跨空间域数据管理分布式共识算法：现状、挑战和展望

李伟明[1]，李彤[1,2]，张大方[1]，戴隆超[1,2]，柴云鹏[1,2]
1. 中国人民大学信息学院，北京 100872；
2. 数据工程与知识工程教育部重点实验室，北京 100872

摘要
随着数字经济的飞速发展，以及“全国一体化数据中心”和“东数西算”等基础设施的不断完善，数据要素流通的大趋势使数据服务逐步由面向单一空间域的数据管理转变为面向跨空间域的数据管理。跨域数据管理需要通过分布式共识算法使数据一致。然而，已有的分布式共识算法仅考虑单数据中心的情况，没有考虑跨数据中心之间的网络通信的不确定性，从而在跨空间域场景下面临日志同步时延大、系统吞吐量低下等问题。系统地梳理了跨空间域下的分布式共识算法的现状以及面临的新挑战，并针对解决这些挑战的技术路线进行了展望。

关键词
跨空间域数据管理；分布式共识算法；日志复制；领导者选举

中图分类号：TP319 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023040

Distributed consensus algorithms for cross-domain data management: state-of-the-art, challenges and perspectives

LI Weiming[1], LI Tong[1,2], ZHANG Dafang[1], DAI Longchao[1,2], CHAI Yunpeng[1,2]
1. School of Information, Renmin University of China, Beijing 100872, China
2. Key Laboratory of Data Engineering and Knowledge Engineering, Beijing 100872, China

Abstract
With the exponential growth of data and the company's cross-domain disaster recovery requirements, companies increasingly need to manage data across spatial domains. Cross-domain data management requires a distributed consensus algorithm to make the data consistent. However, the existing distributed consensus algorithms only consider the situation of a single data center, and do not consider the uncertainty of network communication between data centers, so they face long log synchronization delays and low system throughput in cross-space region scenarios and other issues. The current status and new challenges of distributed consensus algorithms in the cross-space domain were sorted out systematically, and the technical route to solve these challenges was looked forward.

Key words
cross-domain data management, distributed consensus algorithm, log replication, leader election

0 引言

分布式共识算法是分布式系统用于维护多个节点上的多个副本保持一致的算法。该算法能让多副本架构下的分布式系统对外表现为只有一个逻辑副本，同时保障系统的读写操作都满足原子性。进而，应用层就可以忽略分布式系统底层多个数据副本间的同步问题。除此之外，当分布式系统中少部分节点出现异常时，分布式共识算法还能保证整个系统依然能够像单机系统一样提供正确的服务，保证系统的高可用性。

分布式共识算法包含两个核心模块：日志复制（log replication）与领导者选举（leader election）。日志复制指系统中的节点接收到数据后会将数据复制至一个或多个其他节点。通过日志复制可以实现数据的备份和系统的高可用。领导者（leader）是系统中的一个特殊节点，由所有节点选举产生。如果某个节点获得大多数节点的投票，它即可当选为领导者。领导者负责与客户端通信和协调其他节点上数据副本的复制与同步。引入领导者机制可以简化算法，使算法在工程上更易实现。

2020年年末，国家正式发布了《关于加快构建全国一体化大数据中心协同创新体系的指导意见》，该意见指出：“到2025年，全国范围内数据中心形成布局合理、绿色集约的基础设施一体化格局。”2021年发布的《全国一体化大数据中心协同创新体系算力枢纽实施方案》进一步强调加快实施“东数西算”工程，提高跨区域算力调度水平。这一系列重大举措，为数字世界的数据跨域共享与协同提供了重要的基础设施，为跨域数据管理提供了基本条件[1]。数据管理正从单一数据中心管理转为跨域的共享与协同管理。因此，分布式数据库也将越来越多地在跨空间域下的场景中使用与部署。

与在单数据中心内部部署不同，跨空间域部署分布式数据库引入了很多新的挑战。首先，跨空间域网络时延绝对值增大，这将导致算法进行日志复制时产生大量的网络开销。当分布式系统的节点处于同一数据中心时，节点之间的网络时延较低，复制操作的耗时不会对系统性能表现产生较大的影响。然而，当物理位置位于不同洲际的多个数据中心之间复制数据时，每次跨数据中心的通信时延都会高达数十甚至几百毫秒，这是单数据中心内部网络通信时延的数十甚至上百倍。其次，跨空间域节点间网络时延差异性不可忽略。在跨空间域背景下，不同节点部署的位置距离不同，不同节点间的网络时延也不同。因此按照之前的共识算法设计思路随机选举领导者节点可能使算法变得低效。最后，跨空间域网络时延存在动态变化。跨空间域的网络并非如同单数据中心内的网络一样稳定，其网络通信时延不定，例如存在网络的波动现象。这也将对共识算法的日志复制和领导者选举造成新的挑战。

针对上述分布式共识算法在跨空间域下面临的新挑战，可以从算法的日志复制和领导者选举两个方面进行优化。为了优化日志复制，可以采用及时发送数据和分时发送数据等方法。在领导者选举方面，可以通过选举最优领导者和领导者主动禅让等策略来进行优化。这些优化措施可以提高分布式共识算法在跨空间域环境中的性能和可靠性。

