

基于 TS-MLG 的网络状态融合调度方法

林霄¹, 张佳¹, 岳胜男², 孙卫强², 胡卫生²

(1. 福州大学 物理与信息工程学院, 福州 350116;

2. 上海交通大学 区域光纤通信网与新型光通信系统国家重点实验室, 上海 200240)

摘要: 现有存储转发(SnF)调度方法的计算复杂度、状态维护难度大, 严重制约了其在大数据传输领域的应用, 且传统全局建模导致调度问题中出现大量冗余状态。针对此问题, 提出了一种基于时移多层图(TS-MLG)的网络状态融合调度方法。该方法通过对预选路径的状态融合, 减少了冗余状态, 提高了算法的调度效率。仿真结果表明: 相比传统调度方法, 该方法的计算时间短、调度性能好, 且状态维护数量少, 能够为大规模网络提供实时、高效的调度服务。

关键词: 大数据传输; 数据中心间互联网络; 波长路由; 存储; 调度方法

中图分类号: TN915.41 文献标志码: A 文章编号: 1002-5561(2020)09-0052-06

DOI: 10.13921/j.cnki.issn1002-5561.2020.09.013

开放科学(资源服务)标识码(OSID):



Network state fusion scheduling method based on TS-MLG

LIN Xiao¹, ZHANG Jia¹, YUE Shengnan², SUN Weiqiang², HU Weisheng²

(1. College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China;

2. State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: The existing store and forward (SnF) scheduling methods suffer from high computational complexity and heavy state maintenance, which seriously restricts its application in the field of big data transfer. Moreover, the traditional global modeling results in a lot of redundant states in the scheduling problem. Aiming at this problem, this paper proposes a state fusion scheduling method based on time-shifting multi-layer graph (TS-MLG). The method reduces redundant states and improves the scheduling efficiency of the algorithm by fusing the states of preselected paths. The simulation results show that compared with the traditional scheduling method, the proposed method has the advantages of short computing time, good scheduling performance and less state maintenance, which suggests that it can provide real-time and efficient scheduling services for large-scale networks.

Key words: big data transmission; inter-data center networks; wavelength routing; storage; scheduling method

0 引言

新兴云服务的飞速发展催生了海量的数据中心间大数据传输需求^[1]。然而, 数据中心间网络(inter-DCNs)中带宽的使用呈现典型时空不均衡, 使得传统

端到端(E2E)传输难以满足大数据对大带宽、高时长的传输需求^[2]。数据中心(DC)备份、数据迁移等大数据传输应用通常不需要数据传输立即开始, 数据可暂时缓存于中继 DC 节点, 等到后续网络相对空闲或有足够的带宽时再继续传输。这常被称为存储转发(SnF), 广泛应用于 inter-DCNs, 以提高带宽资源的利用率和网络的大数据传输效能^[3-8]。

然而, 存储的引入使得传统的路由问题变为一个同时包括空间和时间 2 个维度的调度问题。求解该问题, 不仅需要考虑带宽和存储资源的分配, 还需要同时寻找空间路由、规划时间调度^[9]。多数文献将调度问题建模为优化问题(如网络流问题^[3-4]、线性规划问题^[5-6]等), 简称优化法。在建模过程中, 优化法将全局的网络状态信息转变为约束函数, 利用经典优化算法, 获

收稿日期: 2020-04-05。

基金项目: 国家自然科学基金青年科学基金项目(61901118)资助; 国家自然科学基金重点项目(61433009)资助; “上海交通大学区域光纤通信网与新型光通信系统国家重点实验室”开放基金(2019GZKF03003)资助。

作者简介: 林霄(1988—), 男, 博士, 助理教授, 现就职于福州大学物理与信息工程学院, 主要从事大数据网络传输、智能光网络及边缘计算方面的研究, 主持或参与新型光网络、光传输系统领域的国家级、省部级和国际项目 8 项。



