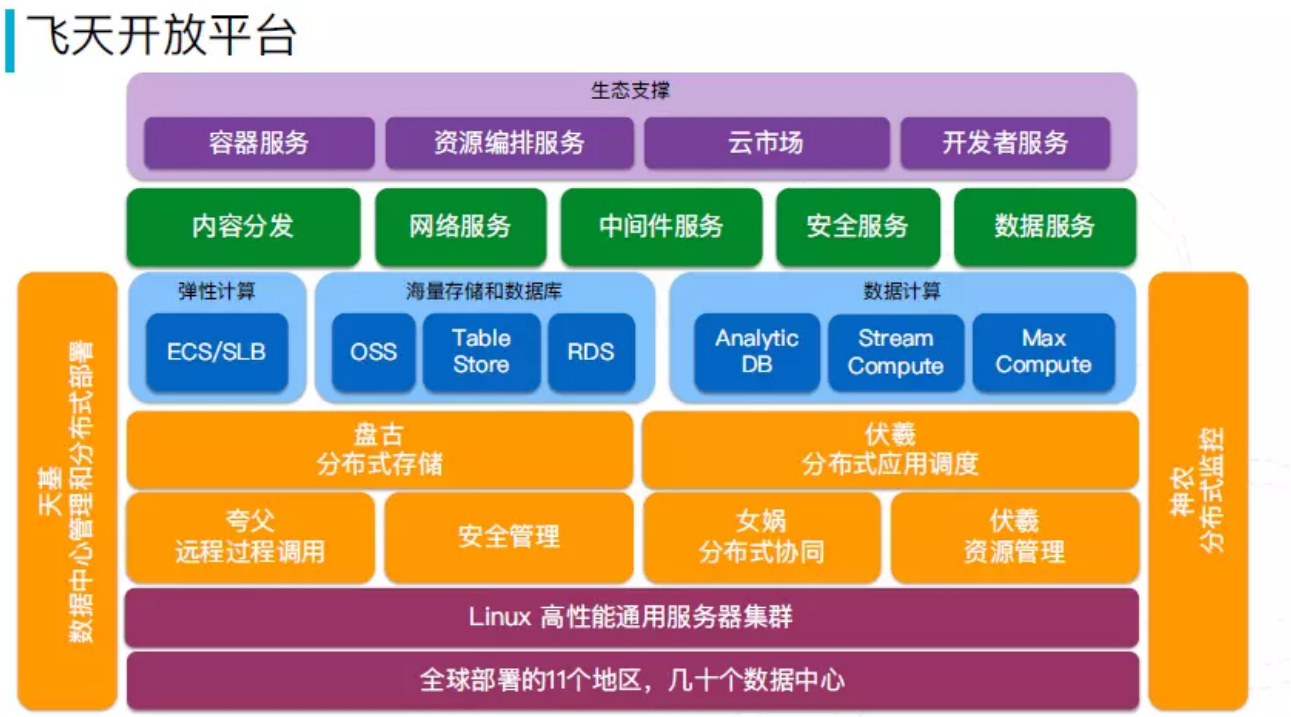
阿里云的分布式存储技术——盘古 和 GFS 的对比

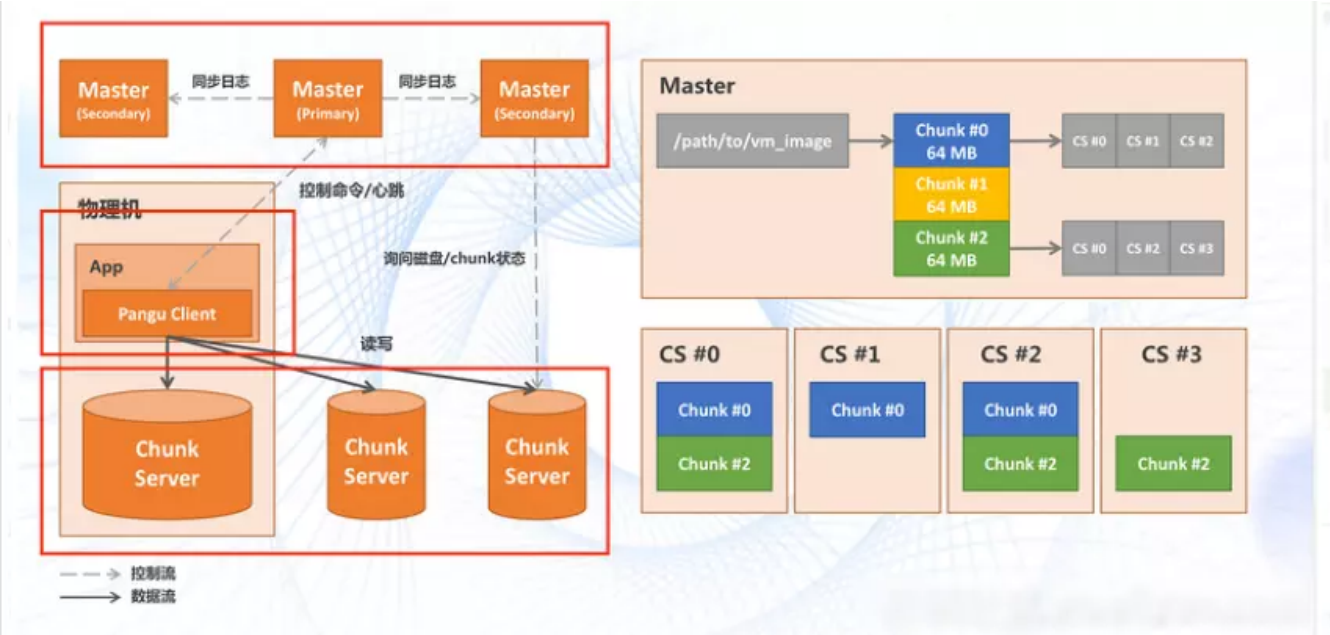
张俊华 16030199025

阿里云的云存储产品都拥有一个公共的底层存储平台，叫做盘古。阿里云在2009年成立之初就开始了盘古分布式文件系统的研发，经过多年的发展，盘古文件系统已经能够支持块存储云盘、对象存储、文件存储、大规模数据处理、数据库等各种复杂场景，并且针对离线和在线等不同场景做了精细的优化。



盘古架构

「盘古」分布式存储架构共有三个模块：Master、Client、Chunk Server



盘古Master

对应GFS的NameNode，负责元数据管理，最主要的就是维护两个映射关系：

- 文件名到数据块；
- 数据块到Chunk Server列表

其中文件名到数据块的信息保存在磁盘上(持久化)；但Master不保存数据块到Chunk Server列表，这个是通过Chunk Server在启动时的上报数据块信息，更新Master上的映射表。 Master暴露了文件系统的名字空间，用户可以以文件的形式在上面存储数据。

盘古采用基于Paxos协议的盘古Master来管理元数据， 通常配置为5个实例，可以同时容忍两台机器出故障。采用Paxos一致性协议，保证了高可用和快速切换的能力，减少了外部的依赖，做到了独立自包含，在保障高稳定性和高性能前提下能够容忍复杂故障。 支持按照Namespace来分区， 支持EB级别容量和万亿级别文件数的线性扩展能力。

盘古 ChunkServer

对应 GFS 的DataNode，负责数据存储，一般是多台组成集群。 存储过程中，一个文件被分成一个或多个数据块(至少一个)，这些块存储在多个Chunk Server上，每块数据通过多副本来保证可靠性以及加快后期的读取速度。Chunk Server负责处理分布式文件系统客户端的实际的读写数据请求。在Master的统一调度下进行数据块的创建、删除和复制。

ChunkServer负责管理存储空间和数据读写。首先，ChunkServer支持分级存储，针对不同的存储介质如NVMe SSD, SATA SDD, HDD等，根据相关配置的策略，把数据写入对应的存储介质，同时支持基于策略的迁移。比如说在混合存储云盘，数据先写入来自三台不同机器的SSD盘后就返回，后台异步地将数据迁移到HDD盘。