1 分布式共识算法简介

分布式共识算法的目标是使一组分布式的进程达成共识，即所有节点在某个时

间点都认为该值是最终一致的。这个过程通常涉及消息传递、数据同步、节点状态更新等操作。分布式共识算法的实现对于构建可靠、安全和高性能的分布式系统至关重要。在分布式共识算法的组织协调下，分布式系统就可以为上层应用提供单一逻辑副本抽象。

分布式共识算法分为拜占庭类分布式共识算法和非拜占庭类分布式共识算法。拜占庭类分布式共识算法需要考虑系统中可能存在的恶意行为，并对这些行为进行适当的容错处理。针对跨空间域下的拜占庭类分布式共识算法已经有一些研究，如Zamani等人[2]首次提出了基于分片的具有拜占庭容错能力的公有区块链协议RapidChain；Amiri等人[3]提出一种针对边缘计算网络优化的许可链（permissioned blockchain）系统Saguaro，通过充分利用边缘计算网络的层级性结构特点来减少广域通信的开销；Amiri等人[4]提出的Ziziphus将具有拜占庭容错能力的服务器划分为若干个拜占庭容错域，每个域用于处理邻近客户端产生的事务请求，并减少跨域的事务数量。

非拜占庭共识算法假设系统中的节点不会出现恶意行为，而这些节点只能由系统的拥有者或运营者进行操作。需要注意的是，本文涉及的所有共识算法均为非拜占庭类分布式共识算法。

分布式共识算法通过投票使多个节点对某个值达成一致。在分布式共识算法中，一组独立运行的进程或者节点之间进行通信投票，一旦投票的结果被超过一半的节点或者进程同意，那么就可以认为完成了共识过程。其中又涉及节点掉线、节点故障、消息重发等问题。目前主流的分布式共识算法有Paxos/Multi-Paxos[5-7]、ZAB[8]、Raft[9]等算法，这些算法都已经有了成熟的工业实现。后文主要从共识算法中的日志复制和领导者选举方面对算法进行介绍。

1.1 Paxos/Multi-Paxos算法

1990年，Lamport提出了Paxos算法，并给出了在任意情况下都能保证一致性的完备性证明，该方法是目前可有效解决分布式共识的算法之一。然而由于Paxos难以理解并且难以在工程上实现，因而Lamport等人对Paxos算法做出改进，提出了Multi-Paxos算法。与Paxos不同的是，在Multi-Paxos算法中存在一个领导者进程或节点，所有的请求都由领导者处理和转发。Multi-Paxos算法的主要内容为日志复制和领导者选举。

1.1.1 日志复制

Multi-Paxos的日志复制流程如**图1**所示。客户端将请求发送给算法中的领导者，领导者在本地形成提案，并将提案发送给跟随者（follower）。具体地，领导者向其他副本发送确认接收（accept）消息，

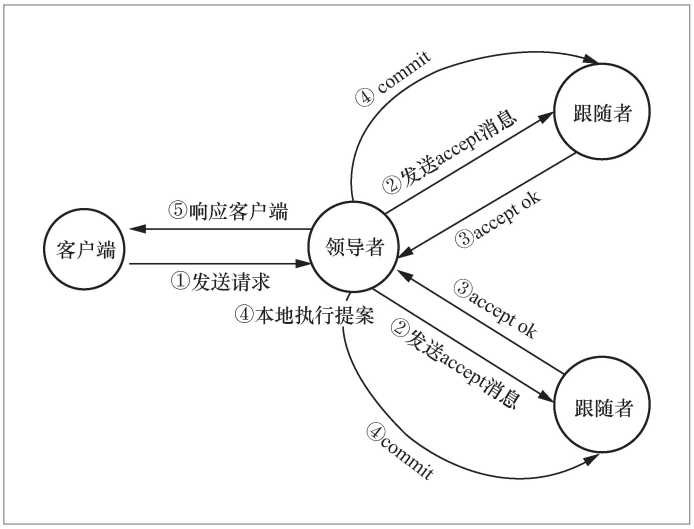

图1 Multi-Paxos 算法

accept消息中包含提案编号和提案内容。其他副本收到提案后，若此前并没有响应过更大提案编号的提案，那么副本接收这个提案并返回回复消息。领导者收到半数以上副本的响应成功消息后，在本地执行提案并向客户端发送成功回复消息，之后向其他副本发送提交日志（commit）消息，使其他副本在本地执行提案。

1.1.2 领导者选举

在Multi-Paxos算法的选举过程中，每个副本会产生一个任期编号（term），表示当前选举周期的序号。副本会向其他副本广播该任期编号，并等待它们的回复。如果有超过一半的副本回复同意，那么该副本将成为新的领导者，并开始执行相应的任务。同时，领导者选举的触发通过超时机制来实现，即如果副本在一定时间内没有收到其他副本的回复，就会发起新一轮的选举过程。