得路由、调度和资源配置的最优解。优化法被证明能够有效求解小规模网络或静态网络负载下的调度问题。但是, 随着网络规模、存储节点数量和传输请求数量的增加, 优化问题的规模呈现指数增长。因此, 优化法难以解决大规模网络和动态网络负载下的调度问题。为了减小调度问题规模, 文献[7]提出了时空解耦的调度(TSD)方法, 该方法利用时移多层图(TS-MLG)^[8]将时空二维的调度问题建模为路由问题, 简化了调度问题的求解过程。在此基础上, TSD 方法引入固定-替代路由(FAR), 仅使用预选路由上的状态信息建模调度问题, 由此减小了调度问题的规模, 降低了问题求解难度。TSD 方法使用备选路由上链路的双向状态信息建模。但是, 引入双向状态信息可能导致数据在节点之间的传输形成环路(loop), 从而降低带宽的使用效率^[9]。此外, 优化法和 TSD 方法都集中维护网络的全局状态信息, 这意味着任意时刻某链路状态发生改变, 即使其它链路的状态保持不变, 也需要被重复记录。随着网络规模扩大, 这种状态信息维护方式将造成大量冗余状态, 增加状态信息的维护成本和维护难度。

为了解决上述问题, 本文以光电路交换(OCS)与 DC 存储的结合为研究场景, 量化分析冗余状态对调度性能和复杂度的影响, 提出一种基于 TS-MLG 的网络状态融合调度方法(简称 SFS 方法)。

1 冗余网络状态分析

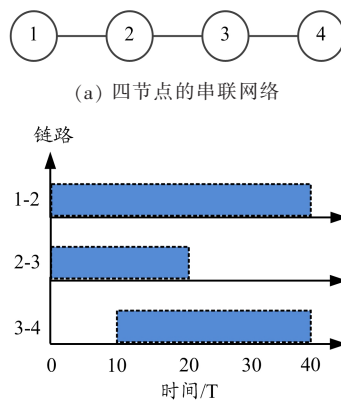
文献[8]提出一种路由框架, 即 TS-MLG。借助 TS-MLG, SnF 调度问题可以被转换为路由问题, 从而降低问题建模与求解的难度。本文将利用 TS-MLG 构建求解调度问题所需的网络状态信息, 从而揭示导致调度问题规模激增的原因。

1.1 冗余状态的产生

本文首先举例说明 SnF 调度问题的规模是如何随着网络规模改变的。例如, 一个四节点三链路的串联网, 如图 1(a)所示。假设每条链路上只有 1 个波长, 每条链路的波长资源占用情况如图 1(b)所示。当 SnF 调度问题包含不同链路状态信息时, 其构建的不同规模 TS-MLG 如图 2 所示。TS-MLG 引入时间链路和空间链路的概念。请求穿越时间链路、空间链路分别表示数据缓存、数据传输。通过对 TS-MLG 进行最短路计算, 即可获得一条“穿越时空的端到端”路径, 同时实现空间路由与时间调度的融合调度、带宽与存储资源的联合分配, 极大简化了 SnF 调度问题的求解过程。在 TS-MLG 中, 一条路径可能穿越多条时

空链路, 这与传统网络中的路径含义不同。本文使用“路由”代指传统网络中的路径。因此, 路由仅在空间上表明请求如何从源节点到达目的节点, 并未考虑时间调度。

当 SnF 调度问题只考虑链路 1-2 的状态信息时, 所构建 TS-MLG 如图 2(a)所示。该 2 层 TS-MLG 表明链路 1-2



(b) 每条链路上波长资源占用情况
图 1 四节点串联网及其链路波长资源占用情况

上的波长在 $t=0$ T (T 为任意时间单位) 时被占用, 在 $t=40$ T 时被释放。当 SnF 调度问题同时考虑链路 1-2 和链路 2-3 的状态信息时, 所构建 TS-MLG 如图 2(b)所示。位于 $t=20$ T 的层表明链路 2-3 上的波长一直被占用直到 $t=20$ T 为止。然而, 对于链路 1-2, 位于 $t=0$ T 和 $t=20$ T 时刻的层代表状态信息相同。同样, 对于链路 2-3, 位于 $t=20$ T 和 $t=40$ T 层的状态信息也相同。当 SnF 调度问题考虑 3 条链路时, 所构建的 TS-MLG 进一步扩展如图 2(c)所示。由于链路 3-4 上的波长在 $t=10 \sim 40$ T 期间被占用, 因此 TS-MLG 在 $t=10$ T 时刻添加了新层。然而, 对于链路 1-2 和链路 2-3, $t=10$ T 时刻的新层不能提供任何新的状态信息。

