文章编号:1671-4598(2020)03-0227-04 DOI:10.16526/j.cnki.11—4762/tp.2020.03.047 中图分类号:TP393.0 文献标识码:A

数据中心网络中基于传输速率分配的 TCP 协议

侯文放，王 辉，曾 波
（河南科技大学 信息工程学院，河南，洛阳 471000）

摘要：传输控制协议（Transmission Control Protocol，TCP）是数据中心网络中常用的传输协议；然而，同一个网络环境下，由于 TCP 协议的公平共享原则，TCP 无法保障不同优先级业务的服务质量；针对该问题，提出了基于传输速率分配算法的 TCP（Transmit Rate Allocation based TCP，TRA－TCP)；汇聚节点实时监测记录不同优先级业务的数据传输速率，并为不同优先级的业务流分配不同的传输速率，以保障高优先级业务的 QoS；实验表明，与已有协议相比，TRA－TCP 协议最高能够将高优先级业务的吞吐量提升 25%，并且将高优先级业务的时延保持在 0.1 秒以下，丢包率保持在 5%左右。

关键词：数据中心网络；服务质量保障；传输速率

Transmission Rate Allocation Based TCP Protocol in Data Center Network

Hou Wenfang，Wang Hui，Zeng Bo
（College of Information Engineering，Henan University of Sience and Technology，Luoyang 471000 China）

Abstract：Transmission control protocol (TCP) is a commonly used transmission protocol in data center networks. However，under the same network environment，due to the principle of fair sharing of TCP protocols，TCP cannot guarantee the quality of service (Qos) of different priority services. Aiming at this problem，a TCP protocol based on the transmission rate allocation algorithm (TRA－TCP) is proposed. The aggregation node monitors and records the data transmission rate of different priority services in real time，and allocates different transmission rates for service flows of different priorities to ensure the QoS of high priority services. Simulation experiment results show that TRA－TCP protocol can improve the throughput of high priority traffic to 25% compared to the ohter existing protocols，keep the average packet delay of high priority traffic below 0.1 seconds，and keep the packet loss rate around 5%.

Keywords：data center network；QoS；transmission rate

0 引言

数据中心网络技术在网络搜索、在线零售、广告系统、社交网络等方面的推广与应用为用户提供了高质量的网络服务。这些服务都使用在线交互式应用 OLDI（online data－intensive)[1]，且具有两大特点：1）应用程序都具有不同的截止时限，无法在截止时限内完成传输的流将不被统计在反馈的结果中；2）这些应用采用基于树的设计模式。通常情况下，用户的请求会被分配给多个工作站并行处理。数据中心网络服务的上述特点为传输层协议的设计提出了巨大挑战。

针对数据中心网络的应用需求，传输层协议设计人员提出了两点目标：1）降低流的平均完成时间；2）根据不同流的截止时限来限制发送窗口。在前述目标的指引下，目前面向数据中心网络的传输层协议研究的主流方向之一是在传统的 TCP（transmission control protocol）协议的基础上，针对数据中心网络的特性进行适应性改进，以适应数据中心网络的需求。比如：DCTCP（data center TCP）协议[2]、D^3 协议[3]以及很多对 TCP 协议的优化[4-7]。上述协议均从数据中心网络的特性出发，通过改进 TCP 协议的传输流程或者优化 TCP 协议参数，从而提高了数据中心网络中 TCP 协议的性能。但是，上述协议均未能考虑保障业务的服务质量（quality of service，QoS)。然而，在数据中心网络中，如果无法保障用户在发出请求后的截止时间内获得相关数据，则直接影响网络服务的性能以及用户体验，导致用户的流失以及网络收益的下降。

研究者针对保障数据中心网络中的 QoS 问题，提出了不少协议设计的方案[12-15]。但是，这些研究工作主要考虑了单条业务流的优化，而没有考虑到网络的整体负载状况以及拥塞状况，无法满足用户 QoS 需求。针对目前的数据中心网络传输机制无法有效地保障业务流 QoS 的问题，本文提出了基于传输速率分配算法的 TCP 协议（Transmit Rate Allocation based TCP，TRA－TCP)。

1 相关工作

目前数据中心网络中的在线数据密集型应用通常使用分割－聚合的设计模式，汇聚节点接收到用户发出的请求，分配多个工作站共同完成工作，在计算完成后，工作站返回结果。一般而言，数据中心网络的结构如图 1 所示：共