1.2 ZAB算法

ZAB算法是由雅虎提出的分布式共识算法。ZAB算法与Multi-Paxos算法类似，主要分为日志复制和领导者选举两部分。

1.2.1 日志复制

与Multi-Paxos类似，ZAB算法中所有的写请求一律由领导者处理。如**图2**所示，领导者收到客户端请求后，将请求封装为一个事务（proposal），并且分配给事务一个编号。事务编号包括任期和任期内的事务编号。然后领导者将事务发给所有的副本。副本收到事务后，将事务写入本地磁盘，然后向领导者发送回复消息。如果有半数以上副本响应成功复制，那么领导者提交该事务，同时给所有副本发送消息提交该事务。

1.2.2 领导者选举

在分布式系统中，节点随时有可能宕机，当领导者宕机或者网络掉线后，ZAB算法需要重新选举出新的领导者。在ZAB算法中，进程具有3种状态：领导状态、追随状态、选举状态。在正常情况下，副本向领导者发送心跳消息，当领导者在一段时间内没有收到一半以上副本的心跳消息时，领导者转换自己的状态为选举状态。而当副本发现领导者进入选举状态后，副本也进入选举状态，将自己转换为候选者（candidate）角色。

候选者向其他副本发送请求投票消息，消息中包括副本进程号和副本进程本地的最后一个事务号。副本接收到候选者的投票消息后，如果候选者进程的事务号更大，那么副本进程给其投票。当有半数以上进程给同一个候选者投票时，候选者成为新的领导者。

1.3 Raft算法

Raft算法是2014年由斯坦福大学提出的分布式共识算法，旨在设计一个易于理解的分布式共识算法。Raft算法具有与Paxos算法相同的功能，即使在网络分区和节点故障的情况下，都可维护数据在多个副本上的一致性。自从Raft被提出以来，已经被TiDB[10]、Polardb[11]、CockroachDB[12]等众多实用的分布式系统广泛采用。Raft算法给系统中的每个节点赋予一个角色，共有领导者、跟随者和候选者3种角色。

领导者主要负责接收客户端命令，并将接收到的命令转发给跟随者。跟随者接

收并保存领导者发送的命令，并对领导者进行响应。候选者由跟随者转变而来，当系统中没有领导者的时候，会有跟随者将自身角色转变为候选者，并向所有其他节点发送投票请求信息，选举产生新领导者。

1.3.1 日志复制

Raft算法的日志复制流程如**图3**所示。一旦系统中的某个节点被选为领导者，它便开始接收客户端的请求。领导者会将这些请求作为日志条目添加到本地的日志中，然后并行向其他服务器发起远程过程调用（remote procedure call，RPC）方法以复制这些日志条目。只有当这些日志条目被成功复制到大多数服务器节点上后，领导者才会将这些日志条目应用到其状态机，并向客户端返回执行结果。如果跟随者发生故障、运行缓慢或出现丢包等问题，领导者会不断重试，直到所有的跟随者最终都复制了所有的日志条目。

在实际的日志复制过程中，为了提升性能，Raft算法支持批处理（batch）和流水线（pipeline）技术。具体来说，领导者不会立即将每个接收到的请求转发，而是会将多个请求组合成一个批次再发送给跟随者。此外，领导者也不会等待上一个批次的结果返回后才继续发送下一个批次，而是会连续发送多个批次。Raft算法的批处理技术结合了心跳机制，领导者会在发送心跳信号的同时将日志发送给其他跟随者。当领导者没有新的日志可发送时，会向其他跟随者发送空的心跳包。

