分布式存储必读论文

作者：余利华 ┆ 2015-06-29 14:10

分布式存储泛指存储存储和管理数据的系统， 与无状态的应用服务器不同， 如何处理各种故障以保证数据一致， 数据不丢， 数据持续可用， 是分布式存储系统的核心问题，也是极具挑战的问题。 本文总结了分布式存储领域的经典论文， 供大家参考。

The Google File System. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung。 基于普通服务器构建超大规模文件系统的典型案例，主要面向大文件和批处理系统， 设计简单而实用。 GFS是google的重要基础设施， 大数据的基石， 也是Hadoop HDFS的参考对象。 主要技术特点包括： 假设硬件故障是常态（容错能力强）， 64MB大块， 单Master设计，Lease/链式复制， 支持追加写不支持随机写。

Bigtable: A Distributed Storage System for Structured Data. Fay Chang, Jeffrey Dean, Sanjay Ghemawat, et. 支持PB数据量级的多维非关系型大表， 在google内部应用广泛，大数据的奠基作品之一 ， Hbase就是参考BigTable设计。 Bigtable的主要技术特点包括： 基于GFS实现数据高可靠， 使用非原地更新技术（LSM树）实现数据修改， 通过range分区并实现自动伸缩等。

Spanner: Google's Globally-Distributed Database. James C. Corbett, Jeffrey Dean, et. 第一个用于线上产品的大规模、高可用， 跨数据中心且支持事务的分布式数据库。 主要技术特点包括， 基于GPS和原子钟的全球同步时间机制TrueTime， Paxo， 多版本事务等。

PacificA: Replication in Log-Based Distributed Storage Systems. Wei Lin, Mao Yang, et. 面向log-based存储的强一致的主从复制协议， 具有较强实用性。 这篇文章系统地讲述了主从复制系统应该考虑的问题， 能加深对主从强一致复制的理解程度。 技术特点： 支持强一致主从复制协议， 允许多种存储实现， 分布式的故障检测/Lease/集群成员管理方法。

Object Storage on CRAQ, High-throughput chain replication for read-mostly workloads. Jeff Terrace and Michael J. Freedman. 支持强一直的链式复制方法， 支持从多个副本读取数据。

Ceph: Reliable, Scalable, and High-Performance Distributed Storage. Sage A. Weil. 功能强大的开源海量存储系统， 支持文件系统、块设备、以及S3接口。 主要技术特色： CRUSH数据对象定位算法， 基于动态子树的文件系统元数据管理。

Finding a needle in Haystack: Facebook's photo storage. Doug Beaver, Sanjeev Kumar, Harry C. Li, Jason Sobel, Peter Vajgel. Facebook分布式Blob存储， 主要用于存储图片。 主要技术特色： 小文件合并成大文件， 小文件元数据放在内存因此读写只需一次IO。

Windows Azure Storage: A Highly Available Cloud Storage Service with Strong Consistency. Brad Calder, Ju Wang, Aaron Ogus, Niranjan Nilakantan, et. 微软的分布式存储平台， 除了支持类S3对象存储，还支持表格、队列等数据模型。 主要技术特点： 采用Stream/Partition两层设计（类似BigTable）；写错（写满）就封存Extent， 使得副本字节一致， 简化了选主和恢复操作； 将S3对象存储、表格、队列、块设备等融入到统一的底层存储架构中。

The Chubby lock service for loosely-coupled distributed systems. Mike Burrows. Google设计的高可用、可靠的分布式锁服务， 可用于实现选主、分布式锁等功能， 是ZooKeeper的原型。 主要技术特点： 将paxo协议封装成文件系统接口， 高可用、高可靠，但是不保证有很强性能。

Paxos Made Live – An Engineering Perspective. Tushar Chandra, Robert Griesemer，Joshua Redstone. 从工程实现角度说明了Paxo在chubby系统的应用， 是理解Paxo协议及其应用场景的必备论文。 主要技术特点： paxo协议， replicated log， multi-paxo。

Dynamo: Amazon's Highly Available Key-Value Store。 Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, et. Amazon设计的高可用的kv系统， 主要技术特点：综和运用一致性哈希，vector clock， 最终一致性构建一个高可用的kv系统， 可应用于amazon购物车场景。

标签： 服务端 分布式

文章来源

服务端性能优化

1144人

相关圈子

云计算

912人

加关注

产品前端技术交流

234人

加关注

移动端公共组件

615人

加关注

服务器端的那些事

911人