分布式系统第一次调研

美团技术团队分布式系统相关探索

2022

美团集群调度系统的云原生实践

有关云原生操作系统的概念解释，可参考这里

> 多年之后，也许大学课程“操作系统原理”里的内容会因为这次变革而发生翻天覆地的改变，但我们坚信，这就是云带给我们这一代操作系统研发人员的使命：重新定义操作系统。只有通过因云而生的技术创新打造出来的操作系统，才是真正的云原生操作系统。

相关概念——云计算

2021

美团图数据库平台建设及业务实践

基于技术调研后选择了 NebulaGraph 图数据库

> NebulaGraph 基于 C++ 实现，架构设计支持存储千亿顶点、万亿边，并提供毫秒级别的查询延时。我们在 3 台 48U192G 物理机搭建的集群上灌入 10 亿美食图谱数据对 NebulaGraph 的功能进行了验证。
>
> - 一跳查询 TP99 延时在 5ms 内，两跳查询 TP99 延时在 20ms 内，一般的多跳查询 TP99 延时在百毫秒内。
> - 集群在线写入速率约为20万 Records/s。
> - 支持通过 Spark 任务离线生成 RocksDB 底层 SST File，直接将数据文件载入到集群中，即类似 HBase BulkLoad 能力。
> - 提供了类 SQL 查询语言，对于新增的业务需求，只需构造 NebulaGraph SQL 语句，易于理解且能满足各类复杂查询要求。
> - 提供联合索引、GEO 索引，可通过实体属性或者关系属性查询实体、关系，或者查询在某个经纬度附近 N 米内的实体。
> - 一个 NebulaGraph 集群中可以创建多个 Space （概念类似 MySQL 的DataBase），并且不同 Space 中的数据在物理上是隔离的。

解决了美团实际应用下在集群中部署图数据库的几个问题：高可用模块设计（不怕挂）、每小时百亿量级数据导入模块设计（大规模数据导入集群）、实时写入多集群数据同步模块设计（实时性）、图可视化模块设计（用户交互）

图数据库在美团中的应用：知识图谱的建立，千亿级数据下多跳查询的需求，代码依赖分析等等

OCTO 2.0：美团基于Service Mesh的服务治理系统详解

2020

Kubernetes如何改变美团的云基础设施?

同样是云原生操作系统相关

2019

保障IDC安全：分布式HIDS集群架构设计

分布式系统安全相关

HIDS全称是Host-based Intrusion Detection System，即基于主机型入侵检测系统。作为计算机系统的监视器和分析器，它并不作用于外部接口，而是专注于系统内部，监视系统全部或部分的动态的行为以及整个计算机系统的状态。

通俗理解我们可以这样认为：一个成功的入侵者一般而言都会留下他们入侵的痕迹。这样，计算机管理员就可以察觉到一些系统的修改，HIDS亦能检测并报告出检测结果。

主要工作：HIDS 面对几十万台甚至上百万台规模的 IDC 环境时，系统架构的设计

2018

美团即时物流的分布式系统架构设计

深度剖析开源分布式监控CAT

CAT（Central Application Tracking）是一个实时和接近全量的监控系统，它侧重于对Java应用的监控，基本接入了美团上海侧所有核心应用。目前在中间件（MVC、RPC、数据库、缓存等）框架中得到广泛应用，为美团各业务线提供系统的性能指标、健康状况、监控告警等。

2017

Mt-Falcon——Open-Falcon在美团点评的应用与实践

监控系统是整个业务系统中至关重要的一环，它就像眼睛一样，时刻监测机房、网络、服务器、应用等运行情况，并且在出现问题时能够及时做出相应处理。

在一个大型分布式系统中，任何一个client的请求调用，会在分布式系统中产生上百次的调用，一旦一个请求异常，那具体到哪个系统服务异常就变得很重要的。

Leaf—美团点评分布式ID生成系统

在复杂分布式系统中，往往需要对大量的数据和消息进行唯一标识。如在美团点评的金融、支付、餐饮、酒店、猫眼电影等产品的系统中，数据日渐增长，对数据分库分表后需要有一个唯一ID来标识一条数据或消息，数据库的自增ID显然不能满足需求；特别一点的如订单、骑手、优惠券也都需要有唯一ID做标识。此时一个能够生成全局唯一ID的系统是非常必要的。

如果ID生成系统瘫痪，整个美团点评支付、优惠券发券、骑手派单等关键动作都无法执行，这就会带来一场灾难。

如何在分布式系统中实现 ID 的快速可靠的生成

2016

分布式块存储系统Ursa的设计与实现

云硬盘对IaaS云计算平台有至关重要的作用，几乎已成为必备组件，如亚马逊的EBS(Elastic Block Store)、阿里云的盘古、OpenStack中的Cinder等。云硬盘可为云计算平台带来许多优良特性，如更高的数据可靠性和可用性、灵活的数据快照功能、更好的虚拟机动态迁移支持、更短的主机故障恢复时间等等。随着万兆以太网逐渐普及，云硬盘的各项优势得到加强和凸显，其必要性变得十分强烈。

云硬盘的底层通常是分布式块存储系统，目前开源领域有一些此类项目，如Ceph RBD、Sheepdog。另外MooseFS和GlusterFS虽然叫做文件系统，但由于其特性与块存储系统接近，也能用于支持云硬盘。我们在测评中发现，这些开源项目均存在一些问题，使得它们都难以直接应用在大规模的生产系统当中。例如Ceph RBD的效率较低（CPU使用过高）；Sheepdog在压力测试中出现了数据丢失的现象；MooseFS的POSIX语义支持、基于FUSE的架构、不完全开源的2.0版本等问题给它自身带来了许多的局限性；GlusterFS与Ceph同属红帽收购的开源存储系统，主要用于scale-out文件存储场景，在云计算领域使用不多。此外，这些存储系统都难以充发挥用万兆网卡和SSD的性能潜力，难以在未来承担重任。

由于以上原因，美团云研发了全新的分布式块存储系统Ursa，通过简单稳固的系统架构、高效的代码实现以及对各种非典型场景的仔细考虑，实现了高可靠、高可用、高性能、低开销、可扩展、易运维、易维护等等目标。Ursa的名字源于DotA中的熊战士，他具有极高的攻击速度、攻击力和生命值，分别隐喻存储系统中的IOPS、吞吐率和稳定性。

分布式队列编程优化篇

分布式系统互斥性与幂等性问题的分析与解决

分布式会话跟踪系统架构设计与实践

MTDDL—美团点评分布式数据访问层中间件

个人想法

方向：云计算 + 操作系统；监控系统；云计算 + 监控系统；分布式存储系统