1.3.2 领导者选举

在系统初始状态下，所有节点的角色均为跟随者。每个跟随者都拥有一个时钟，该时钟的值是随机生成的，用于表示

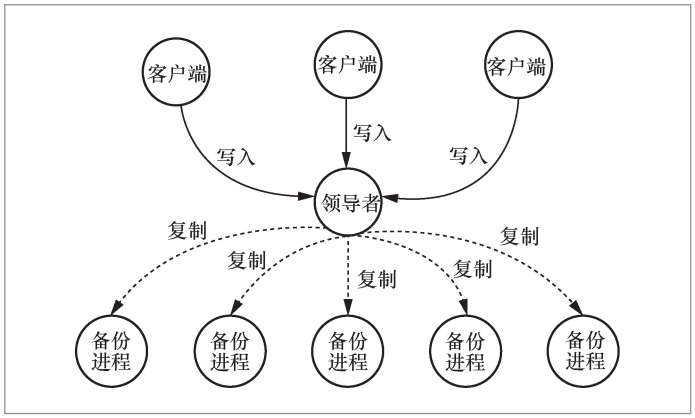

图 2 ZAB 算法

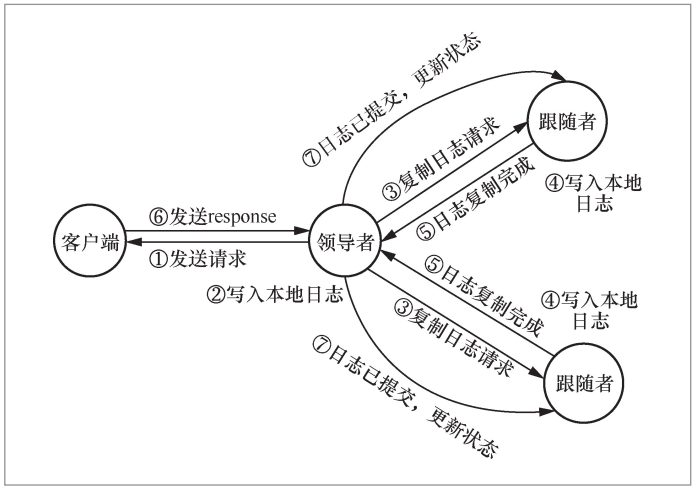

图 3 Raft 算法

跟随者成为领导者所需等待的时间。当某个跟随者的时钟倒计时结束时，该节点会启动领导者选举过程，并在赢得多数选票后成为领导者。随后，领导者会定期向跟随者发送心跳信号，以重置跟随者的时钟，并避免其再次启动选举过程。同时，当进行一次选举时，节点会将自身的任期号加1。任期号是每个节点在本地维护的一个变量，其设计灵感源自总统选举的任期。当跟随者转变为候选者并启动选举过程时，其会计算本地任期号，并将其递增1。

总结起来，Multi-Paxos算法、ZAB算法、Raft算法有很多相似之处，算法主

要内容都是日志复制和领导者选举。事实上，Multi-Paxos算法、ZAB算法和Raft算法都是类Paxos算法，其在Paxos的基础上做了更严格的限制以及更详细、更规范的描述。

2 跨空间域下的分布式共识算法面临的新挑战

在设计前述分布式共识算法时，仅考虑到分布式节点在同一数据中心内的情况。然而，单一数据中心的技术架构目前面临许多问题。首先是性能问题。随着业务量的增加，单一数据中心的基础设施已经无法应对业务量增长带来的挑战。其次是容错问题。如果因不可抗力（例如地震、爆炸等）导致单一数据中心故障，那么将导致业务和系统的不可用性，对公司的形象和收入造成严重的损失。同时，对于单个数据中心的架构，当用户需要跨地域访问时，会遇到较大的延迟，这将影响用户的体验。因此，许多公司需要跨地域部署分布式系统，分布式共识算法也需要在跨空间域的场景下工作。然而，在跨空间域的场景下，共识算法面临着新的问题和挑战。如**图4**所示，在跨国链路、跨省链路以及数据中心内部链路网络的情况下，网络时延的变化情况是不可忽略的。跨省链路的网络时延为几十毫秒，而跨国链路的网络时延则可能超过一百毫秒。同时，不同节点之间的网络时延存在差异性，且节点间的网络时延会出现动态变化。如**图4**所示，在跨国链路中，网络时延可能会突然增加至两百多毫秒，也可能会突然降低至一百多毫秒，这对共识算法的正确性和性能提出了新的挑战。总体来说，广域网的数据传输具有不确定性，这种不确定性体现在以下3个方面。

2.1 跨空间域网络时延绝对值增大

在单数据中心内，节点间的网络时延较低，通信时延通常为几毫秒。然而，在跨地域场景下，节点之间的网络时延将显著增大，一般为数百毫秒甚至数秒。一些实验测试表明，同机房或同地域机房间的网络时延通常在毫秒级别，而跨地域访问时延则上升了一个数量级。

网络时延的增加将对分布式共识算法的性能产生巨大影响。以Raft算法为例，在日志复制过程中，分布式系统中的节点或进程需要与其他节点进行大量数据复制和同步操作。当网络时延增加时，数据复制和同步的效率将受到极大的影响，从而导致分布式系统的性能下降。同时，Raft在进行领导者选举时，候选者需要向其他节点发送请求并获得其他节点的投票，网络时延的增加将严重影响领导者选举的性能，从而导致系统存在长时间的不可用问题。

2.2 跨空间域节点间网络时延差异性不可忽略

在跨空间域的背景下，不同节点间的网络情况是多种多样的。一些地理位置相近的节点之间的网络时延相对较低，而存在一些节点距离其他节点的地理位置很远，因此具有较大的网络时延。

在这种场景下，对于Multi-Paxos、ZAB、Raft等算法，由于领导者是影响系统性能的核心，当领导者距离其他节点很远时，进行日志复制并获得超过半数节点的响应将变得更加困难，算法性能将变得很差。但是，前述的分布式共识算法中并没有机制来阻止这一情况的发生。

2.3 跨空间域网络时延动态变化

在单数据中心内，节点间的网络情况不仅时延低，而且非常稳定。然而，在跨空间域场景下，节点间的网络状况存在一定的波动性。因此，在某些时段，节点间的网络时延可能会出现极端高峰值，而在另一些时段则表现正常。网络状况的大幅波动也会对分布式系统的性能产生影响。

