互联网环境下分布式事务处理系统现状与趋势

马鹏玮，魏凯，姜春宇
中国信息通信研究院云计算与大数据研究所，北京 100191

摘要
传统集中式数据处理系统由于摩尔定律的逐渐失效，开始显露出性能瓶颈，无法有效支撑当前业务的快速发展。同时由于安全、成本、政策等原因，原有集中式数据处理系统向分布式数据处理系统转型成为必经之路。从分布式事务处理系统的发展背景、优势及挑战、技术特点、应用要点、未来发展方向等方面介绍了分布式事务处理系统的意义、现状及趋势，为后续分布式事务处理系统的应用及发展提供了理论基础和指导。

关键词
分布式；数据库；事务

中图分类号：TP392 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018021

State of the art and future research on distributed processing system in Internet environment

MA Pengwei, WEI Kai, JIANG Chunyu
Cloud Computing and Big Data Research Institute of China Academy of Information and Communications Technology, Beijing 100191, China

Abstract
Due to the gradual failure of Moore's law, traditional centralized data processing systems begin to show performance bottlenecks, which can't effectively support the rapid development of current business. At the same time, due to the security, cost, policy and other reasons, the transformation of the original centralized data processing system into a distributed data processing system has become the only way. The significance, present situation and prospect of distributed transaction data processing system from its development background, advantages and challenges, technical characteristics, application points, future development direction were introduced. A theoretical basis and guidance for the application and development of distributed transaction data processing system was provided.

Key words
distributed, database, transaction

1 分布式事务处理系统发展背景

中国互联网持续快速发展，带来了当前信息的快速消费以及数据的快速增长，这种优势为中国创造了巨大的信息红利。与此同时，指数级增长的互联网业务量也为背后的信息支撑系统带来了巨大的压力与挑战。以阿里巴巴淘宝网为例，2009—2017年“双十一”交易额如**图1**所示。

巨大交易增长量的背后，后台信息支撑系统除了直接面向用户的电商网站系统外，相关的物流系统、支付系统、仓储系统、订单系统等也都接受了巨大的考验。而上述各类信息系统的核心数据处理能力，几乎均与数据处理系统的性能息息相关。

各类数据处理场景可以分为事务场景和非事务场景两大类。事务场景即数据处理过程需要满足原子性、一致性、隔离性、持久性（atomicity、consistency、isolation、durability，ACID）4个特性。典型的场景如银行账户间转账、物流仓库的运输等。非事务场景对数据处理过程没有对ACID特性的严格要求。典型的场景有搜索展现场景、报表分析场景等。

因为事务场景具有ACID特性，所以相比于非事务场景，事务场景对底层数据处理系统提出了更高的技术要求。以银行支付系统为例，目前核心的支付系统经过几十年的发展，形成了采用所谓的IOE（即IBM公司的大中型服务器、Oracle公司的数据库软件、EMC公司的存储设备）结构作为事务场景的支撑系统。这套系统结构用于支撑银行传统的柜台业务没有大问题，但是随着国内移动互联网以及电商业务的快速发展，其中存在的问题逐渐暴露。

（1）性能上限

传统的IOE结构为集中式系统，即所有的软件系统都运行在一个性能十分强大的服务器上，例如IBM公司的AS 400、AS 390等。当业务压力变大时，一般采用纵向扩展（scale up）方案，即继续增加单台服务器的性能上限。但是随着摩尔定律逐渐失效，靠增大压缩器件尺寸提高单台服务器性能上限的方案已经逐渐失效，纵向扩展路线即将走到尽头。

（2）成本巨大

当业务量上升时，不仅硬件系统的规模需要呈指数级增长，同时软件系统相应的服务费用也会随之增长，中国银行、中国农业银行、中国建设银行、中国工商银行每年相应的支出达到数亿元，且仍在持续增长。

