基于 Ceph 的企业分布式存储方案

文/刘军军

摘要

本文首先通过对比主要的开源分布式存储系统的优缺点和适用场合，确定 Ceph 是目前适合企业需要的分布式存储系统。之后，提出一套基于 Ceph 的企业分布式存储解决方案，给出软硬件组件选型、构架设计，并对系统细化调优。在这套方案基础上，企业用户可以很容易构建出一套高效运行的分布式存储系统。

【关键词】分布式存储 Ceph Openstack 构架设计 性能优化

1 分布式存储Ceph

1.1 简介

Ceph 是一个开源的、理论上可无限扩展的、具有高可靠性、高性能的分布式存储解决方案。目前很多商业分布式存储解决方案是在开源的 Ceph 基础上发展来的，如 Bigtera 的 VirtualStor 系列产品，Hope Bay 的 ArkFlex 云端大规模数据存储平台产品，SanDisk 的 InfiniFlash 的产品 IF500 等产品都使用了 Ceph。

1.2 Ceph与其他分布式存储方案的对比

分布式存储除了 Ceph 还有 Moosefs（MFS）、Glusterfs、HDFS、Lustre 等很多种。本文综合分析熊文等的论文和其他一些讨论，发现 Moosefs 的优点是实施简单，缺点是存在单点故障和性能瓶颈；Ceph 的优点是扩展性好，可以很好的与 OpenStack 集成，发展很快，缺点是部分功能还不够成熟，通过 POSIX 接口访问 CephFS 时候，底层不稳定性使得不适合应用于生产环境；Glusterfs 的优点是扩展性好，缺点是没有 MDS，因此增加了客户端的负载，占用相当的 CPU 和内存，同时遍历文件目录时，实现较为复杂和低效，需要搜索所有的存储节点；HDFS 的优点是适合部署在大量通用、廉价硬件上，缺点是只适合特定应用场景，即一次写入，多次读出，做数据分析类应用；Lustre 的优点是成熟，缺点是复杂，同时 MDS 无法扩展，存在性能瓶颈。从适用场合方面分析：Moosefs 适合企业小型应用环境，存储小文件；Ceph 适合一般企业使用，如私有云平台应用，存储小文件；Glusterfs 适合一般企业中型应用，存储大文件，下文件读写效率很低；HDFS 适合存储超大数据集，做数据分析类应用；Lustre 是一个并行文件系统，做高性能计算（HPC）类应用，存储大文件，适

<< 上接 97 页

3.3 指标权重计算分析设计

在完成所有判断矩阵的创建后，需通过幂法针对判断矩阵计算所有指标权重，计算出权重后进行一次性检验，经检验满足一致性条件，故判断矩阵合理。

CR=CI/RI= (λ-N)*RI/(N-1)

=（7.259-7）*1.32/6=0.057<0.1 （4）

权重可通过软件查看 , 详见图 4。

3.4 效能评估计算

在进行效能评估计算时，还需对所有根节点指标（不含有子项）的指标能力进行评估，选中指标进行指标能力进行打分。详见图 5。

在完成所有指标能力输入之后即可针对通讯系统进行效能评估，计算系统效能需自下逐层向上进行计算，计算每一个树形节点效能时，需计算其下层所有指标的能力的权重加权和，遍历至通讯系统节点结束。

4 层次分析法模型验证

通讯系统在不同场景下飞行试验获取的结果如表 3 所示。

在场景 1 下，通信系统发挥了它 99% 的效能，场景 2 下，通信系统仅发挥了 65% 的效能，该评估结果与实际情况是相符的。

场景 1 的试验地点地形平坦，有利于电磁波的传播，且当时的飞行季节为冬季，植被稀少，为经济欠发达地区，电磁环境简单。飞行员的主观评述也认为在该地区飞行时，该通信指挥机的通信系统使用效果良好。

