Accepted Manuscript

Congestion control in high-speed lossless data center networks: A survey

Shan Huang, Dezun Dong, Wei Bai

PII: S0167-739X(18)30032-3
DOI: https://doi.org/10.1016/j.future.2018.06.036
Reference: FUTURE 4298

To appear in: *Future Generation Computer Systems*

Received date : 26 January 2018
Revised date : 21 May 2018
Accepted date : 20 June 2018

Please cite this article as: S. Huang, D. Dong, W. Bai, Congestion control in high-speed lossless data center networks: A survey, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.06.036

Congestion Control in High-speed Lossless Data Center Networks: A Survey

Shan Huang, Dezun Dong, Wei Bai

***Abstract*—In data centers, packet losses cause high retransmission delays, which is harmful to many real-time workloads. To prevent packet losses, the lossless fabrics have been deployed in many production data centers. However, when network congestion happens, the lossless fabric also causes many problems like saturation tree and unfairness, which seriously degrade the performance of data center applications. Therefore, how to control the congestion in high-speed lossless data center networks is a significant problem.**

In this paper, we first introduce link layer flow control schemes to provide lossless fabrics. Then we survey congestion control schemes in high-speed lossless data center networks. In particular, we classify existing congestion control schemes into two categories: reactive and proactive. Finally, we present the challenges and opportunities for future research in this area.

***Index Terms*—High-speed Lossless Data Center Network (DCN); Network Congestion; Congestion Control.**

I. Introduction

Many companies, such as Microsoft, Google, Amazon and Alibaba, have built lots of data centers around the world to provide high-quality service to global users [1][2][3][4][5][6]. Inside a data center, the hundreds of thousands of servers are inter-connected by a data center network (DCN) with high link speed (10-100Gbps) and low base latency (10-100μs) [7][8].

Today's data centers host many real-time applications, e.g., web search, retail and social networking [9][10][11][12][13][14]. To serve a user request, these applications will generate lots of small latency-sensitive request and response messages in DCNs. The user experience is determined by how fast the application collects all (or most of) the response messages. Hence, the high delay in DCNs can seriously degrade the performance of these applications, resulting in poor user experience and operator revenue. However, data center traffic is well known for its burstiness and data center switches typically have very shallow buffers [1][15][16]. In DCNs, the bursty traffic is likely to overfill the shallow switch buffers, leading to high retransmission delay and poor user experience.

To minimize retransmission delay, lossless networks have been deployed in production data centers. In contrast to traditional lossy networks, lossless networks employ the hop-by-hop Link-layer Flow Control (LFC) to avoid packet congestion losses. However, like lossy networks, when the congestion happens, lossless networks also suffer from high queueing delay [17]. Furthermore, due to the LFC, lossless networks may spread the congestion, resulting in congestion tree [18], causing unfairness problem (e.g., victim flows) and degrading throughput [19] (more details in §*III*). To mitigate above problems, we need advanced congestion control solutions in lossless DCNs.

Various survey papers related to data center congestion control have been proposed. But most of them emphatically summarized the techniques deployed in lossy data center networks. [20] presented a survey of end-to-end TCP congestion control schemes which do not require explicit feedbacks. [21] surveyed the congestion control schemes which aims at reducing the Flow Completion Time (FCT) and presented them according to four major techniques they used. [22] comprehensively surveyed several congestion control schemes for fast transmission of flows in DCNs, and compared the surveyed schemes based on their objectives, features and working mechanisms. Most recently, [23] surveyed a large number of traffic control schemes, and analyzed the characteristics, techniques and trade-offs of the surveyed schemes.

In this paper, we focus on the congestion control schemes for lossless data center networks. We classify the surveyed congestion control schemes into two categories: reactive congestion control (§*IV*) and proactive congestion control (§*V*). Reactive congestion control schemes iteratively react to congestion signals, such as Explicit Congestion Notification (ECN) and delay. They are reactive in nature as they only take effect after the congestion happens. In contrast, proactive congestion control schemes only send packets when they think that the network pipe has enough capacity to hold them. We further classify the reactive schemes into two categories: switch-based, host-based, and classify the proactive schemes into three categories: centralized, end-to-end decentralized and hop-by-hop decentralized. We make comparisons among different schemes based on several properties, *e.g.*, *facilitating small message*, *achieving fairness*, *implementation complexity*, *reaction speed* (for reactive schemes) and *preprocessing overhead* (for proactive schemes).

The rest of the paper is organized as follows. In §*II*, we introduce the network congestion, and present the criteria

Shan Huang was with the Department of Computer Science, National University of Defense Technology, Changsha, China, e-mail: angry.shanhuang@gmail.com

Dezun Dong was with the Department of Computer Science, National University of Defense Technology, Changsha, China, e-mail: dong@nudt.edu.cn

Wei Bai was with Microsoft Research, Beijing, China, e-mail: webai@microsoft.com