（3）严重捆绑

采用IOE结构的公司，因为各环节信息系统供应商的垄断化及单一化，会造成严重的服务提供商捆绑问题。传统的IOE

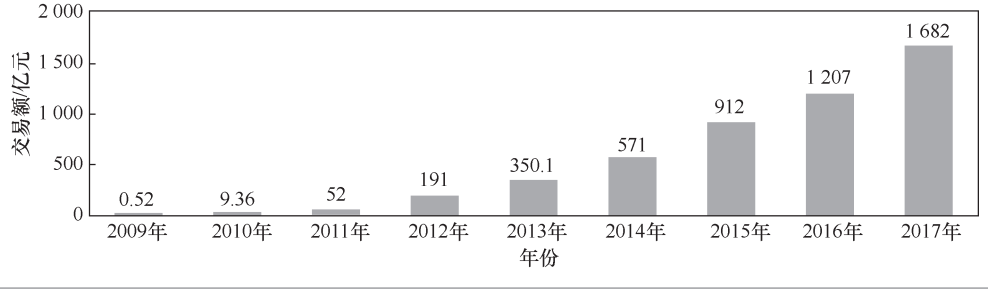

图1 淘宝网“双十一”交易额

服务提供商会对相应功能实施生态封闭策略，造成服务消费方对其服务的严重依赖，这为未来埋下了严重的隐患。例如IBM公司近年来市场份额逐渐下降，未来方向尚不明确，如果将来IBM公司不再提供服务器的供应或维护，那当前相应的IOE体系服务消费方将会遭受严重打击。

（4）非自主化产品

IOE体系均为国外的技术与产品，这方面需要进行自主化改造，从而掌握核心技术主动性。

（5）适应新业务环境

随着业务产品的逐渐丰富，底层数据系统架构需要增加扩展能力，从而适应多种业务类型的灵活性要求，而集中式架构在这方面具有先天缺陷。

为解决以上各类问题，各大技术公司逐渐开始研发分布式事务数据系统技术，从而逐渐在事务场景中替代传统的IOE结构。

2 分布式事务数据处理系统优势及面临的挑战

分布式事务数据处理系统是利用高速通信网络以及多台服务器的分布化方式，解决数据处理事务场景的数据处理系统，其具有以下特性。

● 以横向扩展路线代替纵向扩展路线。通过集群模式，将任务和数据分发到多台服务器上，多台服务器共同协作，通过增加横向扩展服务器的数量达到提高单台服务器性能上限的目的，从而解决性能上限问题。

● 使用廉价服务器及开源产品解决成本问题。分布式事务处理系统在底层不再采用成本巨大的超级服务器以及专有商业事务数据库，而是采用多台廉价的x86服务器以及开源的数据库产品达到相应效果，预计单一企业在该方面的成本将从千万级下降到百万级。

● 多元化、开源化解决服务商捆绑问题。分布式事务处理系统可从硬件、软件等多个环节对接业内多种企业或开源产品。例如硬件支撑系统可以采用华为、浪潮、戴尔等多种x86服务器提供商提供的相应产品，软件支撑系统可以采用MySQL、MariDB、PostgreSQL等多款开源数据库产品，从而解决严重捆绑的问题。

● 国内企业研发解决非自主可控问题。当前国内涌现出多款自主研发分布式事务处理系统产品，例如阿里巴巴的OceanBase数据库、腾讯的TDSQL数据库、中兴通讯的GoldenDB数据库、PingCAP的TiDB数据库等。这样，可以通过国内企业的创新力量解决非自主可控的问题。

理论上分布式事务数据处理系统横向扩展路线可以达到数据处理能力上限。但是为了保证事务场景ACID特性，分布式事务数据处理系统无法达到其理论上限。深层次的原因为分布式事务数据处理系统在数据处理过程增加了网络传输环节，而根据统计分析，数据在网络中传输处理的速度大概为单机系统内传输处理的千分之一，所以传统事务数据处理系统中成熟的保证事务场景ACID的技术方案无法直接应用到分布式事务数据处理系统中。

为了研究如何使分布式事务数据处理系统保障事务场景ACID特性的问题，学术界提出：一致性、可用性、分区容错性（consistency、 availability、partition tolerance，CAP）无法同时全部满足，只能同时满足两个特性。通俗解释为：为了满足分区容错性，数据库系统中的数据需要同时保证多副本分散分布，而多副本分散分布会带来无法保证各副本间数据一致性的问题，而严格保证各副本间数据一致