举例来说，Multi-Paxos、ZAB、Raft等算法在系统初始化时，如果领导者节点的网络状态与其他节点的网络时延相对较低，则系统性能会非常好。但是，随着时间的推移，领导者节点的网络状况可能会恶化，与其他节点的网络时延变得很大，这将极大地降低系统的性能。这种情况在以前的分布式共识算法中未被考虑，因此需要新的机制来应对这种情况。同时，网络状况波动对Raft等算法数据复制过程的影响是显著的。在某些时候，网络时延会很高，网络拥塞，此时领导者节点向所有副本节点发送数据的速度较慢。而在其他时刻，网络较为空闲，网络时延较低，此时领导者节点向所有副本节点发送数据的速度较快。因此，可以充分利用网络波动的这一特性来优化数据复制过程。

综上所述，跨空间域网络时延绝对值增大、跨空间域网络时延差异性不可忽略、跨空间域网络时延动态变化将对分布式共识算法造成严重的影响。目前已经有一些针对跨空间域下的共识算法研究，本文第3节将介绍现有的跨空间域共识算法的研究进展。

3 跨空间域分布式共识算法研究进展

针对跨空间域分布式共识算法优化的技术路线有两种，一种是确定性网络技术，另一种是优化分布式共识算法以更适应广域网。

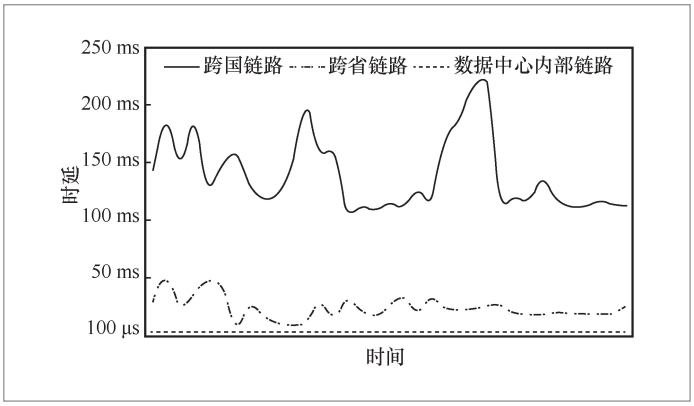

图 4 网络时延对比

3.1 确定性网络技术

广域网数据传输存在不确定性，这种不确定主要体现在跨空间域网络时延绝对值增大、跨空间域网络时延差异性不可忽略以及跨空间域网络时延动态变化3个方面。现在已经有一些确定性网络技术试图解决这些问题。确定性网络用于提供实时数据传输，保证确定的通信服务质量，如超低上界的时延、抖动、丢包率，上下界可控的带宽，以及超高下界的可靠性。确定性网络能够满足高质量通信需求。

最早的确定性网络技术为IEEE 802.1 TSN（time sensitive network）技术。TSN技术是IEEE基于OSI参考模型的数据链路层（L2）设计的相对成熟的确定性网络标准，产业界已推出了支持TSN的芯片、交换机和工业终端等。然而，TSN无法很好地适用于广域网。首先在开销方面，TSN技术需要进行逐流状态维护，这对于广域网传输而言，维护开销可能无法接受；其次在部署方面，TSN技术依赖于精准的时间同步，在跨域场景下的大规模、

远距离传输和复杂组网情况下，难以实现精准时间同步。因此，无法直接将TSN应用于广域网，实现广域网的确定性低时延。

TSC（time sensitive communication）[13]是3GPP在2020年7月发布的R16标准中引入的5G相关确定性网络技术。TSC将TSN的应用范围从有线扩展到了无线。具体地，TSC将5G系统作为一个TSN网桥集成在TSN系统中，通过网络切片、确定性转发、TSN管理协同及网络拓扑发现等能力，在固网覆盖困难或存在移动性要求的业务场景中辅助TSN，提供确定性网络传输服务。TSC是TSN的扩展，因此它同样无法直接应用于广域网。

DetNet（deterministic networking）[14]将确定性网络技术扩展到OSI参考模型的网络层（L3），通过资源分配、服务保护和显式路由等技术[15]，实现了确定性报文转发和路由，为实现跨域的确定性传输提供了技术基础。DetNet适用于受单一管理控制或封闭管理控制组内的网络，例如校园网络和专用广域网。针对公共广域网，DetNet面临着与TSN一样的开销大、部署难问题。

New IP[16]是华为提出的确定性网络技术，给出了确定性低时延的初步框架。New IP不仅包含L3的确定性IP技术，还对基于确定性IP的新传输层（L4）技术进行了初步定义。针对确定性IP技术，New IP引入异步的周期调度机制来严格避免微突发的存在，从而保证确定性低时延数据转发能力。但一方面，New IP需重新设计网络中间节点，规模部署阻力较大；另一方面，New IP只是一个初步的基础网络架构，仍然存在大量技术细节有待产业界共同完成。

综上所述，TSN只适用于局域网，TSC同样只适用于无线局域网络，DetNet只适用于校园网，NewIP的部署阻力较大。因此，目前已有的确定性网络技术很难在广域网上实现，解决确定性广域网数据传输这一任务仍然任重道远。与此同时，针对跨空间域分布式共识算法的优化更具有现实操作性和可部署性，本文之后将介绍目前针对跨空间域分布式共识协议的优化进展，并给出一些优化思路。

