分　数：	
评卷人：	

華中科技大學

研究生（数据中心技术）课程论文（报告）

题　目：分布式存储系统优化模型

学　　号　M202173812

姓　　名　姜志鹏

专　　业　电子信息

课程指导教师　施展 童薇

院（系、所）　计算机科学与技术学院

2022 年 1 月 3 日

分布式存储系统优化模型

姜志鹏[1)]

[1)](华中科技大学计算机科学与技术学院，武汉 430074)

摘 要 数据中心是信息技术和云服务的基础设施，本质上是一种分布式存储系统，因此数据中心的发展与分布式存储系统的发展紧密相关，并成为研究热点。故本文从分布式存储系统模型优化出发。本文调研了三篇文献，Hellings 等人提出了一种可以线性编排、集中式编排和分布式编排事务所涉及分区的框架 BySHARD，在使用分布式编排数据分区时可以有效降低时间消耗和系统的吞吐量。Abebe 等人介绍了一种分布式数据库系统 MorphoSys，它可以根据工作负载动态选择和修改其物理设计。Durner 等人提出了一种新的与计算并行智能缓存存储系统的架构，可以显著改善未修改 Spark 和 Greenplum 上的查询延迟。

关键词：数据中心；分布式存储；分布式数据库；区块链

Distributed storage system optimization model

Jiang ZhiPeng[1)]

[1)](School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

Abstract Data center is the infrastructure of information technology and cloud services, and it is a distributed storage system in essence. Therefore, the development of data center is closely related to the development of distributed storage system, and has become a research hotspot.Therefore, this paper starts from model optimization of distributed storage system. In this paper, Hellings et al. propose a framework BySHARD that can be used for linear, centralized and distributed orchestration transactions involving partitioning, which can effectively reduce time consumption and system throughput when using distributed orchestration data partitioning. Abebe et al. introduced MorphoSys, a distributed database system that can dynamically select and modify its physical design based on the workload.Durner et al. propose a new architecture for a computationally parallel intelligent cache storage system that can significantly improve query latency on unmodified Spark and Greenplum.

Keyword data center；distributed storage；distributed database；blockchain

一、引言

现代数据库系统通过数据副本和数据分区来存储和管理大量的数据，将它们的事务处理分布到多个节点上。为分布式数据库选择的数据副本和分区方案形成了它的物理设计。目前，区块链技术的出现激发了人们对新的分布式存储系统的兴趣，这些系统可以在拜占庭式行为的情况下提供数据和事务处理，例如，来自软件、硬件或网络故障的错误行为，或来自恶意结点攻击的行为。因此，受区块链技术启发的系统可以防止由于系统部分故障而导致的服务中断，并可以提高由许多独立方管理的数据质量，潜在地降低与两者相关的巨大成本。另外，分析数据库存储系统逐渐向云存储转变，这是由云的灵活性和随用随付功能驱动的。此类数据库采用分级或分类存储模型，弹性计算层访问存储在独立可伸缩的远程云存储上的数据。

总之，对传统分布式系统、拜占庭环境下的分布式系统和云存储都面临者同样的问题，即数据副本与数据分区在进行事务处理的时候如何决策，一个不好的决策会显著地降低系统地性能，因此在进行多分片事务处理时，如何对事务所涉及地数据副本和分区进行编排，这一研究方向有了很大地意义，在综述地三篇文章中，Hellings 等人提出了一种可以线性编排、集中式编排和分布式编排事务所涉及分区的框架 BySHARD，其实验结果表明在使用分布式编排数据分区时可以有效降低时间消耗和系统的吞吐量。Abebe 等人介绍了一种分布式数据库系统 MorphoSys，它可以根据工作负载动态选择和修改其物理设计，MorphoSys 使用学习到的成本模型对所有的数据分区、复制和放置决策进行集成设计，MorphoSys 通过一种新颖的并发控制和更新传播方案，在面对设计更改时提供了高效的事务执行，实验表明与使用了几个基准工作负载和最先进系统相比，MorphoSys 提供了优秀的系统性能。Durner 等人提出了一种新的与计算并行智能缓存存储系统的架构，称为 Crystal。Crystal 的客户端是具有下推谓词的特定于 dbms 的“数据源”。在本质上类似于 DBMS, Crystal 集成了查询处理和优化组件，专注于高效缓存和服务于称为区域的单表超矩形。结果表明，使用一个小型的 dbms 特定数据源连接器，Crystal 可以显著改善未修改 Spark 和 Greenplum 上的查询延迟，同时还可以节省带宽从远程存储。