性，又会导致数据库系统性能下降。所以，CAP 3个特性无法同时得到满足。

在分布式体系结构中，分区现象是默认存在的，所以分区容错性默认必须得到满足。而事务特性要求一致性必须得到满足，所以相应的分布式事务数据处理系统必须保证分区容错性和数据一致性，酌情降低了系统可用性。

同时，分布式体系结构存储数据的方式相对于集中式存储系统存储数据的方式来说，增加了两种新型的扩展形式。

第一种数据扩展形式为数据分区处理，即“大表变小表”。当一张表的容量达到上限或者出现业务交叉时，常进行水平拆分处理，即将一张表的数据拆分成几张表共同存储，各表分别放在不同的服务器上，从而针对不同小表访问的业务可以分散到不同的服务器上，减少单点问题，提高可用性。

第二种数据扩展形式为数据镜像处理，即“一份变多份”。一张表进行数据镜像，从而形成内容为同一份数据的多张表，但是从物理位置上看，却分散在不同的服务器上。针对某一份数据的业务可以分散到这份数据其余的镜像数据服务器上，从而提升服务性能。同时，当其中某台服务器发生故障、无法提供服务时，它的数据镜像所在的服务器仍然能够提供服务，提升了分区容错性。

新增的数据扩展形式在带来优点的同时，也为分布式事务性的保障提出了新的挑战。

- 首先，针对数据分区扩展形式产生了分布式事务中的多机协作问题。当某个事务处理过程涉及多个分区的数据，而这多个分区的数据分散在不同的服务器上时，通过服务器之间的网络通信，保证事务的ACID属性成为挑战。
- 其次，针对数据镜像扩展形式产生了分布式事务中的数据同步问题。当某个事务针对某份数据进行更改操作时，利用服务器间网络通信使这份数据的镜像数据与之保持同步，从而维护数据镜像的特点，使对外提供服务的镜像数据与源数据之间不会出现数据不一致的现象，这是另一个挑战。

3 分布式事务数据处理系统各类技术方案

数据分区扩展形式带来的挑战通常称为分区一致性（consistency）问题，数据镜像扩展形式带来的挑战通常称为副本一致性（consensus）问题。

针对分区一致性问题，核心解决方案依然是沿袭传统集中式数据库系统的事务问题解决方案，即处理好事务过程中锁与并发的控制问题。传统集中式数据库系统解决此问题的方式为两阶段锁（2 phase lock，2PL）方案，即通过两阶段锁控制本次事务涉及的资源，从而达到事务属性。在分布式系统中，原有的2PL方案转化为两阶段提交（2 phase commit，2PC）类方案，即通过分别向同一事务涉及的多台服务器发送两阶段的提交要求，完成同步提交的效果。2PC方案根据各类事务隔离级别要求，后续又被改进为3PC、2+XPC、异步消息队列、TCC（try confirm cancel）方案等。目前各类应用都是针对不同的实际业务场景，抽象出业务的实际隔离级别需求，然后在上述各种方案中选择合适的分区一致性解决方案。

副本一致性的核心问题为分布式系统中的数据复制问题。由于分布式系统增加了网络传输的延迟性和不确定性，一致性方案主要分为服务器逻辑架构设计解决方案和数据自组织解决方案。常见服务器逻辑架构解决方案设定为主从模式或者主主模式，通过掐断数据复制过程中从服务

器的数据输出能力，达到副本一致性的目的。常见的数据及组织解决方案为Paxos、Raft、Zab等一致性算法协议，通过自由选举、过半投票确认、分区容错的解决方式自组织解决副本一致性问题。