3.2 跨空间域分布式共识算法优化

现有的跨空间域分布式共识算法（如CURP[17]、EPaxos[18]）聚焦于解决上述的网络时延高的问题，试图减少分布式共识算法中的数据复制的通信次数，以提高系统性能。同时，针对系统的网络环境与硬件环境，以Raft-Plus[19]为主要代表的算法对算法的领导者以选举方式进行优化，选取网络更好、性能更好的节点作为领导者，以提高系统性能。DPaxos[20]则将用户所需数据分片分配到距离用户请求发出位置最近的数据中心，从而降低跨地域的数据访问时延。

CURP和EPaxos都是基于“大多数的操作是可交换的”这一假设提出的。CURP算法中增加了见证者（witnesses）角色，使算法能够减少数据通信次数。CURP将所有可交换的操作复制到见证者，见证者只保证数据的持久性而不对数据进行排序。在CURP算法中，客户端将每个操作请求复制到一个或多个见证者，同时将请求发送到领导者。领导者可以执行操作并返回到客户端，而无须等待数据复制到其他跟随者。这允许数据操作在一轮通信内完成，从而提高系统的性能。

与CURP算法不同，EPaxos是一种无领导者（leaderless）的分布式共识算法。所有的副本都可以从客户端接受请求，并且只需一轮通信即可提交请求，因此客户端可以选择较近的副本发送请求。为了使不同副本提出的提案不会相互冲突，EPaxos设计了一个二维矩阵的日志，每个副本在属于自己的一维数组内放置日志。

每个副本都维护这样一个矩阵。当操作无冲突时，EPaxos可以只进行一次副本间通信，否则需要进行两次。

Raft-Plus在选举时，跟随者并不将选票直接投给第一个到达的候选者，而是收集一段时间内的候选者的请求。跟随者与这些候选者进行测速后将票投给网络时延最低的候选者。同时候选者向跟随者发送请求时会携带一些参数，跟随者会将选票投给处理能力最强、当选领导者次数最多，以及在上一任期内收到最多客户端请求的节点。同时Raft-Plus引入一种反对票机制。如果跟随者发现领导者的网络时延超出阈值，则跟随者向领导者发送反对票。当领导者收到一半以上的反对票时，领导者角色切换为跟随者。

DPaxos（dynamic Paxos）是针对时延敏感性高的应用场景（如AR/VR等），基于Paxos的、应用于边缘计算系统的共识协议。该协议动态地将用户所需数据分片分配到距离用户请求发出位置最近的数据中心，从而降低了跨地域的数据访问的时延。DPaxos提出了区域中心仲裁群，使复制仲裁群小而且靠近用户。同时DPaxos启用了扩展仲裁群，使复制和领导者选举仲裁群都可以动态增长，并且在存在冲突时可以快速扩展。这些改进措施使DPaxos能够更好地管理数据，并且在实际部署中取得了显著的性能提升。

目前针对广域网的优化仍然不够完善，网络时延高、网络时延差异性显著、网络时延动态变化等问题仍然难以得到根本性解决。现有针对跨空间域的分布式共识算法，如CURP、EPaxos，非常依赖于操作的交换性假设，缺乏通用性。同时，Raft-Plus对Raft的领导者选举的优化也未经具体分析和方案论证。DPaxos适用于边缘计算系统，同样缺乏通用性。因此，尽管这些算法为解决跨空间域共识问题提供了一些解决方案，但仍然需要更通用和更有效的优化思路来提高跨空间域分布式共识算法的效率和可靠性。

4 跨空间域分布式共识算法研究展望

如前所述，现有的跨空间域分布式共识算法并不能很好地解决网络时延高、节点间网络时延差异大、网络波动等问题。本文从分布式共识算法的日志复制和领导者选举两方面对跨空间域下分布式共识算法的优化给出一些建议和思路。

如图5所示，对跨空间域共识算法的优化从日志复制和领导者选举两方面入手，可以通过尽早发送数据、分时发送数据等机制来优化日志复制，也可以通过选举最优领导者和领导者主动禅让等机制来优化领导者选举。

4.1 日志复制优化

在跨空间域背景下，因为跨空间域网络时延高，分布式共识算法中最受影响的便是日志复制模块。缩短日志复制时间的思路主要有两方面：一是减少数据传输的通信轮数；二是降低跨空间域网络时延。如此便可减少领导者和跟随者之间的通信

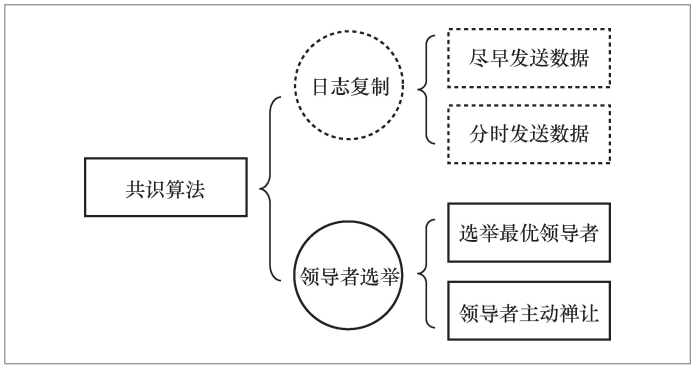

图 5 跨空间域共识算法优化

次数，缩短数据传输时间。本文提出了尽早发送数据和分时发送数据来缩短数据的请求响应时间。

4.1.1 尽早发送数据

一个缩短日志复制时间的思路是尽早发送数据。大多数共识算法采用心跳机制来触发数据的发送，其中心跳时间是一个固定的系统参数。然而，在跨空间域的场景下，领导者与其他节点之间的网络时延差异性较大。因此，领导者可以对每个节点设置不同的心跳间隔。

