 Ceph 存储池；

7. 添加步骤 6 创建的新存储池到 ZStack 的 Ceph 主存储作为数据云盘池，创建数据云盘时使用，在业务使用时，可将部分业务，部署在后 4 台机器的存储池中；

8. 添加新购入的三台服务器到 ZStack 的计算节点集群中，同时用于提供计算资源。

使用上述方案变更，可以解决当前业务场景下，针对原本前 5 台服务器的每天硬盘写入量 3 遍，导致 SSD 寿命加速衰减的情况，又新增了三台服务器进行了计算存储的超融合扩容。针对 Ceph 容量存储 IO 请求不均衡的场景，也使用单独的存储池，进行规划，相同容量的硬盘规划到同一存储池，可以实现 IO 请求的均衡，IO 数据的均衡，各 SSD 的使用也相对均衡，即 8 台服务器的使用损耗一致，后 4 台服务器的使用损耗也一致。

结语

综上所述，分布式存储在规划部署时，需要考虑以下方面：

1.同一存储池的硬盘型号容量应一致，否则不同容量的硬盘在同一存储池，会导致 IO 请求的不均衡，导致存储分布不均衡，在使用 SSD 缓存盘的场景会导致使用大容量硬盘对应的 SSD IO 请求更多，损耗会更快；

2.业务规划需提前做好评估，针对 IOPS，带宽写入进行提前规划，高 IO 的业务需进行评估，准备的硬件是否可满足业务需求，如果业务需求较高，需使用更高配置硬件或进行相应的硬件扩容；

3.分布式存储选择 SSD 时，建议关注 SSD 的 PBW(生命周期总写入量)和 DWPD(每天硬盘可写入量)，SSD 寿命的损耗与其总写入量需要规划考虑业务类型，IO 密集型业务应选择更高 DWPD 的 SSD。