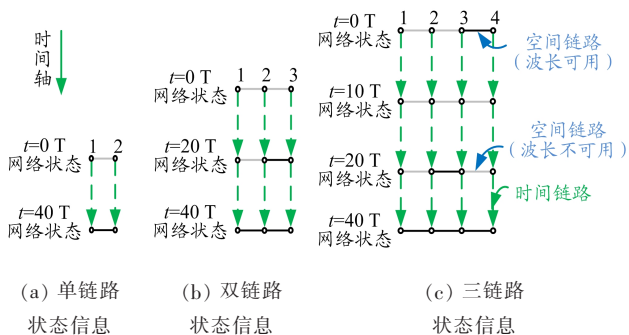


图 2 当 SnF 调度问题包含不同数量链路状态信息时, 构建的不同规模 TS-MLG

显然, 由于每条链路上的网络状态变化的时间不一致, 某时刻任意一条链路的状态发生改变 (即使其它链路的状态不变), 也需要被重复记录。SnF 调度问题考虑的链路数量越多, 单链路状态上信息被不断复制的可能性就越大, 最终造成 SnF 调度问题的规模随着网络规模的扩展而迅速增大。对于从节点 1 到节点 4 的大数据传输请求, 图 2(c)中 TS-MLG 提供了网络

的完整信息, 有助于获得最优调度解决方案。但是, 对于从节点 1 到节点 2 的请求, $t=10T$ 和 $t=20T$ 层是无用的。此时, 重复的状态不仅不能提供有用的信息, 还导致问题规模的扩大和求解难度增大。这些重复的状态就称为冗余状态。

1.2 冗余状态的影响

本文采用文献[7]提出的模型来分析冗余状态数量对调度性能和计算复杂度的影响。假设固定路由 $R=\{s, i_1, \dots, i_{N-2}, d\}$, 其中 s 表示源节点, d 表示目的节点, N 表示固定路由上节点总数 ($N \geq 2$), $i_1 \sim i_{N-2}$ 表示路由 R 途径的中继节点。固定路由

SnF 调度模型如图 3 所示。假设 TS-MLG 有 L 层, $d^{(l)}$ 表示位于第 l 层的目的节点 ($l \in [1, L]$)。对于从 s 传送数据

到 d 的请求 r , 只需对该模型进行路由搜索, 找到一条 s 传送到 $d^{(l)}$ 的有效路径, 即该路径的时空链路均有足够资源, 则 r 可以被成功传输。

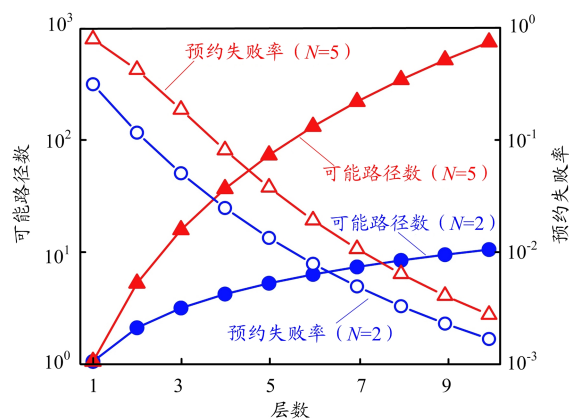
$P_{(s,d)}(N, L)$ 定义为从 s 到 $d^{(L)}$ 的所有可能路径的总数 ($L=[1, \dots, L]$)。TS-MLG 的规模越大, s 与 d 之间存在的可能路径就越多, r 成功传输的可能性越大。然而, 搜索如此庞大的可能路径集合意味着高昂的计算成本。因此, $P_{(s,d)}(N, L)$ 被用于衡量 SnF 调度问题潜在计算复杂度的重要指标, 其表达式参见文献[7]。