收稿日期:2019-08-08； 修回日期:2019-09-05。
基金项目:河南省科技计划项目(172102210255);赛尔网络下一代互联网创新项目(NGII20160517)。
作者简介:侯文放(1993-),男,河南郑州人,硕士研究生,主要从事数据中心网络拥塞控制方向的研究。
王 辉(1966-),女,河南洛阳人,教授,博士生导师,主要从事数据挖掘,网络性能改善方向的研究。

有 4 个工作站，工作站通过汇聚节点向数据中心传输数据分组，当汇聚节点发生拥塞之后，则汇聚节点通过显示拥塞通告（explicit congestion notification，ECN）机制向工作站反馈拥塞信息。下面将分别介绍上述两种数据中心网络协议类型。

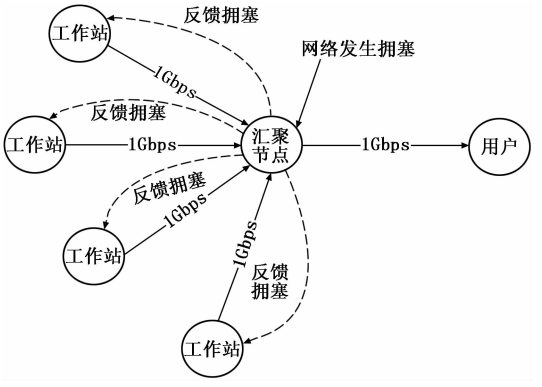

图 1 数据中心网络协议的网络结构

在这种分割－聚合的工作模式下，多个计算结果同时反馈用户会造成汇聚节点的缓存溢出，导致丢包。但是数据中心网络的在线数据密集型应用往往具有软实时性[8]，受到截止时限的约束。不能及时的返回结果，大大影响了用户体验和运营商的投资回报[9]。

目前，针对数据中心网络的 TCP 协议改进主要包含两类方案：无 QoS 保障的 TCP 改进协议以及基于 QoS 保障的 TCP 改进协议。

DCTCP[2]是最早提出的面向数据中心网络的 TCP 改进协议。该协议利用 ECN（explicit congestion notification）[10]机制，根据网络的拥塞程度调节发送速率，减少了拥塞时的丢包。在 DCTCP 的基础上，D^3 协议[3]采用了集中式的调度和主动式的速率请求，交换机贪婪的在先到先得的原则上分配带宽。这种贪婪的方式导致了一些离时限更近的流不能分配到合理的带宽。文献［11］针对数据中心网络中因为 SYN 包丢失而引起的 TCP 连接延迟问题，提出了一种基于加权随机早起检测的改进协议以优化 TCP 协议的同步过程。然而，此类解决方法虽然能够有效地提升网络整体性能，比如：DCTCP 降低了网络的整体冲突概率，提升了整体的吞吐量，但是，却没有考虑业务的 QoS 需求。

针对上述协议缺乏网络服务质量的问题，出现了一些以保障数据中心网络业务 QoS 的研究。D^2TCP（Deadline－aware Data－center TCP）协议[12]在 DCTCP 拥塞控制的基础上考虑了流的时限感知，以保障业务的时延需求。LSTCP（Least Slack－aware TCP）协议[13]通过采用最小空闲时间优化调度策略对 TCP 流进行优先级划分，提出了一种基于数据中心网络闲时感知算法，从而减小了短流的完成时间，并提高了长流的吞吐量。PD^2TCP 协议[14]针对 TCP 流无法在截止时限内完成传输的问题，提出了基于优先级时限感知的 TCP 改进协议，通过感知瞬时队列长度以及 ECN 标记对于拥塞窗口进行调整。文献［15］中针对多媒体业务的 QoS 保障，提出了源端多媒体数据流带宽控制策略以及基于动态部分缓存共享的丢包控制方法，从而实现了对于 TCP 协议拥塞控制机制的改进。

虽然，已有部分研究成果考虑了数据中心网络的 QoS 保障问题，但是主要针对于单 TCP 流进行优化以保证单 TCP 流的 QoS，缺乏对整体网络状况的考虑。基于此，本文提出了 TRA－TCP，从网络整体负载能力方面，根据高优先级和低优先级业务的比例调整拥塞窗口，旨在为不同优先级业务分配不同传输速率，并且保障高优先级业务的 QoS。

2 算法设计

本节主要描述了 TRA－TCP 协议的核心算法以及具体的工作流程。首先采用伪代码方式描述传输速率控制算法的思想和总体流程，其次描述了汇聚节点检测网络不同优先级业务传输速率的方法，最后描述了汇聚节点为不同优先级业务分配传输速率的算法以及拥塞窗口的调节方法。

