POLITECNICO DI TORINO
Repository ISTITUZIONALE

Traffic Optimization in Data Center and Software-Defined Programmable Networks

Original
Traffic Optimization in Data Center and Software-Defined Programmable Networks / Sviridov, German. - (2021 Mar 04), pp. 1-155.

Availability:
This version is available at: 11583/2875743 since: 2021-03-23T09:46:46Z

Publisher:
Politecnico di Torino

Published
DOI:

Terms of use:
Altro tipo di accesso

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 September 2022

Summary

In the past decade, the field of telecommunications went through a radical change in the way users interact with the network and the way networks evolved in the function of user needs. Data Centers have become the backbone of modern digital society and industry. This led to their rapid spread with the goal of providing fast and readily-available services to customers around the globe. Consequently, wide-area networks have grown exponentially more complex to accommodate the surge in bandwidth demand and new paradigms such as Software Defined Networking emerged to combat the increase in complexity. Nevertheless, even nowadays, most of the wide-area networks still rely on legacy network management algorithms and protocols which significantly limit the pace at which current network infrastructures are able to adapt to modern traffic scenarios. Similarly to what happens in the wide-area networks, data center networks still employ classic flow management mechanisms which, although being able to guarantee a good level of performance, do not fully exploit the potential of modern network architectures.

In this Thesis a major focus is devoted towards analyzing the impact of the radical change introduced by modern network infrastructures on the traffic flow performance. As a first contribution, we show that programmable data planes, although being able to overcome some of the shortcomings of traditional Software Define Networking, still fall short for many applications relevant to the operations of wide-area networks. This in turns contributes to the deterioration of traffic flow performance, for a wide range of applications. We address this issue by proposing a novel paradigm of designing programmable data planes-ready network applications that exploit replicated states inside the network, ultimately permitting to improve traffic flow performance.

A second contribution relates to the performance optimization of data center networks. Traditional flow scheduling mechanisms employed in data center networks are unable to keep up with the ever-increasing demand for highly responsive applications deployed in modern data centers. At the same time, solutions proposed in the literature rely on complex control mechanisms. Those solutions are capable of significantly improving flow scheduling policies, thus traffic flow performance. Yet, even by relaxing some of the requirements of those solutions, they still remain too complex and require complex modifications to the hardware of underlying devices composing the network. This ultimately results in prohibitively expensive solutions, thus inapplicable in realistic scenarios. We propose a flow scheduling mechanism based on aggregate flow statistics that is capable of achieving traffic flow performance close to state-of-the-art solutions while keeping the complexity low, thus making it accessible for the already available network infrastructures.

While our contributions in the field of programmable data planes and data center flow scheduling are capable of considerably improving traffic flow performance we show that blindly optimizing aggregate traffic flow performance does not

lead to optimal results in terms of user experience. Indeed, observing aggregate flow information does not give any insight on the nature of the service carried in single flows, thus precluding any possibility of understanding the impact of the network parameters on those flows. Our final contribution addresses this issue by proposing an efficient way of assessing the impact of the latency on interactive applications which dominate the modern Internet. We consider cloud gaming as a reference application and highlight the heterogeneity in network requirements for apparently similar flows. This is done by developing an automatic Quality of Experience assessment procedure which exploits modern advances in the field of artificial intelligence and deep reinforcement learning. Finally, we show that the proposed methodology can be employed to define and enforce fine-grained flow optimization policies capable of taking into account service-level network requirements.