SnF 的核心思想是用存储交换带宽。因此, 分析模型需要考虑网络资源的可用性。预约失败率 $F_{(s,d)}(N, L)$ 定义为请求未能从一组总数为 $P_{(s,d)}(N, L)$ 的可能路径集合中找到有效路径的概率。假设 p_b 表示 r 未能在一条空间链路上找到所需带宽的概率, p_s 表示 r 未能在一条时间链路上找到所需存储的概率。 $F_{(s,d)}(N, L)$ 被用于衡量 SnF 调度问题潜在调度性能的重要指标, 其表达式参见文献[7]。由于典型的 inter-DCNs 中带宽资源通常比存储资源稀缺, 因此为了简单起见, 假设 $p_b=0.3, p_s=0.1$ 。不同数值的 p_b 与 p_s 对 $F_{(s,d)}(N, L)$ 的影响可参考文献[7]。

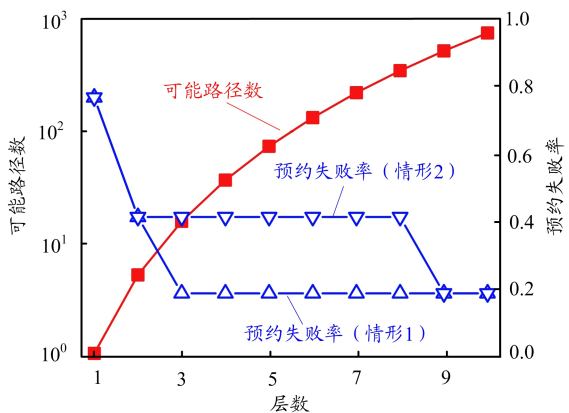
假设 TS-MLG 中不存在任何冗余状态。图 4(a) 显示 $P_{(s,d)}(N, L)$ 和 $F_{(s,d)}(N, L)$ 随 TS-MLG 层数 L 变化的情况, $P_{(s,d)}(N, L)$ 随 L 增加而增大, N 值越大, $P_{(s,d)}(N, L)$ 增加幅度越大。选择更长的路由 (即增加 N) 或扩展

时间调度的范围 (即增加 L) 都可以提供更多的可能路径。 $F_{(s,d)}(N, L)$ 随 L 增大而减小, N 值越大, $F_{(s,d)}(N, L)$ 的减小幅度越小。相比扩展时间调度的范围增大请求成功预约到所需资源的可能性, 选择更长的路由反而使得请求难以预约到所需资源。

假设 TS-MLG 中 3 层为有效状态, 其它均为冗余状态。情形 1: 假设 TS-MLG 中的最高三层是有效的。如图 4(b) 所示, 当 L 增加超过 3 时, $F_{(s,d)}(N, L)$ 不再继续减小。情形 2: 假设 TS-MLG 中的最高 2 层和第 9 层都是有用的。在图 4(b) 中, 当 L 在 2~8 的范围内或超过 9 时, $F_{(s,d)}(N, L)$ 不再继续减小。然而, 不论何种情形, $P_{(s,d)}(N, L)$ 都随着 L 不断增大而增大。上述研究表明: 冗余状态的增加不仅无法改善调度性能, 而且加重了路由计算的负担。



(a) 不存在冗余状态 ($N=2$ 和 5)



(b) 存在冗余状态 ($N=5$)

图 4 不同节点数 N 下, 可能路径数 $P_{(s,d)}(N, L)$ 和预约失败率 $F_{(s,d)}(N, L)$ 随层数 L 变化情况 ($p_b=0.1, p_s=0.3$)

2 SFS 方法

本文上节内容研究了大规模网络下调度问题难以求解的根本原因在于大量冗余状态的存在。此外,

网络负载的动态性要求调度方法能够实时更新和维护网络状态信息。然而, 不断增加的冗余状态也给传统集中式状态维护造成了沉重的负担。因此, 减少冗余状态是设计高效调度方法的关键, 本文提出 SFS 方法。

2.1 网络模型与主要假设

假设每个 DC 节点都能存储集群缓存大数据, 如图 5 所示。文献[10]介绍了存储集群的结构。在图 5 中, 来自 DC1 的大数据通过光路 1 传送并存储在 DC2 的存储集群。当网络空闲后, 数据再通过光路 2 传送到 DC3。假设每个节点都具备波长转换功能, 每个传输请求占用一个波长。相比传输延迟, 数据传播延迟和处理开销(例如, 路由决策和光路建立)可忽略不计。存储服务器集群的读/写速度与一个波长的传输速率相当。