2.1 算法设计

本节主要描述了论文算法的构成及运作机制。算法 1 中详细介绍了算法流程。算法的输入参数为高优先级业务的目标传输速率以及网络调节的周期；算法输出为高优先级以及低优先级业务调节后的拥塞窗口。

算法 1：Transmission Rate Control Algorithm

输入：target transmit rate $\alpha(0 \leqslant \alpha \leqslant 1)$, the circle of updating transmit rate α

Variable definition：

$T(i)$：transmit rate of high priority service F

g：congestion window of high priority service i

$T(j)$：transmit rate of low priority service j

$W(j)$：congestion window of high priority service j

N：the number of high priority services

M：the number of low priority services

β：the percentage of the transmit rate of high priority service to the transmit rate of low priority service

γ：the improving ratio of high priority services

η：the reducing ration of low priority services

输出：$W(i)*$ and $W(j)*$

1) initialization：convergence node collects transmit rate and congestion window information；

2) convergence node starts updating transmit rate process every S time；

3) $T_{\text{high}} = sum(T(i))$, and $T_{\text{low}} = sum(T(j))$；

4) if $T_{\text{high}} \geqslant NT*$, go to step 1；

5) else $\beta = T_{\text{high}}/T_{\text{low}}$；

6) $\gamma = (NT* - T_{\text{high}})/T_{\text{high}}$；

7) calculate the congestion window the low priority services need to reduce W_r；

8) $\eta = W_r/M$；

9) $W(i)* = W(i) \times (1+\gamma)$；

10) $W(j)* = W(j) \times (1-\eta)$。

2.2 监测业务的传输速率

对于汇聚节点而言，由于其接收所有工作站的业务分组，故汇聚节点可以获取高优先级业务和低优先级业务的比例（IP 包头中有优先级标志位），并且监测不同优先级业务的传输速率。假设数据中心网络中存在 N 个高优先级业务，每个高优先级业务传输速率为 $T(i)$，拥塞窗口为 $W(i)$，M 个低优先级业务，每个低优先级业务传输速率为 $T(j)$，拥塞窗口为 $W(j)$，并且汇聚节点计算业务传输速率的周期为 S，即每隔 S 时长汇聚节点会统计这段时间内监测到的不同优先级业务的传输速率。

汇聚节点获取到不同优先级业务的传输速率之后，可以定义高优先级业务和低优先级业务比例关系 β 为：

$$\beta = \frac{\sum_{j}^{N}(i)}{\sum_{j}^{M}(j)} \tag{1}$$

因为网络设定的高优先业务的目标传输速率为 T^*，故定义高优先业务传输速率提升系数 γ 来描述当前网络的配置能否满足高优先级业务的传输速率需求。γ 的定义如下：

$$\gamma = \frac{NT^* - \sum_{i}^{N}T(i)}{\sum_{i}^{N}T(i)} \tag{2}$$

当 γ 大于 0 时，表示网络无法满足高优先级业务的目标传输速率，反之表示高优先级业务的目标传输速率得到满足。因此，当 γ 大于 0 时，汇聚节点需要重新进行传输速率分配。

此外，汇聚节点采用主动队列调整算法（RED）[16]，因此当缓存队列大于门限值时就会发生丢包事件，故汇聚节点还需要记录网络整体的拥塞状况。根据文献［2］定义拥塞参数 $\alpha(0 \leqslant \alpha \leqslant 1)$ 表示网络的拥塞程度。α 参数的计算周期同样为 S，计算方法如下所示：

$$\alpha = (1-g) \times \alpha + g \times F \tag{3}$$

其中：F 表示 S 时间内发生拥塞的分组比例，g 表示上一个 α 参数所占权重比例。

2.3 传输速率分配

为了满足网络需求以及高优先级业务的需求，网络需要重新进行传输速率分配。具体而言，汇聚节点会根据获取的不同业务的传输速率，计算出不同 TCP 流拥塞窗口的调节方法，从而达到传输速率分配的目的。

首先需要保障高优先级业务的传输速率。根据之前的计算结果，可以知道高优先级业务需要提升 γ 倍的传输速率才能满足 QoS 需求。因为拥塞窗口与传输速率的关系是正比关系，故高优先级业务的拥塞窗口需要提升为之前的 $(1+\gamma)$：

$$W(i)^* = W(i) \times (1+\gamma) \tag{4}$$

其次由于网络整体发生了拥塞，根据 DCTCP 协议的建议[2]，网络整体的拥塞窗口需要降低 $\alpha/2$。另外，由于高优先级的业务的拥塞窗口增加到之前的 $(1+\gamma)$ 倍，故也需要降低这部分增量。因此，需要减小的总体拥塞窗口值为：

$$W_{all} = (\frac{\alpha}{2}+\gamma)\sum_{i}^{N}W(i) + \frac{\alpha}{2}\sum_{j}^{M}W(j) \tag{5}$$