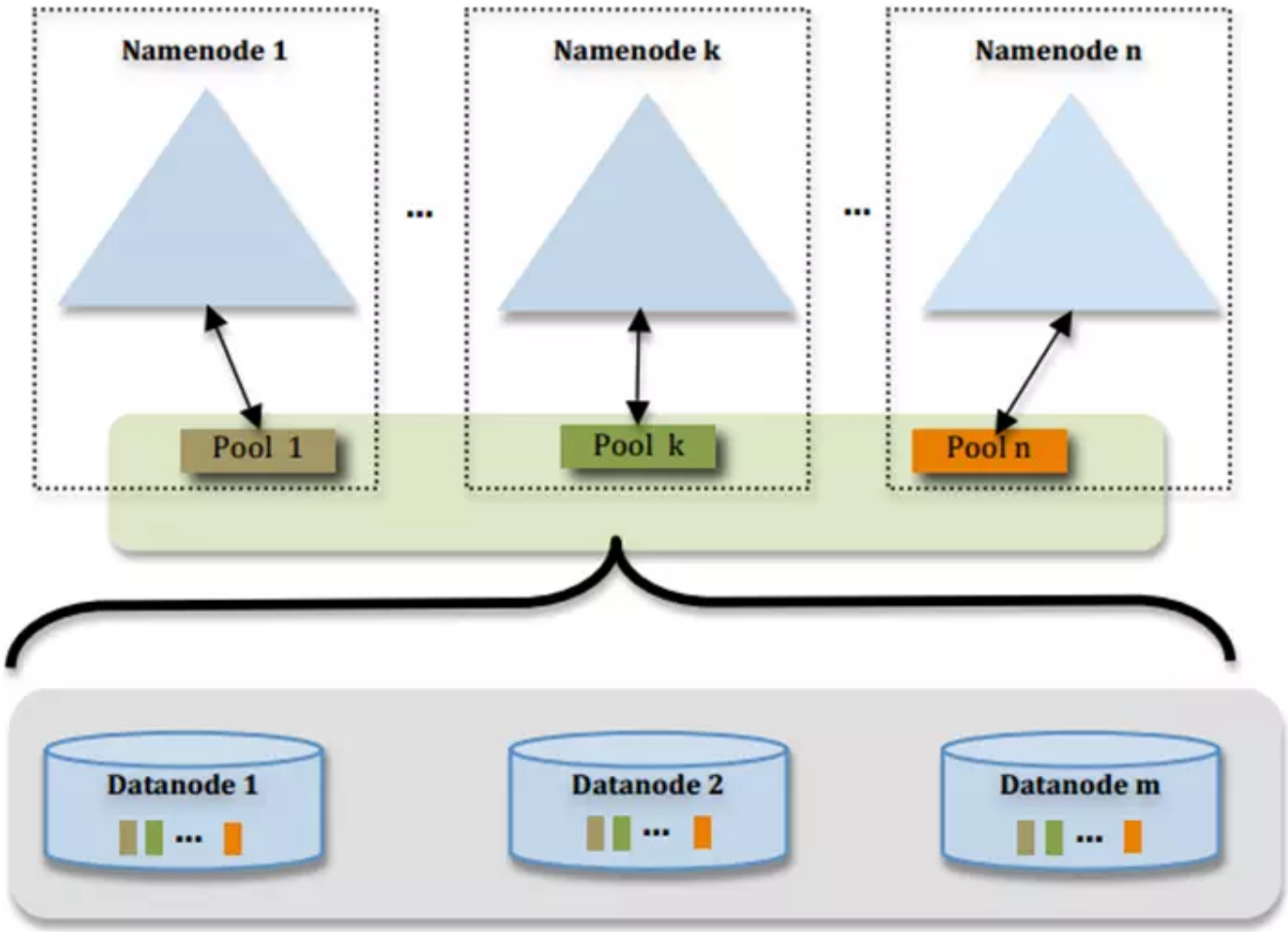
其次，ChunkServer采用了一系列技术来提供稳定的性能：1）服务分级，对请求队列和网络流量设定不同的优先级；2）管理好昂贵的后台活动；3）热点负载平衡；4）增加副本来应对重度使用的数据；5）缓冲来加速；6）备份请求（Backup Requests) 来规避慢盘等。这些技术的本质目标就是基于无法预估的资源来打造可以预测的整体，提供稳定的性能，通常用99.9%分位和99.99%分位的性能来表征。