场景 2 的试验地点为丘陵地形，多径效应复杂，影响电磁波的传播；当时的飞行季节为夏季，植被繁茂，对超短波通信造成不利影响；为经济发达地区，来自工业及民用通信的电磁波干扰较大；以上众多因素导致通信系统效能发挥率不高。同时，由飞行员给出的主观评述也认为在南京地区飞行时，该通信指挥机的使用效果一般。

5 结束语

本文介绍了如何利用层次分析法进行武器装备效能评估，并设计了效能评估软件，以通讯系统为例对效能评估的软件模型进行了验证，本文的软件模型正确，能够通过该软件模型对武器装备效能进行评估。

本文用于验证的模型建立的指标体系较简单，实际指标类型更为复杂，但本文中的效能评估软件的基本思路和方法已初步形成，对基于层次分析法的效能评估具有推广借鉴意义。

参考文献

[1] 王强，周怀军，吴成富．基于 AHP 算法的相控阵雷达系统效能评估 [J]. 舰船电子对抗，2003,32(3):81-82.

[2] 许黎黎． 基于多层次综合评价的指挥控制系统评估技术研究 [D]. 沈阳：沈阳理工大学，2007:15-16.

[3] 孟超．机载数据链系统效能评估研究 [J]. 硅谷 ,2014,163(19):55-57.