根据公式（4）化简如下：

$$W_{all} = (\frac{\alpha\beta}{2}+\gamma\beta+\frac{\alpha}{2})\sum_{j}^{M}W(j) \tag{6}$$

最终可以得到每个低优先级业务需要减小的塞窗口数值为：

$$W(j)^* = W(j) \times (1-\eta) \tag{7}$$

其中 η 等于：

$$\eta = \frac{\alpha\beta}{2}+\gamma\beta+\frac{\alpha}{2} \tag{8}$$

汇聚节点通过公式（4）和公式（7）可以获取到高优先级和低优先级拥塞窗口的调节方法，之后需要向工作站反馈调节结果。参考 DCTCP[2] 以及 D²TCP[11]，汇聚节点在 ACK 数据包中携带拥塞调节信息。具体而言，ECN 为 1 表示减小拥塞窗口，ECN 为 0 表示增大拥塞窗口，ACK 中的 4 比特保留位用于携带 γ 以及 η 参数，即 γ 和 η 参数需要量化为离散值。

3 仿真实验

为了验证 TRA－TCP 的协议性能，通过 NS2 网络仿真软件搭建了相应的仿真平台，主要比较了 TRA－TCP 协议与 DCTCP 协议以及 D²TCP 的性能差异，所采用的性能评价指标为：网络吞吐量、平均时延、丢包率。

3.1 实验环境

实验参数设置如表 1 所示。高优先级业务数量固定为 5，低优先级业务数量则为［2，4，6，8，10］。高优先级业务和低优先级业务的业务速率均为 100 Mbps。

表 1　实验参数配置

参数名称	参数设置
高优先级业务速率	100 Mbps
高优先级业务数量	5
低优先级业务速率	100 Mbps
低优先级业务数量	2,4,6,8,10
数据包尺寸	1460 字节
工作站链路时延	0.025 ms
数据中心链路时延	0.025 ms
工作站带宽	1.1 Gbps
数据中心带宽	1 Gbps
工作站队列模型	先入先出模型
数据中心队列模型	随机丢包模型
数据中心队列长度	250 数据包
仿真时间	10 秒
吞吐量的目标值	100 Mbps

3.2 实验结果分析

从图 2 可以看出，对于 DCTCP 协议以及 D²TCP 协议而言，随着低优先级业务数量的不断增加，高优先级业务的吞吐量不断下降，而低优先级业务吞吐量不断提升。另

外，D²TCP 协议的吞吐量低于 DCTCP 协议，这是因为 D²TCP 协议需要保证业务的截止时间，故降低了吞吐量。相比之下，由于 TRA－TCP 协议将高优先级业务的传输速率目标 T^* 设置为 100 Mbps，并通过拥塞窗口的调整，保证了高优先业务流的吞吐量，因而，能够获得比 DCTCP 和 D²TCP 较高的网络吞吐量。

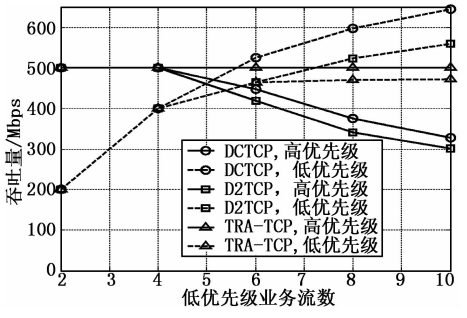

图 2 吞吐量的仿真结果

图 3 展示了业务流的时延变化情况。从图中可以看出 DCTCP 协议不关注业务的截止时间，因此其平均时延最大。D²TCP 考虑了业务的截止时间，故能够一定程度保障业务的平均时延，且高优先级业务的平均时延低于低优先级业务的平均时延。对于 TRA－TCP 协议而言，充分保障了高优先级业务的平均时延，但是相对应的低优先级业务的平均时延较高。

