CEPH的性能调优实践分享

孙杰

TOPIC

- 硬件选型
- 性能调优
- 小结

硬件选型

- **把握一个原则**：Ceph的硬件选型需要根据存储需求和企业的使用场景来制定。
- **企业需要什么渴望什么：TCO低、高性能、高可靠**
- **一般企业使用Ceph的历程**：硬件选型— 部署调优— 性能测试—架构灾备设计—部分业务上线测试—运行维护（故障处理、预案演练等）。

企业里的典型场景

- **高性能场景**：这种场景类型的亮点在于它在低TCO下每秒拥有最高的IOPS。典型的做法是使用包含了更快的SSD硬盘、PCIe SSD、NVMe做数据存储的高性能节点。通常用于块存储，也可以用在高IOPS的工作负载上。
- **通用场景**：这种场景类型的亮点在于高吞吐量和每吞吐量的低功耗。通用的做法是拥有一个高带宽、物理隔离的双重网络，使用SSD和PCIe SSD做OSD日志盘。这种方法常用于块存储，如果你的应用场景需要高性能的对象存储和文件存储，也可以考虑使用。
- **大容量场景**：这种场景类型的亮点在于数据中心每TB存储的低成本，以及机架单元物理空间的低成本。也被称为经济存储、廉价存储、存档/长期存储。通用的做法是使用插满机械硬盘的密集服务器，一般是12～20台服务器，每台服务器4～6TB的物理硬盘空间。通常用于低功耗、大存储容量的对象存储和文件存储。

硬件选型依赖的几个因素

- 1、CPU

Ceph OSD运行RADOS服务，需要通过CRUSH来计算数据的存放位置，复制数据，以及维护Cluster Map的拷贝。通常建议每个OSD进程至少有一个CPU核。 Metadata和Monitors?

- 2、RAM内存

OSD在日常操作时不需要过多的内存（如每进程500MB）；但是，在执行恢复操作时，就需要大量的内存（如每进程每TB数据需要约1GB内存）。通常来说，内存越多越好。

- 3、数据存储
- 规划数据存储时要考虑成本和性能的权衡。进行系统操作，同时多个后台程序对单个驱动器进行读写操作会显著降低性能，也有文件系统的限制考虑。

硬件选型依赖的几个因素

- 4、网络

网卡能处理所有OSD硬盘总吞吐量，所以推荐最少安装两个千兆网卡，但是一般生产环境会建议部署万兆网卡。

- 5、硬盘

Ceph集群的性能很大程度上取决于存储介质的有效选择。应该在选择存储介质之前了解集群的工作负载和性能需求。

- 6、CEPH OSD日志盘
- 如果工作环境是通用场景的需求，那么建议使用SSD做日志盘。使用SSD，可以减少访问时间，降低写延迟，大幅提升吞吐量。使用SSD做日志盘，可以对每个物理SSD创建多个逻辑分区，每个SSD逻辑分区（日志）映射到一个OSD数据盘。通常10～20GB日志大小足以满足大多数场景。

调优前的分析

- 根据业务类型迚行分析
- 分析业务是偏向于容量还是性能，分析对数据安全性，可靠性的需求。
- 分析业务IO类型，是吞吐要求较高还是IOPS要求较高。
- 分析业务是多点访问存储还是单点访问，是否对单点突发性能有较高要求。
- 分析业务是否需要扩展，规划crushmap，减少在扩展时的数据迁移。

性能调优之一：硬件层面

- 硬件规划
- SSD选择
- BIOS设置
- NUMA设置

性能调优之一：系统层面

- Linux Kernel。
- 内存。
- Cgroup。

性能调优之一：网络层面

- 巨型帧。
- 中断亲和。
- 硬件加速

性能调优之一：ceph层面

- Ceph参数。
- PG Number调整。
- Ceph参数配置示例。

CEPH优化之阀值

- filestore_queue_max_ops = 65536
- filestore_queue_max_bytes = 536870912
- filestore_queue_committing_max_ops = 65536
- filestore_queue_committing_max_bytes = 536870912
- journal_queue_max_ops = 65536
- journal_queue_max_bytes = 536870912
- osd_client_message_cap = 65536
- osd_client_message_size_cap = 536870912
- ms_dispatch_throttle_bytes = 536870912
- 调整之后会占用更多内存，配合其他优化大概可以提升10%左右性能

Ceph优化之杂项

- osd_enable_op_tracker = false #默认开启，可以跟踪op执行时间
- throttler_perf_counter = false #默认开启，可以观察阈值是否是瓶颈
- 当在特定环境调整到最佳性能后，建议关闭，tracker对性能影响较大。
- cephx_sign_messages = false #默认开启，如果对安全要求不高建议关闭
- filestore_fd_cache_size = 4096 #默认256
- filestore_fd_cache_shards = 256 #默认16，修改后，略有提升

CEPH测试

- ·测试对象：要区分硬盘、SSD、RAID、SAN和云硬盘等，因为它们有不同的特点。
- ·测试指标：IOPS和MBPS（吞吐率），下面会具体阐述。
- ·测试工具：Linux下常用Fio、dd工具，rados bench，Windows
- 测试参数：IO大小、寻址空间、队列深度、读写模式和随机/顺序模式。
- ·测试方法：也就是测试步骤。

Ceph运维

- Ceph运维要做到能文能武!
- 文能提笔写Ceph运维手册、预案手册等；武能挥手部署Ceph、进行预案演练、故障处理、集群扩容等。来保证Ceph整个集群的高可用性，确保数据不丢失，同时也进行一些常规故障的预案演练，保证出现故障后能够有序的进行故障恢复。
- 三分技术七分运维!