作者简介

李太平（1982-），男，湖北省宜昌市人。硕士学位。现供职于中国飞行试验研究院，主要从事综合航电系统试飞技术研究工作。

作者单位

中国飞行试验研究院 陕西省西安市 710089

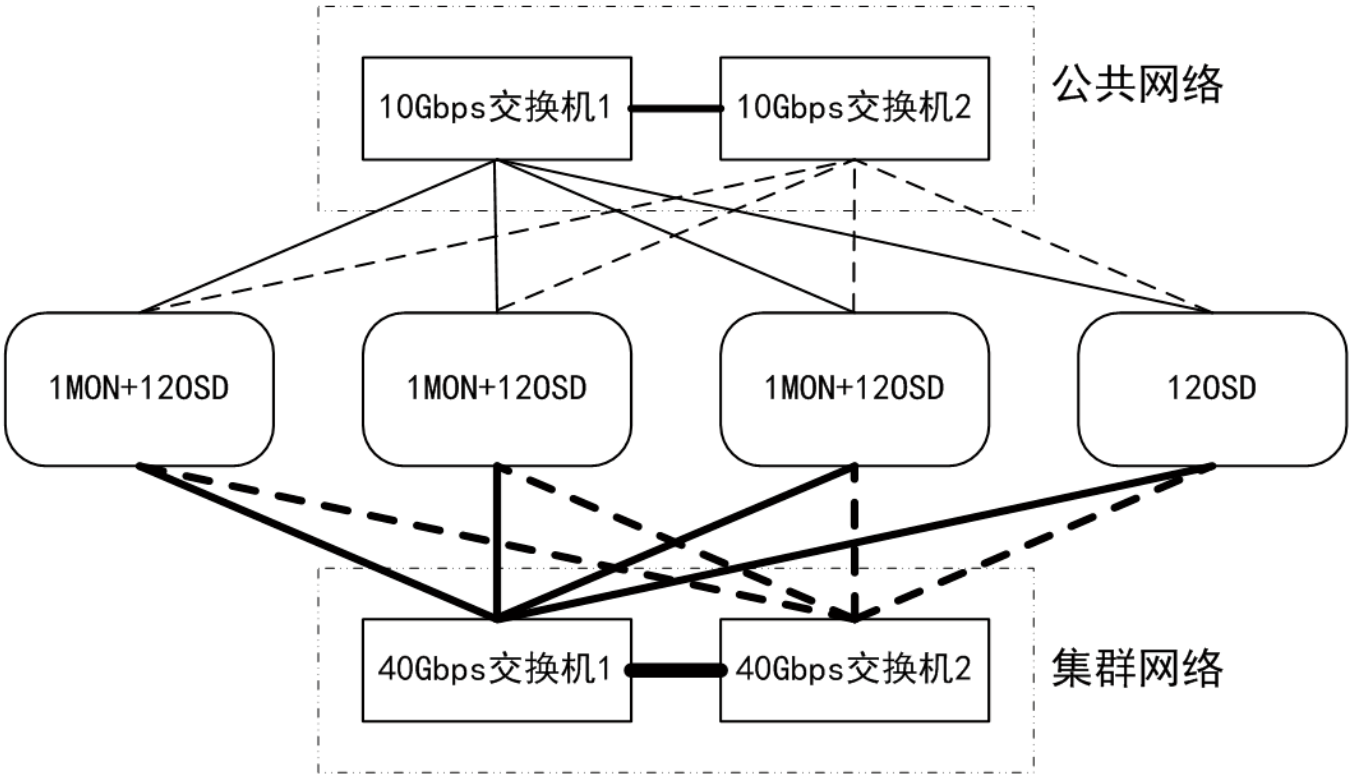

图 1：网络构架图

合大型科研、企业应用，一般 HPC 具有计算密集型、海量数据处理等特点，在石油勘探、地震预测、气象预报、航天航空、科学研究、机械制造、动漫渲染等领域都有应用。

2 基于Ceph的企业分布式存储方案

本文给出一种基于 Ceph 的企业分布式存储解决方案以供参考。方案描述总体设计架构、软硬件设计、性能调优这三个方面。

2.1 软硬件设计

设计目标：创建一个包含 4 个存储节点的分布式存储集群。

2.1.1 软件配置

操作系统建议选择最新的，本文选择 CentOS7.1 操作系统，将内核升级为 4.5 版本。由于目前 btrfs 文件系统还不稳定，因此底层文件系统选择 XFS。

Ceph 发行版本根据目前 Ceph 发行版本情况，选择最新稳定版 9.2.1。

存储的访问方式选择块设备方式访问，由于目前 CephFS 文件系统还不稳定，不建议用于生产环境。

2.1.2 硬件配置需要考虑的因素

（1） CPU。需要考虑数据存储节点，即 OSD 节点和监视器节点，即 MON 节点的 CPU 消耗量。Ceph 的 OSD 进程利用 CRUSH 算法计算数据的存放地址，复制数据，维护自身的集群映射，如果使用纠删码的数据池比使用直接复制数据的数据池 CPU 消耗要多。因此，OSD 要根据数据存储的策略预备足够量的 CPU 资源。监视器只是简单维护集群映射的主拷贝，所以他们一般消耗不了多少 CPU 资源。

（2）内存。对于 OSD，常规操作每个进程需要 500MB 内存，恢复数据适合每个 OSD 进程需要至少 1GB 内存每 1TB 数据。对于 MON 进程，一般每进程需要至少 1GB 内存，以实现对集群映射的快速维护，当然多配置些内存会达到更好的效果。