在跨空间域的共识算法中，领导者可以为网络时延较高的节点设计较小的心跳时间，这样当领导者有数据时，就可以尽快地将数据发送给远程节点，从而减少跨空间域数据请求的时间。这种方法可以有效地缩短日志复制时间，提高系统的性能和可靠性。

4.1.2 分时发送数据

由于跨空间域网络的波动性，网络时延在不同时间段会发生变化，因此领导者向副本发送数据可能成为系统性能的瓶颈。为了解决这个问题，可以采用削峰填谷的策略。在网络阻塞期间，领导者可以只向网络时延较低的一半以上的节点发送数据，而在网络空闲期间，则可以向之前未被发送数据的滞后节点发送最新的快照，使它们能够快速追赶到领导者的最新状态。相比于之前固定同量发送数据的算法，这样做大大减小了领导者在网络阻塞时期发送数据的压力，并减少了发送的数据量。

4.2 领导者选举优化

现有的分布式共识算法，如Multi-Paxos、Raft、CURP等，在设计时，系统中所有节点当选领导者的概率是相同的。然而，在跨空间域场景下，系统中的节点可能具有不同的软硬件和网络条件，因此应该在领导者选举时具有侧重性。

4.2.1 选举最优领导者

在跨空间域场景下，网络条件因节点间的地理位置不同而异。因此，系统中存在一些节点与大部分节点的通信时间很短，而另一些节点由于地理位置过远，与系统中其他节点的通信时延很高。在共识算法中，领导者与其他节点之间的网络时延是影响系统性能的关键因素。为了提高系统性能，共识算法应该设计一些机制，使与其他节点通信时间更短的节点更容易当选为领导者。

在分布式系统中，节点之间可以两两通信测量网络时延。一个节点与其他节点的网络时延的中位数可以代表该节点的网络情况，中位数越小，节点的网络情况越好。因此，可以采用让网络情况更好的节点优先发起选举的方式，使其成为领导者，从而提高系统的性能。

4.2.2 领导者主动禅让

由于网络时延的波动和变化，一个最初被选为领导者的节点，随着时间的推移，其网络情况可能会变差，已经不再适合担任领导者；或者系统中出现了网络条件明显优于当前领导者的节点。因此，需要实现领导者节点的动态切换，以维护系统的高性能。领导者可以监测每个节点与其他节点的网络时延情况。当领导者检测到某个跟随者的网络条件更优于其自身时，例如此节点与大多数节点的网络时延较小，且远小于当前领导者与大多数节点的网络时

延，此时由领导者指定此节点发起选举竞选新的领导者。此外，领导者的切换需要耗费一定的时间成本，因此需要在切换频率与切换时间成本之间做出权衡，以使领导者的切换不会太频繁。

为了更清晰地理解上述的优化思路，**表1**对上述优化思路做了总结与分析。对共识算法的优化思路仍然从日志复制和领导者选举两个方向展开。在日志复制方向，尽早发送数据可以缩短跨空间域请求的耗时，但需要实时监控集群的状态，选择合适的心跳时间。分时发送数据可以减小领导者在网络阻塞时期发送数据的压力，但这需要对网络状况有比较准确的预测。在领导者选举方向，选举最优领导者和领导者主动禅让都可使网络最好的节点当选领导者，提高算法性能，但系统的选举和切换过程将产生额外的开销，需要进一步探索进行选举切换的时机。

5 结束语

网络时延绝对值增大、节点间网络时延差异性不可忽略和网络时延动态变化，使跨空间域分布式共识协议的设计面临新的挑战。本文针对分布式共识协议中核心的日志复制和领导者选举两个模块，提出了跨空间域分布式共识算法的设计思路，可以为跨空间域数据管理领域的研究工作提供参考。

参考文献:

[1] 柴云鹏，李彤，范举，等. 跨域数据管理的内涵与挑战[J]. 中国计算机学会通讯，2022，18(11)：29-33.
CHAI Y P，LI T，FAN J，et al. Cross-domain data management：connotation and challenges[J]. Communications of China Computer Federation，2022，18(11)：29-33.

[2] ZAMANI M，MOVAHEDI M，RAYKOVA M. RapidChain：scaling blockchain via full sharding[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York：ACM Press，2018：931-948.

[3] AMIRI M J，LAI Z L，PATEL L，et al. Saguaro：efficient processing of transactions in wide area networks using a hierarchical permissioned blockchain[J]. arXiv preprint，2021，arXiv:2101.08819.