业内为了综合解决上述两个问题，从各个方向逐渐发展出了3种相应的架构方案。

第一种架构方案为中间件方案，其架构示意如**图2**所示。

整个分布式事务数据处理系统架构分为3层，最上层为应用层，第二层为中间件层，第三层为数据库实例层。其中，底层数据库实例层常采用开源的集中式数据库，

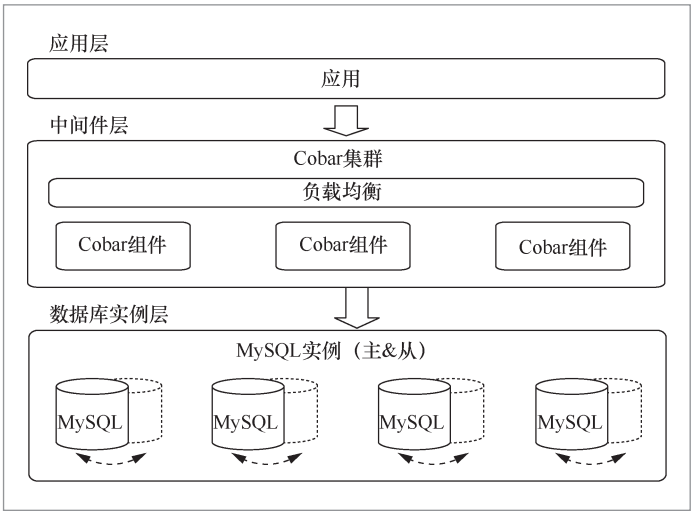

图 2 Cobar 中间件方案架构

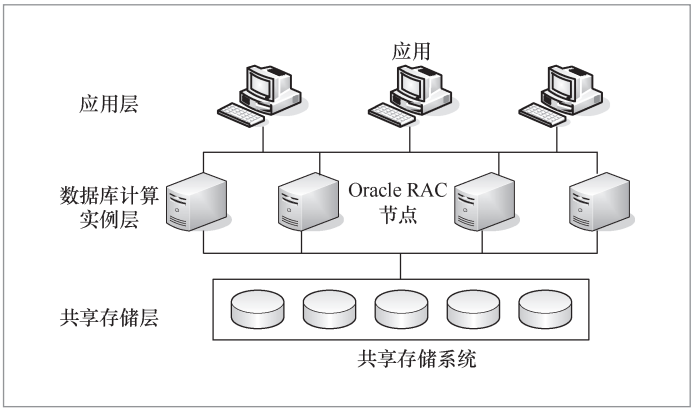

图 3 Oracle RAC 共享存储方案架构

负责本实例上的数据存储与处理；中间件层负责手动数据分区之后的分区信息维护、上层应用下发的SQL语句分布式解析下发、分布式事务的统一执行控制等，从而实现数据分区与分布式事务的实现。

第二种架构方案为共享存储方案，其架构示意如**图3**所示。

整个分布式事务数据处理系统架构分为3层，最上层为应用层，第二层为数据库计算实例层，第三层为共享存储层。数据库计算实例层通过分布式扩展，保证集群的计算性能，增强可用性。同时共享存储层采用新型存储硬件结构，保证高速I/O性能的同时，避免了存储层数据网络通信的环节，通过改进一定量传统的集中式数据库系统的事务解决方案，保障了共享存储架构中的分布式事务能力。但是由于统一式存储带来的并发和I/O能力的理论极限，集群规模无法做到大规模扩展。业内实践案例表明，当此类解决方案集群数量达到百台规模时，性能将会出现比较大的下滑。

第三种架构方案为计算存储均分布式扩展方案，其架构示意如**图4**所示。

整个分布式事务数据处理系统架构分为3层，最上层为应用层，第二层为分布式计算层，第三层为分布式存储层。分布式计算层与共享存储方案的数据库计算实例层相似，均通过分布式扩展能力保证集群的计算性能，增强可用性。而分布式存储层采用零共享（shared nothing）架构，同时大部分产品具备自动化分区功能，不需要人为定义分区模式，各个存储节点之间通过高速互联网络进行通信，通过一致性算法即多版本并发控制（multi-version concurrency control，MVCC）保障数据执行过程中的集群状态正确性以及事务的ACID属性。但是当存储层分布式化之后，事务设计不同节点通信时，由于网络时延问题，与统一存储相比，事务整体执行过