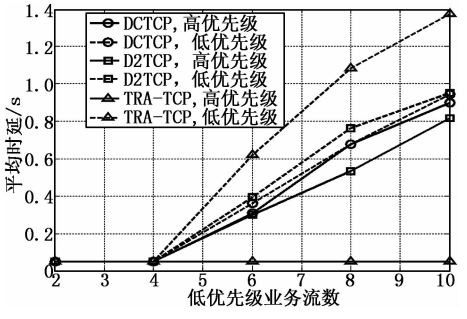

图 3 平均时延的仿真结果

从图 4 可以看出，DCTCP 协议和 D²TCP 协议的丢包率都维持在较高水平，特别是高优先级业务的丢包率不断增加。TRA－TCP 协议中的高优先级业务则保持几乎不丢包的状态，低优先级业务的丢包率则维持较高水平。

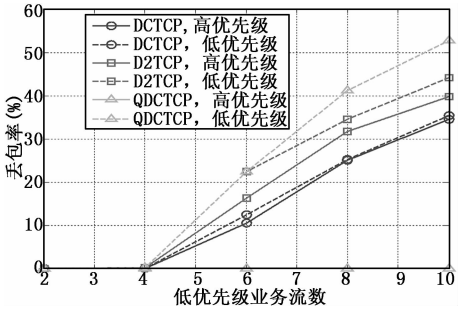

图 4 丢包率的仿真结果

4 总结

针对数据中心网络中业务的优先级和吞吐量的综合性能需求，本文提出了一种基于传输速率分配的服务质量保障 TCP 协议：TRA－TCP 协议。TRA－TCP 协议根据高优先级业务和低优先级业务的数量比例的分配带宽，调节拥塞窗口，保障了高优先级业务的吞吐量。仿真实验结果表明：TRA－TCP 协议不仅能够有效保障高优先业务的吞吐量，还能降低高优先级业务的平均时延以及丢包率。

参考文献：

[1] Meisner D，Sadler C M，Barroso L A，et al. Power management of online data－intensive services [A]. International Symposium on Computer Architecture [C]. ACM，2011：319－330.

[2] Alizadeh M，Greenberg A，Maltz D A，et al. Data center TCP (DCTCP) [J]. Acm Sigcomm Computer Communication Review，2010，40 (4)：63－74.

[3] Wilson C，Ballani H，Karagiannis T，et al. Better never than late：meeting deadlines in datacenter networks [Z]. ACM，2011：50－61.

[4] Wu H，Feng Z，Guo C，et al. ICTCP：Incast Congestion Control for TCP in data center networks [A]. International Conference [C]. ACM，2010：1－12.

[5] Guo C，Yuan L，Xiang D，et al. Pingmesh：A Large－Scale System for Data Center Network Latency Measurement and Analysis [A]. ACM Conference on Special Interest Group on Data Communication [C]. ACM，2015：139－152.

[6] Mittal R，Lam V T，Dukkipati N，et al. TIMELY：RTT－based Congestion Control for the Datacenter [A]. ACM Conference on Special Interest Group on Data Communication [C]. ACM，2015：537－550.

[7] Zhu Y，Kang N，Cao J，et al. Packet－Level Telemetry in Large Datacenter Networks [A]. ACM Conference on Special Interest Group on Data Communication [C]. ACM，2016：479－491.

[8] 邓 罡，龚正虎，王 宏，等. 现代数据中心网络资源管理技术分析与综述 [J]. 通信学报，2014 (2)：166－181.

[9] Kohavi R，Longbotham R. Online Experiments：Lessons Learned [J]. Computer，2007，40 (9)：103－105.

[10] Ramakrishnan K，Floyd S，Black D. The Addition of Explicit Congestion Notification (ECN) to IP [M]. RFC Editor，2001.

[11] 罗万明，林闯，阎保平. 一种支持多媒体通信 QoS 的拥塞控制机制 [J]. 电子学报，2000，28 (z1)：48－52.

[12] Vamanan B，Hasan J，Vijaykumar T N. Deadline－aware datacenter tcp (D2TCP) [A]. Acm Sigcomm Conference on Applications [C]. ACM，2012：115－126.

[13] 刘 洪，伊 鹏，胡宇翔. 基于动态优先级的数据中心网络闲时感知 TCP 协议 [J]. 计算机应用研究，2018 (1)：190－194.

[14] 赵正伟，许 刚，毕经平. 基于优先级的时限感知的数据中心网络拥塞控制算法 [J]. 高技术通讯，2014，24 (6)：587－596.

[15] 罗万明，林 闯，阎保平. 一种支持多媒体通信 QoS 的拥塞控制机制 [J]. 电子学报，2000，28 (z1)：48－52.

[16] Floyd S，Jacobson V. Random early detection gateways for TCP congestion avoidance [Z]. 1993：397－413.