第三，数据可靠性（Durability）和完整性（Integrity）是盘古的生命线。每一份数据成功地写入三台不同的机器（来自不同的可用区，或者不同的机架）后才返回。而且自动巡检系统不停地检测不可用的副本，一旦出现，自动地及时复制，使得每一份数据任何时刻至少有2个及以上的副本，提供至少10个9（11个9如果是多可用区）的可靠性。盘古系统也提供端到端的数据校验，上层的云存储产品提供数据的CRC，盘古在落盘的时候进行校验，并且把CRC和数据一起写入磁盘。后台任务也一直巡检，检查存储介质可能出现的位跳变（bit rot）错误，一旦发现和写入的CRC不符，找到正确的副本，重新复制一份新的副本。

第四，全自动的健康检查和主动规避低性能的机器和磁盘。在盘古所管理的成千上万的机器中，总是会有已经出现了健康问题的磁盘和机器，盘古文件系统根据线上历史上所有的运维操作和硬件故障做了机器学习，自动地将这些盘和机器过滤出来，进行慢盘规避，磁盘打分下线和机器调整，把隐患提前解决。

第五，支持多种访问方式并极致硬件的能力。在线访问追求低延迟，而离线访问追求高吞吐。盘古文件系统既要满足低延迟的在线访问，也要满足大并发大吞吐的离线访问，而传统的多线程系统在线程较多时，切换代价非常高。盘古文件系统从端到端采用协程的方式设计，使得在多任务的情况下，使用盘古文件系统的效率极高。内核态和用户态切换是另一个主要软件开销来源，盘古采用类似SPDK轻量级用户态文件系统来访问单机的磁盘，减少切换。高性能通信库和QoS是另一个重要的方面来提高存储的性能和资源的使用效率。

盘古 Master 扩展性



盘古基于Federation实现Master水平扩展。

Federation使用了多个独立的Namespace，namenode之间相互独立且不需要互相协调，各自分工，管理自己的区域。

分布式的datanode被用作通用的数据块存储存储设备。每个datanode要向集群中所有的namenode注册，且周期性地向所有namenode发送心跳和块报告，并执行来自所有namenode的命令。

一个block pool由属于同一个namespace的数据块组成，每个datanode可能会存储集群中所有block pool的数据块。每个block pool内部自治，也就是说各自管理各自的block，不会与其他block pool交流。一个namenode挂掉了，不会影响其他namenode。

某个namenode上的namespace和它对应的block pool一起被称为namespace volume(命名空间卷)。它是管理的基本单位。当一个namenode/nodespace被删除后，其所有datanode上对应的block pool也会被删除。当集群升级时，每个namespace volume作为一个基本单元进行升级。

总结：阿里云 盘古 和 GFS 的异同

盘古同 GFS 一样采用（主从）master/slave设计。一个存储集群是由一组Master和一定数目的Chunk Server组成，在数据读写时Client先请求Master获取到数据存储的元数据，之后的数据读写，Client就直接跟Chunk Server交互了。Chunk Server和Master保持心跳，向Master反馈数据状态以及接受Master的指令。

在系统中meta管理模块称为盘古master，类似于HDFS中的Namenode，数据存储模块称为Chunkserver，类似于GFS 中的Datanode。

同GFS的主要不同点在于下面几点：

- 在数据安全方面：
  - 采用Paxos协议实现盘古多master，实现meta服务的强一致性和高可靠；
  - Chunkserver自动检测磁盘状态，在磁盘异常情况下，主动通知盘古master进行性有目的备份，提高数据可靠性；
  - 采用嵌入式checksum，保证数据和checksum同时写入同时读取，在提高数据一致性的同时提高数据吞吐率。
- 在高可用和性能方面：
  - 对于每个文件可以设置分布属性，实现数据聚簇，计算节点按照数据locality特性进行调度来达到高性能需求。
  - 盘古client端实现基于数据位置相应时间的统计，可以有效绕过热点机器和磁盘，主动判断慢磁盘，降低了数据的平均响应延迟。
  - 在整个盘古系统中，各个模块使用同一套流控优先级控制策略，可以支持多个不同优先级业务并行使用；
  - 采用对用户透明的混合存储技术，让用户使用SATA磁盘的成本来享受SSD磁盘的高性能，同时做到不适用操作系统memory cache，将掉电丢失数据的概率降低为0。
  - Chunkserver不会定期汇报Replica信息，而由盘古master定期增量获取replica信息，有效防止了由于盘古master被动接受block汇报产生的性能影响。
  - 盘古master在内存组织方式采用更紧凑的内存组织方式，提高了单组盘古master可以承载的数据量；