程的时延势必有一定比例的增加。

最后根据新数据时代下对数据库的功能需求，本文简单从几个维度对上述3类数据库方案进行粗略的对比和总结，对比结果见**表1**。

4 分布式事务数据处理系统应用注意事项

分布式事务数据处理系统相比于集中式数据处理系统具有一系列优点，但是在成功应用之前，仍然需要注意几点问题。

（1）正确认识自身业务需求

分布式系统中的CAP决定了强一致性、高可用性、分区容错性三者不能同时实现。正确地认识自身业务需求，明确当前业务在上述3种特性中的重点，才能切合实际地对后续分布式事务数据处理系统功能做出取舍。而正确认识业务主要是服务规模和服务质量两方面。例如，对于金融行业中的核心业务，服务质量是必须满足的，所以必须要保证强一致性，从而弱化高可用性。而对于互联网系统中的查询业务，服务规模是主要需求，所以高可用性必须保障，而一致性可以相应地进行弱化。

（2）明确分布式事务数据处理系统各产品特点及适用场景

分布式事务数据处理系统因为底层技术路线不同，所以上层适用场景也不尽相同。例如，部分分布式事务数据处理系统需要在管理节点数量大于某个数量时，才能发挥较优性能，这就对应用方集群规模做出了限制；部分分布式事务数据处理系统需要使用特殊的硬件作为存储或计算介质，从而达到较优性能，这就对应用方硬件环境做出了限制。其他的类似区别指标有是否需要人工参与数据分区、适用的服务规模范围、是否具备多中心容灾能力、是否能够提供多种事务隔离级别等。

（3）正确规划已有IT架构向分布式体系迁移的路径及策略

目前现有数据库系统以集中式架构为主，且当前实际业务和第三方应用大部分也是基于集中式数据库结构实现的。如何在数据库架构从集中式向分布式转移过程

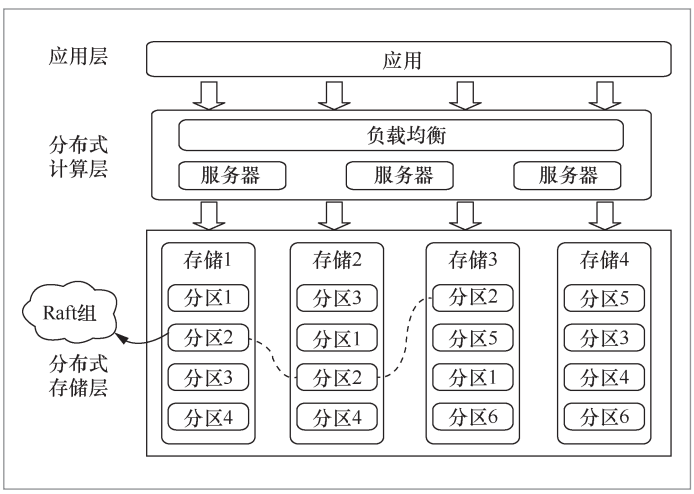

图4 计算存储均分布式扩展方案架构

表1 3类数据库方案对比

对比项	中间件方案	共享存储方案	计算存储均分布式方案
应用场景	OLTP业务为主	OLTP业务为主	OLAP/OLTP
扩展能力	受中间件能力限制	受存储设备能力限制	理论上无限制
高可用性	中间件存在单点故障	存储存在单点故障	理论上无单点故障
使用影响	增加业务规划能力要求	与集中式数据库相比无大变化	与集中式数据库相比无大变化
生态情况	方案稳定，应用案例丰富	成熟商业产品与创新开源产品并存	产品较少，但发展迅速
典型产品	Cobar、MyCat、TDDL	Aurora、Oracle RAC	Spanner、TiDB、OceanBase

中，保证上层业务的正常服务能力以及尽量少的原有业务体系变动，是需要格外注意的问题。为达到此目的，首先需要对转型后系统与原系统的兼容能力做一个充分评估，其次需要合理规划转型过程中存量业务和增量业务的关系，最后需要在新型系统正式上线前进行充分验证，做好原有系统灾备的充足准备。

（4）验证分布式事务数据处理系统产品能力

在正式迁移应用之前，需要正确筛选符合要求的分布式事务数据处理系统产品，同时验证各类分布式事务数据处理系统产品能力范围。笔者所在单位于2017年7月开始分布式事务数据处理系统标准化的尝试，梳理出分布式事务数据处理系统基准测试的方法论和具体实施方法。基准测试主要包括功能和性能两大方向，指标项包括兼容能力、透明能力、管理能力、分布式能力、事务能力、安全能力、高可用能力、场景化性能指标等多个维度，从而全面衡量分布式事务数据处理系统产品能力。