二、原理和优势

2.1 BySHARD

其中，为了处理多分片事务，BySHARD 引入了 orchestrate-execute model (OEM)。该模型可以在每个涉及的分片最多两个一致步骤中合并处理一个多分片事务所需的所有提交、锁定和执行操作。OEM 的第一个组件是 orchestration:在所有涉及的分片之间复制事务，同时就事务是否可以提交达成原子决策。OEM 的第二个组成部分是 execution of transactions。为了提供在分片之间保持数据一致性的执行功能，该方法展示了如何以最小的成本(根据所涉及的分片的一致步骤)将标准的两阶段锁定样式的执行适应于拜占庭环境。在 BySHARD 框架中，组件 orchestration 文章提出了三种方式：线性编排、集中编排和分布式编排，如图 1 所示，其中线性编排是事务 T 涉及的分区 Si，当第一个分区 S1 决策完之后如果同意执行则通知下一个分区进行决策，串行执行分区决策，当中间任意一个结点决策终止执行事务时，立即终止事务的提交，因此在最坏的执行情况下，需要执行分区数+1 次，毫无疑问时间复杂度很高，针对这个问题，文章又提出了集中式编排和分布式编排，前者是先选出一个中心结点，中心结点如果决策执行事务时，向其他分区结点广播投票，其他结点收到广播后将投

票结果返回给中心结点，最后再由中心结点决策是否执行。后者在前者的基础上，改变了其他结点向中心结点返回结果这一步骤，变为了每个结点决策完后将自己的结果广播给除自身外的所有结点，这样就增大了系统的吞吐量，实验证明，集中式编排与分布式编排比线性编排有着更高的性能，而分布式编排比集中式编排有更高的吞吐率。

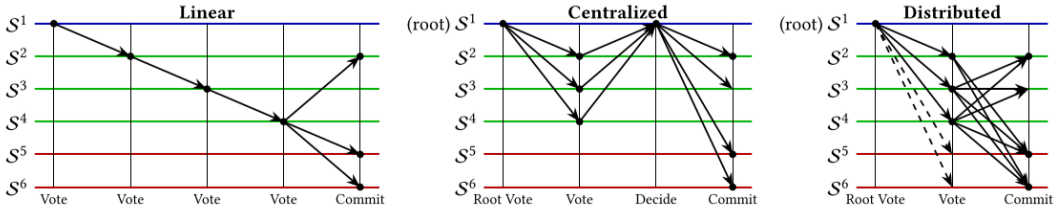

图 1 BySHARD 事务的三种编排方式

2.2 MorphoSys

MorphoSys（如图 2 所示）不断捕获事务工作负载并对其建模，以做出设计决策。事务路由器对提交的事务进行采样，并捕获数据项读写访问频率。对于分区 p，事务路由器维护读 R(p)和写 W(p)到分区的概率，而不是所有的分区访问。事务路由器计算 R(p)(或 W(p))的方法是将每个分区读(写)计数除以所有分区的所有读和写操作的运行计数。给定一个物理设计，事务路由器使用一个学习到的成本模型来估计执行事务和应用物理设计更改的成本。这个成本模型预测设计更改和事务执行操作的延迟。MorphoSys 对这些操作的实现转化为一种自然的系统延迟分解:在站点上等待服务、等待更新、获取锁、读写数据项和提交。因此，我们将成本模型分解为五个相应的成本函数，MorphoSys 学习并结合这些函数来预测操作的延迟。实验证明，与以前的方法相比，MorphoSys 提高了巨大的性能，同时避免了使用需要以前工作负载信息的静态设计。