（3） 磁盘和网络。对于一个小规模的集群，单台存储服务器不要配置过多的磁盘，这种情况的危险在于单台服务器出现故障需要停机时，将造成集群的存储重心转移，出现数据丢失或大量数据恢复的情况发生。存储节点的数据磁盘无需做任何 RAID，直接配置成 JBOD 模式或直通模式，每块数据盘在系统中看起来是一个磁盘设备，在其上运行一个 OSD 进程。考虑磁盘空间和价格因素，选择单块 4TB 大小的磁盘作为数据盘比选择 2TB 大小的磁盘更经济。网络需要确定对外提供服务的公共网络和存储集群网络。选定公共网络网口带宽为 10Gbps，由于 Ceph 是通过存储多份拷贝保证数据安全的，客户的一个写入请求可能触发多个同样的写操作，这些通信是在集群网络内完成的，集群网络带宽要数倍于公共网络，因此，集群网络网口选择 40Gbps。为了防止系统设计出现瓶颈，公共网络带宽确定之后，根据公共网络带宽和存储节点磁盘吞吐量的平均值，容易计算出存储节点上 OSD 的数量为：存储节点上 OSD 的数量 = 公共网络带宽 / 存储节点磁盘吞吐量，根据当前系统的选择，存储节点上 OSD 的数量 =1250MB/110MB=11.4，于是确定存储节点上普通企业级磁盘数量应该配置在这个数值附近。如果在系统中使用 SSD 盘用于存放 OSD 的日志（Journal）信息，我们选择 Intel DC S3500 系列 480GB 的 SSD 盘，其写速度可达到 410MB/s，在其上分区存放几个 OSD 的 Journal 信息，410/110=3.7，于是可确定，一块 SSD 盘可供 4 个 OSD 同时在其上存放 Journal，为每个 OSD 划分 120GB 空间存放 Journal，最终确定系统在公共网络带宽为 10Gbps 前提下，一个存储节点 OSD 数量设置 12 个比较合适，需要 12 个 4TB 的数据盘和 3

个 Intel DC S3500 系列 480GB 的 SSD 盘。

2.1.3 具体配置

（1）存储节点配置

CPU ：2 路 Intel E5-2630V3，8 核 2.4GHz

内存：8 条 16GB DDR4 2133

系统硬盘：2 个 300G SAS 10K 2.5 寸

Journal 盘：3 个 Intel DC S3500 480G SSD

数据盘：12 个 4TB SATA 3.0 6Gbps 7500 RPM

RAID 卡：1 个 LSI MegaRAID SAS 9240-8i，配置成 JBOD 模式

网卡：1 个双口 10Gbps 光纤网卡（含 2 个 SFP+ 模块），1 个双口 40Gbps 光纤网卡（含 2 个 QSFP+ 模块）。

（2） 监视器（MON）节点：

在存储节点上，在安装操作系统的磁盘上启动 MON 监视器进程，监视器的网络就使用集群网络。由于整个集群需要监视器进行仲裁，所以，整个集群需要有大于等于 3 的奇数个监视器进程在运行。本方案中采取在 3 个存储节点上启动 MON 监视器服务进程。

（3）网络交换设备

交换机：2 个华为 CE7850-32Q-EI 用于集群网络，2 个华为 CE6850-48S6Q-HI 用于公共网络。

2.2 构架设计

最终确定如图 1 所示。

构架设计中公共网络、集群网络都采用 2 个交换机组成冗余网络的方式进行构建：网络中 2 个交换机使用华为 iStack[8] 技术进行堆叠，使得两个交换机在逻辑上看起来像是一个整体，在交换机一侧，2 个连接同一个存储节点服务器的端口进行链路汇聚，在存储节点服务器一侧，2 个连接公共网络或集群网络的网口在操作系统中分别做捆绑操作。客户机通过公共网络连接存储进行访问。上述构架设计，系统的裸存储容量为 4TB×12×4=192TB。由于 Ceph 具有横向扩展性，我们只需要添加新的存储节点和网络交换机即可完成存储容量的增加。假设，网络交换机堆叠构建的是无阻塞网络，那么对于集群网络交换机，最多有一半的接口用于与存储节点连接，即 16 个接口，另一半接口用于交换机之间的 iStack 连接，也就是在不添加网络交换的情况下，本系统裸存储容量做大增长到 4TB×12×16=768TB。

2.3 性能优化

2.3.1 软硬件参数调整

（1）硬件参数调整。在服务器的硬件配置中开启 CPU 的 HT（Hyper-Threading）特性，关闭 CPU 的节能模式，关闭内存的 NUMA（Non Uniform Memory Access Architecture）特性。在运行 OSD 进程的操作系统中设置可运行进程数为理论最大值 pid_max=4194303；在 Ceph 存储节点的操作系统中设置 vm.swappiness=0 禁止使用交换分区，完全使用内存；设置 SSD 盘的调度算法为 noop，设置 OSD 的磁盘调度算法为 deadline，所有分区均采用 GPT 分区格式且保证分区是 4K 对齐的；在使用 Ceph 块设备的客户端上设置对应的块设备参数 read_ahead_kb=8192；设置 Ceph 存储节点上所有公共网络和集群网络的接口使用巨型帧，MTU=9000，并在相应的交换机对应接口上也做相应设置；利用 ulimit -n 131072 设置系统打开文件数最大值为 131072。