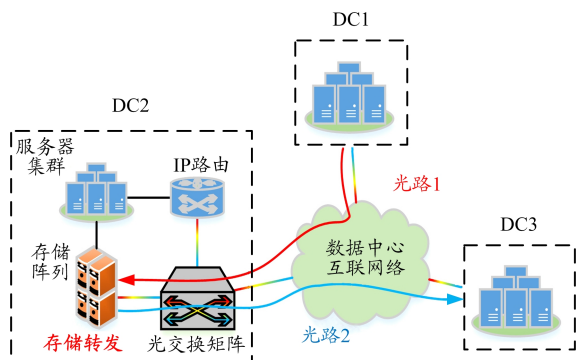


图 5 光互联数据中心网络中 SnF 大数据传输方案

2.2 SFS 方法描述

SFS 方法的特点表现在 2 个方面: ①对于不同传输请求, SFS 方法通过将相关预选路由上的单向链路信息融合, 以构建调度问题。因此, 相比优化法和 TSD 方法, SFS 方法最大限度地减少了调度问题中的冗余状态, 缩小了问题规模。②与集中式状态维护不同, SFS 方法为每条链路单独维护状态信息, 即每条链路都独立维护一个 TS-MLG 来记录该链路状态信息随时间的变化, 从而避免了维护大量冗余状态。因此, 相比优化法和 TSD 方法, SFS 方法维护的状态数量明显减少。

SFS 方法的流程如图 6 所示。假设从节点 s 到节点 d 的传输请求 r , 到达网络, 即 $r=\{s, d\}$ 。首先, SFS 方法选择一条从节点 s 到节点 d 的路由 R_k 。其次, 将路由 R_k 上每条链路 e 的 TS-MLG(即 G_e), 逐一合并, 从而构建辅助图 G' 。然后, 利用宽度优先搜索算法搜索辅助图 G' , 寻找可行路径。如果成功地找到可行路

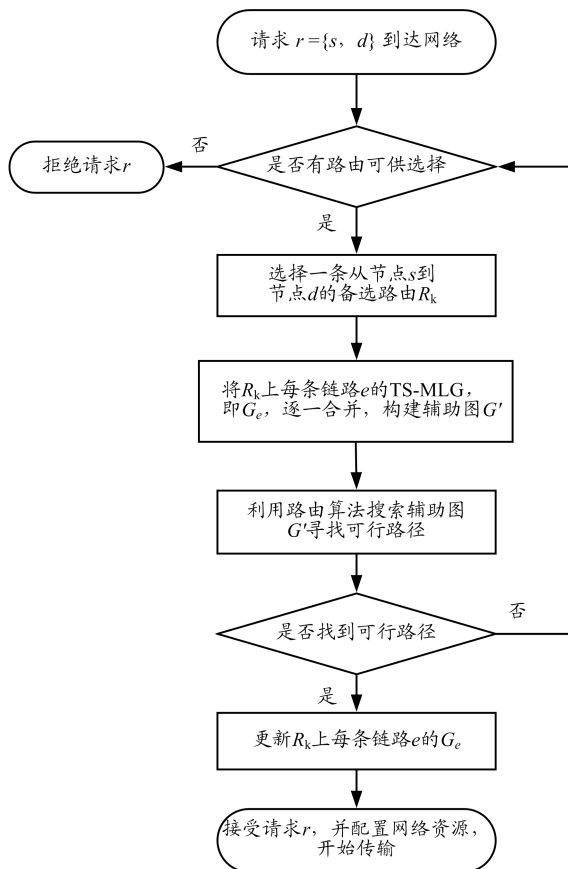


图 6 SFS 方法流程图

径, 则更新 R_k 上每条链路的 G_e , 接受请求 r ; 否则, 选择下一条路由 R_{k+1} 重新运行算法。如果所有路由都无法找到可行路径, 则请求 r 将被拒绝。

