Computer Communications 61 (2015) 58–69

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

OpenFlow-based in-network Layer-2 adaptive multipath aggregation in data centers

Tara Nath Subedi *, Kim Khoa Nguyen, Mohamed Cheriet

Ecole de Technologie Superieure, University of Quebec, Montreal, Quebec H3C1K3, Canada

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form 1 October 2014
Accepted 13 December 2014
Available online 20 December 2014

Keywords:
Multipath
Aggregated path capacity
OpenFlow
Routing
Forwarding

ABSTRACT

In order to satisfy the high bandwidth and performance demands of applications, host servers are built with multiple network interfaces, and a data center network consists of multiple redundant links. It is important to make efficient use of all the available network capacity, using multiple physical paths whenever possible, but traditional forwarding mechanisms using a single path are not able to take advantages of available multiple physical paths. The state-of-the-art MPTCP (Multipath Transmission Control Protocol) solution uses multiple randomly selected paths, but cannot give total aggregated capacity. Moreover, it works as a TCP process, and so does not support other protocols like UDP. This paper presents an alternative solution using adaptive multipath routing in a Layer-2 network with static (capacity and latency) metrics, which adapts link and path failures. This solution provides in-network aggregated path capacity to individual flows, as well as scalability and multitenancy, by separating end-station services from the provider's network. The results of deploying a proof-of-concept prototype on a data center testbed, which show the aggregated path capacity per flow, demonstrate an improvement of 14% in the worst bisection bandwidth utilization, compared to the MPTCP with 5 subflows.

1. Introduction

Server virtualization, consolidation, and cloud computing initiatives are enabling data center providers to pool their computing resources for multiple consumers using a multitenant model. The resources provided are location-independent, as they can be pooled from anywhere. This is reshaping data center traffic flows, and escalating the bandwidth and performance demands on the underlying physical network. In this environment, the traditional tiered tree topology gives poor reliability and leads to oversubscribed any-to-any network design, and forwarding along a tree constrains workload placement.

Rearchitecting the DCN topology to support high bisection bandwidth, as well as flexibility for incremental expansion and fault-tolerance, is an active research area [1–3]. In modern data centers, servers are often built with multiple interfaces, and their network topology consists of multiple redundant links, resulting in a multipath physical network. Fig. 1 depicts the DCN topology: circles and squares, representing switch nodes and host nodes, are connected by links of various capacity weights (in Gbps). A link is a direct connection between two adjacent nodes. A path is a set of continual links interconnecting two different nodes. A multipath network is a network in which there is more than one path between any pair of nodes. For example, in Fig. 1, the route linking nodes X and Y consists of multiple paths. With more paths, nodes have more options for communicating with one another, potentially increasing scalability, reliability, and link load balancing. Examples of multipath network topologies include DCell [1], BCube [2], and Fat Tree [3], as well as the flat-mesh architecture, an Ethernet fabric [4], for example. These topologies are an improvement over the traditional hierarchical tree topology, in which there is only a single path between any pair of nodes in the network, and so only basic connectivity is provided. The use of multiple paths simultaneously provides aggregated capacity, which is useful for applications that demand high bandwidth, such as virtual machine (VM) migration, eScience, and video. Aggregated capacity is the total capacity of all paths linking a pair of nodes. In this paper, the term "flow" refers to a logical connection between a pair of endpoints, and consists of packets sent from a source node to a destination node.

The main challenges in DCN are maximizing network utilization and ensuring fault tolerance to address multiple node and link failures. VL2 [5], TRILL (IETF RFC 5556) [6], and SPB (Shortest Path Bridging IEEE 802.1aq-2012 [7]) use the Equal-Cost Multi-Path

* Corresponding author.
E-mail addresses: tsubedi@synchromedia.ca (T.N. Subedi), knguyen@synchromedia.ca (K.K. Nguyen), mohamed.cheriet@etsmtl.ca (M. Cheriet).

http://dx.doi.org/10.1016/j.comcom.2014.12.006
0140-3664/

(ECMP) to spread traffic across multiple paths. ECMP [8] balances the load across flow-based paths by calculating a hash of every packet header, but uniquely mapping a flow to a single path to prevent out-of-order delivery at the destination. For example, a flow between nodes X and Y (Fig. 1) can be mapped to either the X–a–b–Y or X–e–f–Y path. Thus, a single flow's throughput is limited to single path capacity, not to aggregated path capacity. Although there are many flows in a network, they are not always mapped to the right paths because of hashing collisions. The more links a flow traverses, the more collisions will occur [9]. With ECMP, the overall throughput is not optimal.

The MPTCP with OpenFlow [10] provides a Layer-2 (L2) multipath using multiple subflows as end TCP processes and mapping subflows to VLANs, depending on ECMP hashing. For example, when MPTCP uses 5 subflows to communicate between nodes X and Y (Fig. 1), ECMP hashing can choose both the X–a–b–Y (1 Gbps) and X–e–f–Y (1 Gbps) equal paths with a certain probability (e.g. 95%). In the case of a failed (a, b) link, an unequal path X–a–c–b–Y (1 Gbps) will not be used, which means that the ECMP only provides a single path bandwidth of 1 Gbps instead of available aggregated bandwidth of 2 Gbps.