（2）软件参数调整。Ceph 软件的参数都记录在 ceph.conf 配置文件中，每次 Ceph 的服务进程，比如 MON、OSD 启动，都会首先读取这个配置文件。这个文件分 global、MON、OSD、MDS、client 五部分，global 为全局公共部分，参数定义全局有意义，MON、OSD、MDS 这三部分对应三种服务进程的配置，可以覆盖在 global 部分定义的相同参数，client 部分定义的参数对所有连接 Ceph 存储系统的客户端都生效。根据 Ceph 构架和参数的含义，通常设置从日志到数据盘最大同步间隔 filestore max sync interval=15，根据系统设计原理设置：

osd journal size = {2×(expected throughput× filestore max sync interval)}

对于 OSD，其吞吐量为企业级硬盘的吞吐量，按照上面计算每个存储服务器应该设置多少 OSD 时使用的值为 110MB/s，于是有：osd journal size=2×110×15=3300MB，考虑到 OSD Journal 设置太小，会导致 Journal 文件频繁清空重写，也会造成性能损失，需要设置得大一些，一般设置 osd journal size 为 20480，即 Journal 文件大小为 20GB。

归置组 PG 数量要根据 OSD 的数量进行计算，最后算出的结果取一个最接近的 2 的指数的值：

PG_num = (Total_number_of_OSD×100) / max_replication_count

对于上述 4 个存储节点的方案，系统 Total_number_of_OSD = 4×12 = 48 个，假设系统每份数据存储 1 份副本，即 max_replication_count = 2，那么 PG_num = 48×100/2 = 2400 ~= 2048。

3 总结

目前，Ceph 存在的主要问题仍处于发展阶段。Btrfs 虽然具有写时复制等很好的特性，但是还不成熟，无法在生产系统中应用，一旦它成熟起来可以应用于生产环境，将极大提升 Ceph 的性能。

参考文献

[1]Sage A.Weil.Ceph:Reliable, Scalable,and High-Performance Distributed Storage[D].Ph.D. thesis,University of California, Santa Cruz,2007.

[2] 蔡思萌 . 开源分布式存储当道 Ceph 系统抢占企业应用 [EB/OL].http://toutiao.com/a6218346743769497858,2015.

[3] 熊文等 . 几个常见分布式文件系统特征分析和性能对比 [J]. 集成技术 ,2012, 1(4):58-63.

[4] 朱荣泽 .4 种分布式文件系统比较 [EB/OL].http://blog.csdn.net/metaxen/article/details/7108958,2011.

[5] 陶然 . 基于 Lustre 的 HPC 产品 [EB/OL].http://server.it168.com/a2014/0722/1648/000001648542.shtml, 2015.

[6] 刘明 .Linux 文件系统 btrfs 简介 [EB/OL].http://www.ibm.com/developerworks/cn/linux/l-cn-btrfs/, 2010.

[7] 维基百科 .JBOD 的介绍 [EB/OL]. https://en.wikipedia.org/wiki/Non-RAID_drive_architectures,2015.

[8] 华为技术有限公司 .iStack 技术白皮书 [Z]. 深圳：华为技术有限公司 ,2013.

作者简介

刘军军（1982-），男，内蒙古自治区乌海人。硕士学位。现为神华准格尔能源有限责任公司网络工程师。研究方向为小型机技术、云计算技术。

作者单位

神华准格尔能源有限责任公司　内蒙古鄂尔多斯市准格尔旗薛家湾镇　010300