3 实验结果与分析

本文在 Matlab 中建立模拟仿真动态网络的场景, 对比 SFS 方法与 TSD 方法的性能。仿真实验采用本文 2.1 节的主要假设, 并使用 NSFNET 拓扑结构。请求到达服从泊松过程, 到达率为 λ 。请求持续时间服从负指数分布, 平均持续时间为 $1/\mu$ 单位时间。网络负载 ρ 等于 λ/μ 。 L_R 表示可用于请求路由的层数, L_R 越大获得的调度性能就越好^[9], 调度方法的复杂度也越高。因此, L_R 可用于调度的网络状态数的上限。 w 表示每条链路的波长数。 K 表示每个节点对之间的最短路由数。为了简单起见, 假设每个节点上的存储容量约束是无限的。空间、时间链路的链路代价均设为 1, 路由问题转化为最短跳数问题。每次仿真实验产生 500000 个请求, 独立重复 20 次并取平均。

3.1 网络性能研究

本文首先研究了阻塞率与链路利用率随网络负载 ρ 的变化情况, 当 $w=4, L_R=4$ 时, 仿真结果如图 7 所示。图 7(a) 中, 当 $K=3$ 和 5 时, SFS 方法的阻塞率优于 TSD 方法。当 $\rho=20$ 时, SFS 方法的阻塞率为 0。图 7(b) 中, 当 $\rho=20\sim30$ 时, SFS 方法的链路利用率与 TSD 方法相似。当 ρ 增加到 40 以上时, SFS 方法的链路利用率高于 TSD 方法。由图 7 可知, 与 TSD 方法相比, SFS 方法能够在保持低阻塞率的同时获得高利用率。为了进一步理解上述结果, 本文继续研究了其它网络性能指标随 ρ 的变化情况, 结果如图 8 所示。

活跃请求 (active request) 定义为已被网络接受但尚未完成传输的请求。图 8(a) 中, 当 ρ 超过 40 时, SFS 方法的网络能够比 TSD 方法的网络容纳更多活跃请求。这是因为与 TSD 方法相比, SFS 方法包含更少的冗余状态, 可以更有效地调度请求传输。因此, SFS 方法中请求缓存率也高于 TSD 方法, 如图 8(b) 所示。此