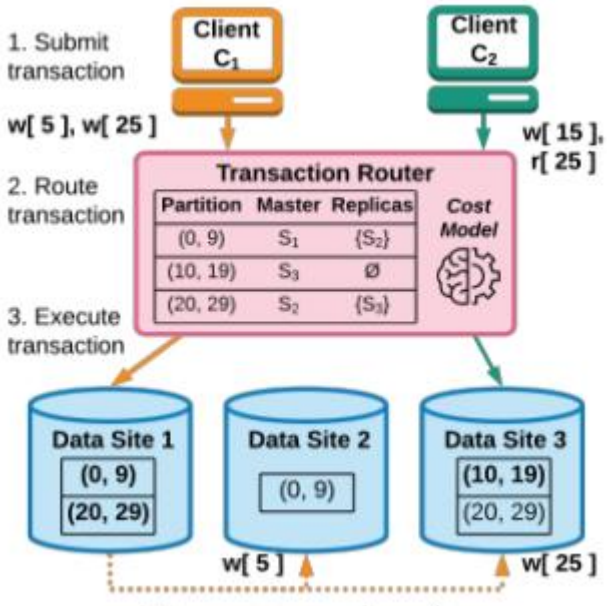

图 2 MorphoSys 架构

2.3 Crystal

Crystal（如图 3）具有足够的通用性，以便它能够以最小的工程工作量插入现有的大数据系统。因此，Crystal 被架构成两个独立的组件:一个轻量级 dbms 特定的数据源连接器和 Crystal CMS 过程。在后台，Crystal 会收集查询的历史轨迹，并调用一个缓存 Oracle Plugin 模块来为 OR 缓存计算最佳内容。使用 RR 和 OR 缓存中的远程存储和现有内容组合填充新内容。Crystal 的架构目的是使其易于与任何云分析系统一起使用。Crystal 提供了三个扩展点。首先，用户可以根据自己的工作负载定制一个自定义实现来替换缓存 oracle。其次，远程存储适配器可以被替换为与任何云远程存储一起工作。第三，可以为每个需要使用 Crystal 的 DBMS 实现定制连接器。

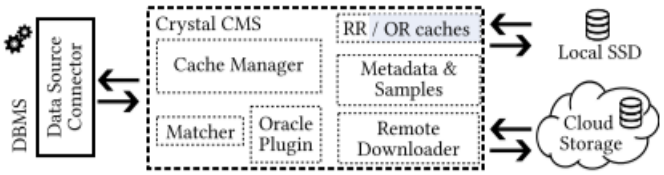

图 3 Crystal 的架构

三、研究进展

关于分布式存储系统，近期的研究热点仍然是针对延迟、性能，稳定性等方面进行更深入的优化。在保留与深化优势部分的同时，还要对缺陷部分各个击破，实现全方位的提升。分布式存储是一个技术难度被显著低估的领域。随着 HDFS、Lustre、GlusterFS、Ceph 等开源分布式软件项目的兴起，一个普通软件工程师可以在一天或一周时间内搭建一个 PoC 分布式存储系统。包括华为、新华三、以及一些初创公司，在开源项目上优化，推出各自分布式存储产品，同质化现象严重，同时给行业带来假象，似乎分布式存储没有什么技术门槛。然而，无论是互联网公司，还是在企事业单位，由于使用分布式存储导致的业务中断、甚至数据丢失的报导屡见不鲜。

分布式存储技术主要针对三个方面，分别是(1)数据副本，(2)数据分区，(3)主数据位置。迄今为止，任何相关工作都不能以动态方式同时实现上述三种方案中的一种以上。目前在数据副本这一块，动态副本调整工作负载的复制内容和复制位置。非事务性存储系统使用的自适应复制算法 和 keyvalue 存储根据工作负载描述和操作成本估算优化副本放置。这些系统只在每个站点的基础上复制，而不是在每个数据项或分区的基础上复制;相比之下，MorphoSys 是具有 ACID 语义的事务性的，在为单个数据分区做出复制决策时考虑了整个工作负载，并保证了单站点执行。

动态分区系统这一块，动态地将数据分组到分区中，以解决访问倾斜问题。这些系统依赖昂贵的两阶段提交协议(2PC)来协调多站点事务执行。MorphoSys 通过使用动态物理设计操作符(改变主数据及其副本的位置)来保证单站点事务。Clay 在生成修订的数据分区方案之前，会在较大的时间间隔内观察工作负载，然后周期性地做出数据分区决策。

动态掌握通过将主数据放在一个站点上来保证单站点事务的执行。MorphoSys 为此目的采用了动态控制，并利用副本来改变主控不同于 STAR 和 DynaMast，它们需要完整的数据复