5 分布式事务数据处理系统未来趋势

近几年，以Google Spanner为代表的NewSQL类数据库发展迅速，其以全球范围内的扩展性、强事务支持能力、灵活的应用形式、廉价的软硬件成本等特点，逐渐成为数据库的最新演进形态，而这也是当前信息社会环境下的必然结果。究其原因，数据库的最初需求就是高性价比地实现数据的管理工作。首先高性价比可以从经济成本和业务成本两方面衡量，NewSQL类数据库实现了基于x86等廉价服务器搭建，并在底层实现数据自动化均衡处理，对业务侵入性不高，不需要应用过多感知，同时节约了经济成本和业务成本。其次，数据的管理工作分为存储和处理两方面。在大数据时代，数据的基本特征为数量大、类型多、价值密度低、速度快、时效高，所以对数据处理的需求为扩展性强、数据兼容能力强、具备分析能力、处理能力强，而NewSQL类数据库具备联机分析处理（online analytical processing，OLAP）/联机事务处理过程（on-line transaction processing，OLTP）混合能力、存储与计算强扩展性、兼容结构化与非结构化数据等多个特点，切合了当前时代数据管理功能需求。

随着大数据、物联网、人工智能等应用的快速兴起，分布式架构取代集中式架构是大势所趋，未来分布式事务数据处理系统将会作为数据时代的技术结构，为未来技术浪潮的发展提供更多的数据基础能力，未来分布式事务数据处理系统将在几个方面继续延伸变化。

首先，分布式事务数据处理系统的功能形态将会更加丰富。随着大数据和人工智能产业的发展，各数据应用企业都利用自身快速增长的业务数据，扩展出核心业务能力之外的数据分析能力，这就需要底层数据处理过程除了满足事务型逻辑之外，分析功能需求占到越来越多的比例。数据库作为数据底层服务设施，保障上层业务事务特性的同时提供高可用的分析能力是未来的发展趋势。目前已有将底层存储结构对接上层两套逻辑系统，完成上述功能的实例。未来如何平衡事务和分析两种业务的支撑占比以及相应技术实现方式，仍然需要继续探索与实践。

其次，新型硬件之下的分布式事务数据处理系统形态得到发展。底层存储与计算硬件（如3D内存、Flash卡、硬件压缩卡、非易失性存储器等）的快速发展，是原

有数据库系统发展过程未曾预想到的。当前数据库系统的底层读写模式仍是以原有磁盘存储介质特点进行设计并发展起来的，例如事务日志刷盘模式、缓冲区大小计算方法、I/O与网络的性能瓶颈评估等，这些传统设计方法造成了新型硬件介质的优势特点无法被现有数据库系统充分应用。未来的分布式事务数据处理系统需要针对新型硬件特性进行相关研发，从而突破传统数据库的思维及认知局限。

最后，利用协调硬件突破软件困境。例如因为物理时钟难以同步，所以传统事务数据库通常采用逻辑时钟维护事务顺序，但是逻辑时钟的维护会限制数据库的规模大小，所以覆盖全球范围的事务数据库几乎不可能实现。而Google公司推出的Spanner数据库直接采用硬件GPS加上原子钟方案，解决了物理时钟同步问题，做到了全球范围的数据库事务一致性。预计未来将会出现更多协调硬件辅助数据库软件系统解决之前无法解决的问题。

6 结束语

IT架构作为企业架构的基础，支撑着上层业务架构的建设与发展，促进了顶层愿景和战略的顺利实施。而传统集中式事务数据处理系统面对新环境下OLTP业务的迅速增长，相关问题逐渐暴露，故需要完成集中式体系向分布式体系的转型，从而释放数据红利，充分发展数字经济。而当前分布式事务数据处理系统正处于快速创新发展时期，多种体系架构并存，在扩展性、性能、可用性等方面优于传统集中式数据处理系统，但稳定性、服务能力等方面仍与传统集中式数据处理系统存在差距。故在转型过程中，应用方需要充分掌握相关技术特点、应用方法、实现方式等知识体系，从而实现快速、平稳的转型进程。

参考文献:

[1] CORBETT J C, DEAN J, EPSTEIN M, et al. Spanner: Google's globally-distributed database[C]//USENIX Conference on Operating Systems Design and Implementation, October 8-10, 2012, Hollywood, USA. Berkeley: USENIX Association, 2012: 251-264.