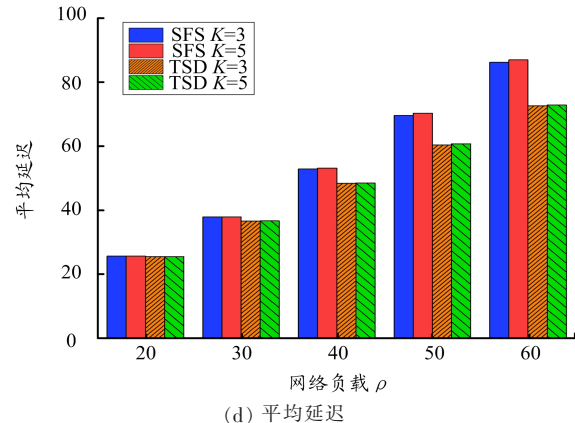
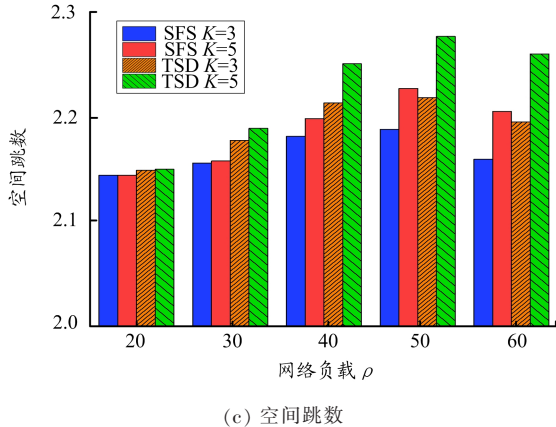
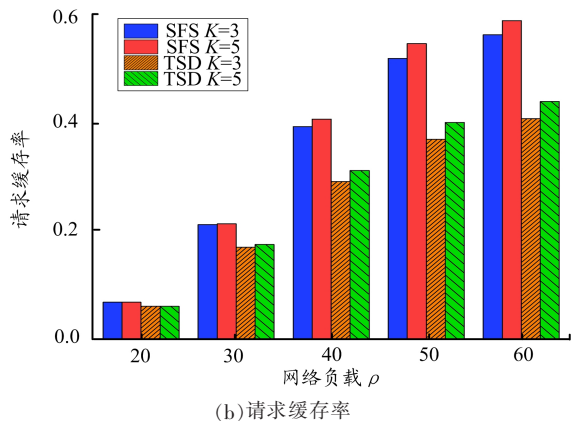
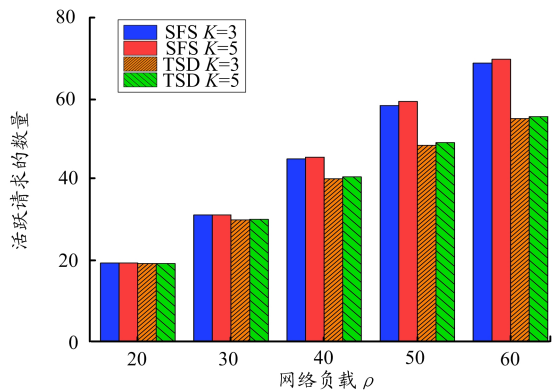
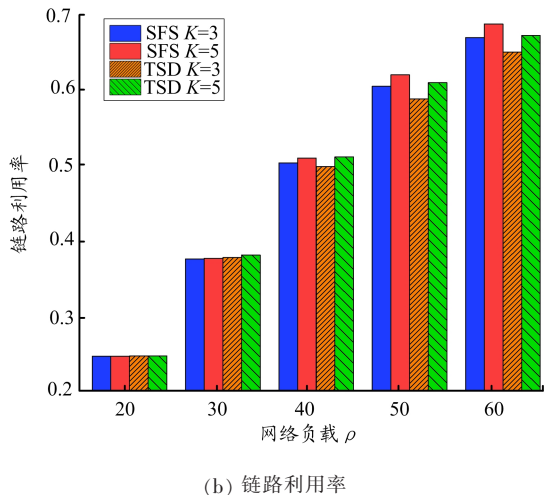
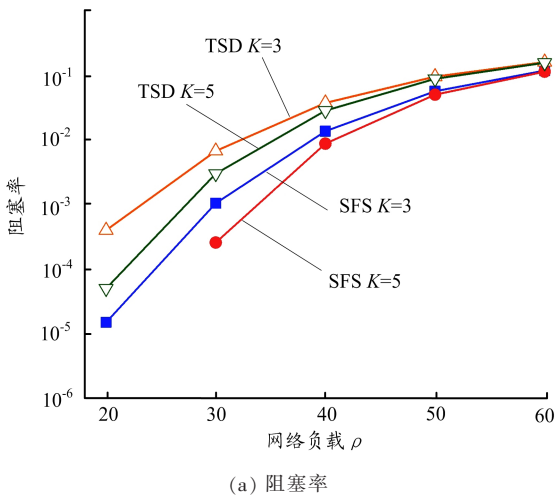


图 7 不同网络负载下, 阻塞率与链路利用率

图 8 不同网络负载下的网络性能指标

外,在 SFS 方法下,请求可以使用更长的缓存时间,等待拥有更短跳数的传输路径出现,如图 8(c)所示,请求在 SFS 方法中所选择传输路径的空间跳数比在 TSD 方法中的更短。虽然 SFS 方法比 TSD 方法具备更有效的调度请求,但代价是平均延迟更长,如图 8(d)所示。

3.2 网络状态维护研究

本文研究对比了 SFS 方法与 TSD 方法在仿真过程中维护的平均网络状态数,即 TS-MLG 的层数。虽然 2 种调度方法可用于请求调度的 TS-MLG 层数是相同的,即均为 L_R ,但是 2 种调度方法都需要根据网络负载的变化动态更新和维护所有链路的状态信息。因此,用于记录网络状态的 TS-MLG 层数在仿真过程中是实时变化的,过时的层或新的层将不断从 TS-MLG 中被删除或加入,其总层数通常大于 L_R ^[7]。由于 2 种调度方法维护状态信息的方式不同,导致需要维护的 TS-MLG 层数也不同。调度方法维护的平均网络状态结果如图 9 所示。可以看出,与 TSD 方法相比,SFS 方法需要维持的网络状态信息更少。