制。所有之前的动态掌握系统都需要静态的先验分组数据到分区，而 MorphoSys 动态分区数据，以减少竞争的方面。STAR 和 slog 采用批处理的事务执行，MorphoSys 在事务到达时执行，以减少延迟。

最近几年分布式存储系统的成果都主要围绕着区块链这一块技术,区块链在加密货币之外被广泛应用于各种新的应用领域，如医药、供应链管理、全球贸易和政府服务。区块链是以比特币为代表的数字加密货币体系的核心支撑技术。区块链技术的核心优势是去中心化，能够通过运用数据加密、时间戳、分布式共识和经济激励等手段，在节点无需互相信任的分布式系统中实现基于去中心化信用的点对点交易、协调与协作，从而为解决中心化机构普遍存在的高成本、低效率和数据存储不安全等问题提供了解决方案。随着比特币近年来的快速发展与普及，区块链技术的研究与应用也呈现出爆发式增长态势，被认为是继大型机、个人电脑、互联网、移动/社交网络之后计算范式的第五次颠覆式创新，是人类信用进化史上继血亲信用、贵金属信用、央行纸币信用之后的第四个里程碑。区块链技术是下一代云计算的雏形，有望像互联网一样彻底重塑人类社会活动形态，并实现从目前的信息互联网向价值互联网的转变。

四、总结和展望

4.1 总结

本文一共引入了三个最新的分布式存储相关的框架：ByShard，这是一个用于分片弹性数据管理系统的通用框架。此外，引入了 orchestrate-execute model(OEM)来处理 ByShard 中的多分片事务。接下来，我们展示了 OEM 可以将处理多分片事务所需的提交、锁定和执行步骤合并到每个涉及的分片最多两个一致步骤中。接着介绍了 MorphoSys，这是一个分布式数据库系统，可以自动修改其物理设计以提供卓越的性能。MorphoSys 集成了分布式设计的三个核心方面:将数据分组到分区中，选择分区的主站点，以及定位复制的数据。MorphoSys 使用学习到的成本模型进行综合设计决策,并使用基于分区的并发控制和更新传播方案动态执行设计更改。MorphoSys 比之前的方法提高了 900 倍的性能,同时避免了静态设计的使用。最后介绍了 Crystal，一个智能缓存存储系统，与计算并行，可以通过数据源连接器客户端被任何未修改的数据库使用。Crystal 在语义数据区域上操作，并不断调整本地缓存的内容以获得最大收益。结果表明，晶体能明显改善曲度。

4.2 展望

国内近几年,从海外 IBM、EMC、HPE 等传统存储强企吸引了不少高端存储人才回国创业，已经产生了一批极具创新力的企业。以南京道熵为例，其铁力士分布式存储采用双重 RAID 架构，通过本地 RAID 与分布式技术相结合，代表了下一代分布式存储的发展方向。分布式存储将会继续落地多领域应用场景：分布式存储的特性，让数据存储、文件传输、网络视频、社交媒体及去中心化交易等多个领域都是分布式存储的应用场景，完善互联网技术设施，推动互联网更好发展：同人工智能和大数据的等，分布式存储存储依然是互联网基础设施，并且在当前推动 5G 新基建的大环境下，分布式存储更能推动互联网的发展，能更好的将互联网提升一个水平。区块链分布式存储对当今的中心化存储是一个非常大的补充，分布式浪潮的来临并不是要取代当下的中心化互联网，而是要让未来数据存储发展的更好，为整个市场生态带来无法想象的活力。另外，对于区块链，分布式存储技术是基础和关键。在此基础上，

通过更安全地存储和使用每个人产生的宝贵数据，并确保用户对其数据的所有权，从信息互连到价值互连。

五、参考文献

[1] Hellings J, Sadoghi M. Byshard: Sharding in a byzantine environment[J]. Proceedings of the VLDB Endowment, 2021, 14(11): 2230-2243.

[2] Abebe M, Glasbergen B, Daudjee K. MorphoSys: Automatic physical design metamorphosis for distributed database systems[J]. Proceedings of the VLDB Endowment, 2020, 13(13): 3573-3587.

[3] Durner D, Chandramouli B, Li Y. Crystal: a unified cache storage system for analytical databases[J]. Proceedings of the VLDB Endowment, 2021, 14(11): 2432-2444.