[4] AMIRI M J，SHU D，MAIYYA S，et al. Ziziphus：scalable data management across byzantine edge severs[C]//Proceedings of the 2023 IEEE 38th International Conference on Data Engineering. Piscataway：IEEE Press，2023.

表1 跨空间域分布式共识算法优化思路比较

优化方向	解决方案	优点	缺点
日志复制	尽早发送数据	缩短跨空间域请求的耗时	需要实时监控集群状态，选择合适的心跳时间
	分时发送数据	减小领导者在网络阻塞时期发送数据的压力	对网络状况预测准确度要求高
领导者选举	选举最优领导者	能够通过网络状态实时选取性能最优的领导者节点	视算法复杂度不同，系统的选举过程将产生额外的开销
	领导者主动禅让	当领导者节点网络状况发生变化时，能够及时切换到当前最优的领导者节点	

[5] CHANDRA T D, GRIESEMER R, REDSTONE J. Paxos made live: an engineering perspective[C]//Proceedings of the 26th Annual ACM Symposium on Principles of Distributed Computing. New York: ACM Press, 2007: 398–407.

[6] LAMPORT L. Paxos made simple[J]. ACM Sigact News, 2001, 32(4): 18–25.

[7] LAMPORT L. The part-time parliament[J]. ACM Transactions on Computer Systems, 1998, 16(2): 133–169.

[8] JUNQUEIRA F P, REED B C, SERAFINI M. Zab: high-performance broadcast for primary-backup systems[C]//Proceedings of 2011 IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN). Piscataway: IEEE Press, 2011: 245–256.

[9] ONGARO D, OUSTERHOUT J. In search of an understandable consensus algorithm[C]//Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference. New York: ACM Press, 2014: 305–320.

[10] HUANG D X, LIU Q, CUI Q, et al. TiDB: a raft-based HTAP database[J]. Proceedings of the VLDB Endowment, 2020, 13(12): 3072–3084.

[11] CAO W, ZHANG Y, YANG X, et al. PolarDB serverless: a cloud native database for disaggregated data centers[C]//Proceedings of the 2021 International Conference on Management of Data. New York: ACM Press, 2021: 2477–2489.

[12] TAFT R, SHARIF I, MATEI A, et al. CockroachDB: the resilient geo-distributed SQL database[C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2020: 1493–1509.

[13] JUN S M, KANG Y, KIM J, et al. Ultra-low-latency services in 5G systems: a perspective from 3GPP standards[J]. ETRI Journal, 2020, 42(5): 721–733.

[14] FINN N, THUBERT P, VARGA B, et al. Deterministic networking architecture[J]. RFC, 2019, 8655: 1–38.

[15] SONG F, LI L T, YOU I, et al. Enabling heterogeneous deterministic networks with smart collaborative theory[J]. IEEE Network, 2021, 35(3): 64–71.

[16] 郑秀丽, 蒋胜, 王闯. NewIP: 开拓未来数据网络的新连接和新能力[J]. 电信科学, 2019, 35(9): 2–11.
ZHENG X L, JIANG S, WANG C. NewIP: new connectivity and capabilities of upgrading future data network[J]. Telecommunications Science, 2019, 35(9): 2–11.

[17] PARK S J, OUSTERHOUT J. Exploiting commutativity for practical fast replication[J]. arXiv preprint, 2017, arXiv: 1710.09921.

[18] MORARU I, ANDERSEN D G, KAMINSKY M. There is more consensus in Egalitarian parliaments[C]//Proceedings of the 24th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2013: 358–372.

[19] XU J J, WANG W, ZENG Y, et al. Raft-PLUS: improving raft by multi-policy based leader election with unprejudiced sorting[J]. Symmetry, 2022, 14(6): 1122.

[20] NAWAB F, AGRAWAL D, EL ABBADI A. DPaxos: managing data closer to users for low-latency and mobile applications[C]//Proceedings of the 2018 International Conference on Management of Data. New York: ACM Press, 2018: 1221–1236.

作者简介

李伟明（1999– ），男，中国人民大学信息学院硕士生，主要研究方向为分布式共识协议。

李彤（1989– ），男，博士，中国人民大学信息学院副教授，主要研究方向为新一代互联网体系结构、跨域数据管理和大数据。

张大方（1998– ），男，中国人民大学信息学院硕士生，主要研究方向为分布式共识协议。

戴隆超（1996– ）男，中国人民大学信息学院硕士生，主要研究方向为跨域数据管理和大数据。

柴云鹏（1983– ），男，博士，中国人民大学信息学院教授、博士生导师，中国人民大学理工学科建设处副处长、计算机科学与技术系主任，主要研究方向为数据库管理系统、存储系统、云计算。

收稿日期： 2023-02-28

通信作者： 李彤，tong.li@ruc.edu.cn

基金项目： 国家自然科学基金资助项目（No.61972402，No.61972275，No.62202473）；中国人民大学建设世界一流大学（学科）基金资助项目

Foundation Items: The National Natural Science Foundation of China (No.61972402, No.61972275, No.62202473), Fund for Building World-Class Universities (Disciplines) of Renmin University of China