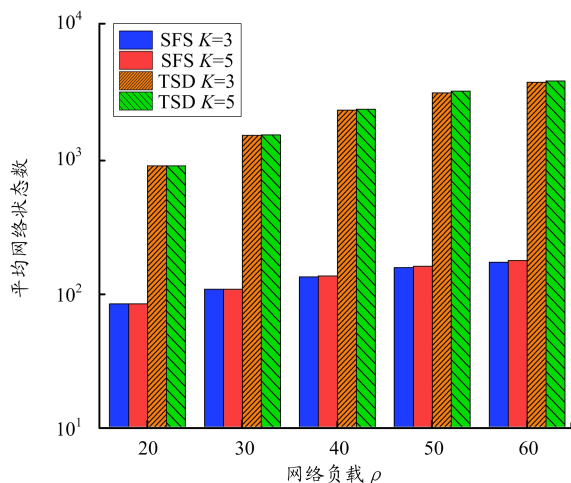


图9 调度方法维护的平均网络状态数

4 结束语

本文首先研究了网络冗余状态对 SnF 调度的影响。主要研究发现:①对于绝大多数传输请求,网络存在的大量状态信息是冗余的。随着网络规模的扩大,

冗余状态的数量会急剧增加。②冗余状态不仅无法提供有效信息,反而增加调度问题的求解难度,以及状态信息的维护成本。因此,本文提出并介绍了 SFS 方法。仿真结果表明:在可用于调度的网络状态数受限的情况下(即 L_R 有限),SFS 方法能比传统调度方法获得更好的阻塞率、维护更少的网络实时状态信息。然而,SFS 方法请求经历的延迟也更高。

参考文献:

- [1] DORADO J L G, RAO S G. Cost-aware Multi Data-Center Bulk Transfers in the Cloud from a Customer-Side Perspective [J]. IEEE Transactions on Cloud Computing, 2019, 7(1): 34–37.
- [2] YANG Z, CUI Y, WANG X, et al. Cost-Efficient Scheduling of Bulk Transfers in Inter-Datacenter WANs [J]. IEEE/ACM Transactions on Networking, 2019, 27(5): 1973–1986.
- [3] LU P, ZHU Z. Data-Oriented Task Scheduling in Fixed- and Flexible-Grid Multilayer Inter-DC Optical Networks: A Comparison Study [J]. Journal of Lightwave Technology, 2017, 35(24): 5335–5346.
- [4] CHHABRA P, ERRAMILI V, LAOUTARIS N, et al. Algorithms for Constrained Bulk-Transfer of Delay-Tolerant Data [C]//IEEE International Conference on Communications, May 23–27, 2010, Cape Town, South Africa. Piscataway: IEEE, 2010: 1–5.
- [5] WANG Y, SU S, LIU A X, et al. Multiple Bulk Data Transfers Scheduling Among Datacenters[J]. Computer Networks, 2014, 68: 123–137.
- [6] IOSIFIDIS G, KOUTSOPOULOS I, SMARAGDAKIS G. Distributed Storage Control Algorithms for Dynamic Networks [J]. IEEE/ACM Transactions on Networking, 2017, 25(3): 1359–1372.
- [7] LIN X, WANG X, YUE S, et al. Design of an SnF Scheduling Method for Bulk Data Transfers over Inter-Datacenter WANs[C]//IEEE 20th International Conference on High Performance Switching and Routing (HPSR), May 26–29, 2019, Xi'an, China. Piscataway: IEEE, 2019: 1–8.
- [8] LIN X, SUN W, VEERARAGHAVAN M, et al. Time-Shifted Multilayer Graph: A Routing Framework for Bulk Data Transfer in Optical Circuit-Switched Networks with Assistive Storage[J]. IEEE/OSA Journal of Optical Communications and Networking, 2016, 8(3): 162–174.
- [9] MORI Y, GANBOLD M, SAYO K. Design and Evaluation of Optical Circuit Switches for Intra-Datacenter Networking [J]. Journal of Lightwave Technology, 2019, 37(2): 330–337.
- [10] DART E, ROTMAN L, TIERNEY B, et al. The Science DMZ: A Network Design Pattern for Data-Intensive Science [C]//Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, November 17–22, 2013, Denver, USA. Piscataway: IEEE, 2014: 1–10.