[2] CHEUNG D W, NG V T, FU A W, et al. Efficient mining of association rules in distributed databases[J]. IEEE Transactions on Knowledge & Data Engineering, 1996, 8(6): 911-922.

[3] ÖZSU M T, VALDURIEZ P. Principles of distributed database systems[M]. Berlin: Springer, 1999, 24(1): 31-41.

[4] BERNSTEIN P A. Concurrency control in distributed database systems[J]. ACM Computing Surveys, 1989, 13(2): 185-221.

[5] PAIVA J, RUIVO P, Romano P, et al. AutoPlacer: scalable self-tuning data placement in distributed key-value stores[J]. ACM Transactions on Autonomous & Adaptive Systems, 2014, 9(4):19.

[6] KUMAR R, GUPTA N, CHARU S, et al. Manage big data through NewSQL[C]// National Conference on Innovation in Wireless Communication and Networking Technology, April 6-9, 2014, Istanbul, Turkey. Piscataway: IEEE Press, 2014: 1-6.

[7] KUMAR R, PARASHAR B B, GUPTA S, et al. Apache Hadoop, NoSQL and NewSQL solutions of big data[J]. International Journal of Advance Foundation and Research in Science & Engineering, 2014, 1(6): 28-36.

[8] ARDEKANI M S, TERRY D B. A self-configurable geo-replicated cloud

storage system[C]//USENIX Conference on Operating Systems Design and Implementation, October 6-8, 2014, Broomfield, USA. Berkeley: USENIX Association, 2015: 367-381.
[9] LING G U, WENG H X, HUA-LIANG H U, et al. OceanBase schema design for OLAP application[J]. Journal of East China Normal University, 2014(5): 301-310.
[10] RAMANATHAN S, GOEL S, ALAGUMALAI S. Comparison of cloud database: Amazon's simpleDB and Google's bigtable[C]// International Conference on Recent Trends in Information Systems, June 3-5, 2011, Chennai, India. Piscataway: IEEE Press, 2011: 165-168.
[11] SUKHWANI B, MIN H, THOENNES M, et al. Database analytics acceleration using FPGAs[C]// International Conference on Parallel Architectures and Compilation Techniques, September 9-13, 2017, Portland, USA. Piscataway: IEEE Press, 2017: 411-420.
[12] ABBAS Q, SHAFIQ H, AHMAD I, et al. Concurrency control in distributed database system[C]// International Conference on Computer Communication and Informatics, January 7-9, 2016, Coimbatore, India. Piscataway: IEEE Press, 2016: 1-4.

作者简介

马鹏玮(1993-),男,中国信息通信研究院云计算与大数据研究所大数据研发与测试工程师,主要研究方向为大数据技术体系以及数据库技术体系。

魏凯(1981-),男,中国信息通信研究院云计算与大数据研究所部门主任,中国信息通信研究院互联网领域副主席,国际电信联盟分布式账本焦点组(ITU-T FG DLT)副主席,ITU-T SG16 Q21报告人,中国通信标准化协会TC1大数据与区块链工作组组长。牵头完成数据中心联盟大数据基准测试、可信区块链评测等标准和评测体系。主要研究方向为大数据和区块链相关技术与标准。

姜春宇(1987-),男,中国信息通信研究院移动互联网与大数据部工程师,数据中心联盟大数据技术与产品工作组组长。主要研究方向为大数据技术体系,从事大数据技术和产品的标准化工作,牵头建立了国内首个大数据产品能力评测体系。参与起草《促进国家大数据发展行动纲要》,参与编写工业和信息化部《大数据产业“十三五”发展规划》。

收稿日期: 2018-02-01