A multitenant and highly dynamic virtualized environment consists of a large number of end-stations, leading to a very large number of flows that challenge the scalability of a solution to network throughput maximization. The challenges are scalability, in terms of address learning, forwarding decision convergence, and forwarding state size, as well as flexibility for workload migration with VM migration; for example, Ethernet address learning by flooding and remembering the ingress port restricts the topology to a cycle-free tree. In forwarding along a tree, switches near the root require more forwarding entries (TCAM).

In this paper, we propose an adaptive multipath routing architecture that takes advantage of in-network multipath mechanisms and provides transparent service to end-stations. In addition, our solution will address the asymmetric link bandwidth issue (as shown in Fig. 1, links may have different capacities), which has never been considered in recently proposed symmetric topologies such as BCube [2] and DCell [1]), as they were both designed with the same capacity in all their links. Our solution allows a flow between nodes X and Y (Fig. 1)) to achieve the aggregated capacity of 2 Gbps along paths X–a–b–Y and X–e–f–Y. In the case of a failed (a, b) link, the flow still achieves the aggregated capacity of 2 Gbps along the unequal paths X–e–f–Y and X–a–c–b–Y. The main contributions of this paper are the following:

- an adaptive multipath routing (AMR) architecture, which dynamically adapts to network states,
- a central application that proactively provisions loop-free multiple paths at network level,
- a solution for out-of-order packet delivery and ensuring per-flow aggregated capacity on multiple paths by applying path capacity-based weighted probabilistic link selection with caching in switches and admission control only at ingress switches,
- a scalable in-network multipath solution for end-stations in a multitenant dynamic virtualized environment by applying the reactive encapsulation of end-station flows to edge switches,
- scalable routing and forwarding solutions, by dividing a large topology into multiple administrative domains and using the prefix MAC as the flow rule.

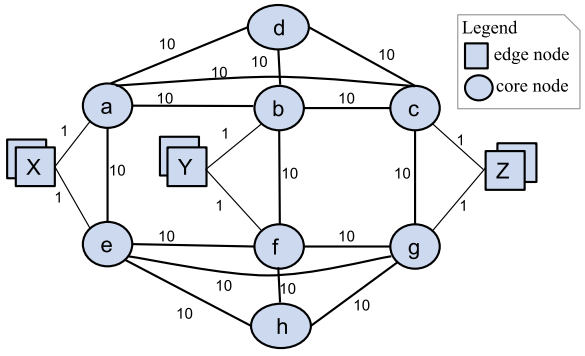

Fig. 1. Multipath topology example.

This paper is organized as follows. In Section 2, we present related work on the multipath concept in the DCN context. In Section 3, the AMR architecture is defined and a controller application is presented that proactively provisions multipath forwarding on OpenFlow switches, based on the proposed multipath algorithm. In Section 4, we describe a link selection algorithm on switches. In Section 5, we show how edge switches map an ingress flow to multiple paths. In Section 6, we describe the scalability of the solution in a large topology. In Section 7, we evaluate our proposed model, in terms of aggregated capacity, bisection bandwidth utilization, forwarding table size, and convergence time. Finally, we conclude the paper and present future work in Section 8.

2. Related work

The current Layer-3 (L3)-routed approach assigns IP addresses to hosts hierarchically, based on their directly connected switch. For example, hosts connected to the same Top of Rack (ToR) could be assigned the same /26 prefix, and hosts in the same row may have a /22 prefix [11]. With such an assignment, the forwarding tables across all data center switches will be relatively small. So, using multiple L2-switched domains and an L3-routed network for IP routing between them is a scalable addressing and forwarding solution. However, configuration and operational complexity are increased in the case of VM migration across L2 domains. VL2 [5] solves this problem and provides virtual L2 service in an L3-routed network by using IP-in-IP as the location separation mechanism and agent/directory service that follows end-system-based address resolution and takes advantage of a scalable L3 design. However, VL2 relies on ECMP, calculated by OSPF in L3 routers, which cannot use multiple paths for a flow.

One of the challenges in L2-switched network deployment in current DCNs is that the spanning tree protocol (STP) will prune paths from the network to ensure a loop-free topology, resulting in a single-tree topology [6]. Moreover, STP effectively wastes much of the potential throughput between any pair of nodes [6], and so a physical multipath design will not be fully exploited, which means that DCN is not scalable. There is growing interest to eliminate STP in L2 networks and enable multipath use in switching networks. There have been several improvements giving multiple STP instances, that is, multiple trees in a network. For example, Cisco's Per-VLAN Spanning Tree (PVST) [12] creates a separate spanning tree for each VLAN in a multi-VLAN network, and the IEEE 802.1s MST (Multiple Spanning Tree) [13] links multiple VLANs into a spanning tree, creating multiple trees in a network. The drawback of the multi-VLAN approach is resource fragmentation and under-utilization [14], because VM consolidation cannot be achieved between different VLANs.

Link aggregation (IEEE 802.3ad) [15] combines multiple links to create a single logical connection between two directly connected endpoints and increases bandwidth. However, this solution does not deal with links traversing multiple switches. There are proprietary multi-chassis Etherchannel (MEC) solutions, VSS, vPC, and MLAG, for example, which allow link aggregation towards different