Computer Communications 53 (2014) 1–12

Contents lists available at ScienceDirect

Computer Communications

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

Virtual machine placement with two-path traffic routing for reduced congestion in data center networks

Renuga Kanagavelu [a,*], Bu-Sung Lee [b], Nguyen The Dat Le [a], Luke Ng Mingjie [a], Khin Mi Mi Aung [a]

[a] Data Storage Institute, A*STAR (Agency for Science and Technology), Singapore
[b] School of Computer Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 29 November 2013
Received in revised form 9 July 2014
Accepted 19 July 2014
Available online 27 July 2014

Keywords:
Data Center
Virtual machine placement
Traffic splitting
Congestion
Protection

ABSTRACT

Virtualization-based Data Centers are increasingly becoming the hosting platform for a wide range of applications. The communication patterns in Data Center networks show the trend towards increasing bandwidth usage between virtual machines (VMs) within the Data Center resulting in higher chance of occurrence of network congestion. Thus, VM placement and routing algorithms are increasingly important to maximize application performance, provide fault tolerance, and reduce network loads. A less-than optimal placement of communicating VMs can cause inter-VM traffic to traverse bottlenecked network paths leading to large cross network traffic. The core network oversubscription and unbalanced workload placement could lead to long-lived congestion in Data Center networks. Multipath routing with traffic distributed in an appropriate proportion helps balance the load and decrease the possibility of congestion. Furthermore, by routing traffic on multiple link-disjoint paths, traffic can be protected against failures. The use of link-disjoint paths ensures the availability of at least one path for the traffic upon a link failure, thus guaranteeing a certain bandwidth (associated with the surviving paths). In this paper, we study the problem of VM placement with traffic routing on multiple paths for reduced occurrence of congestion while satisfying a certain protection grade which is defined as the fraction of rate (or bandwidth) guaranteed to be available in the event of single link failures. We develop an efficient algorithm based on a greedy technique for placing VMs onto servers satisfying the computing and memory resource requirements, taking into account the amount of inter-VM traffic and network load. In addition, we develop a two-path routing algorithm to satisfy the bandwidth and protection grade requirements so as to reduce the network congestion. Our simulation results show the effectiveness of the proposed algorithms in balancing the load and resilient when compared to other first-fit and random algorithms.

1. Introduction

Data centers host a wide variety of applications including web hosting, video services, e-commerce and social networking. In recent years, we have seen the traffic pattern changing from "North–South" to "East–West". This is due to the increased usage of resources in Data Centers to run large-scale data-intensive tasks, such as indexing web pages or analysing large data-sets, often using variations of the Map Reduce paradigm. To provide support for large number of applications, Data Centers require high performance network interconnect to connect tens of thousands of servers. Conventional Data Centers follow to a great extent a common a 2/3-tier single-rooted network topology. Using single path routing in this type of networks cannot fully utilize the network capacity, leading to congestion on the oversubscribed links and underutilizing the resources on other available paths. In such networks, it is critical to employ effective load balancing schemes so that the bandwidth resources are efficiently utilized.

To overcome the problems like poor bisection bandwidth and poor performance isolation, inherent in the traditional Data Centers, new network architectures such as VL2 [1], Portland [2], and BCube [3] have been proposed. These topologies take the form of multi-rooted trees with one or multiple paths between hosts. Optimized routing over multiple paths helps improve link utilization and decrease congestion. Load balancing could be done using multi-path routing by splitting the traffic between a source–destination pair across multiple disjoint paths. While splitting the traffic among different paths makes the network more reliable and load balanced, it requires intelligent mechanisms to choose good paths and traffic-splitting ratios among different paths.

* Corresponding author.
E-mail addresses: renuga_k@dsi.a-star.edu.sg (R. Kanagavelu), ebslee@ntu.edu.sg (B.-S. Lee).

http://dx.doi.org/10.1016/j.comcom.2014.07.009
0140-3664/

Virtualization is being deployed in Data Centers at a rapid pace to consolidate workloads for improved server utilization, for ease of provisioning, configuration management, and more generally, for efficient and flexible use of Data Center resources. It is highly desirable that VMs be placed to maximize application performance, power efficiency, fault tolerance, and reduce network bandwidth usage. Bi-section bandwidth is a critical resource in today's Data Centers because of the high cost and limited bandwidth of higher-level network switches and routers. This problem is aggravated in virtualized environments where a set of virtual machines, jointly implementing some service, may run across multiple layer-2 (L2) hops. Poor placement of communicating VMs can cause inter-VM traffic to traverse bottlenecked network paths leading to unnecessary cross network traffic. In recent years, VM placement problem has received much attention from the researchers. Some works considered the placement of VMs with the constraints of limited resources of servers. The works carried out in [20–22] addressed the VM placement problem considering only the server resources such as CPU and memory, while leaving out networking aspects. In [19,23], network-aware VM placement has been studied. In [24], VM placement with its impact on network traffic has been considered but with the main goal of reducing the total energy consumption in a Data Center.

Since Data Centers carry huge amounts of traffic, component faults have severe consequences. Fault tolerance or survivability is thus critically important to provide reliable services. If we make advance resource reservation, with sufficient backup bandwidth resources, we can provide 100% traffic protection. Since reserving backup resources is expensive, for a cost-effective solution, traffic can be protected partially wherein traffic will get reduced bandwidth in the event of failures. Partial protection ensures service availability in the event of failures but with reduced bandwidth and low performance which could be acceptable for most applications. We use protection grade as a measure of partial protection. We define protection grade as the fraction of bandwidth which is guaranteed to be available in the event of a single link failure. For example, if b units of bandwidth are guaranteed for a flow during the normal working condition and only b' ($\leqslant b$) units are available upon a link failure, the protection grade provided is said to be b'/b. We note that unlike the failures in the wide area networks (where cables are usually laid under the ground and sea), failures inside Data Centers can be restored. Therefore, partial protection is a cost effective solution for providing reliable services (with little or no extra bandwidth) although full bandwidth is not guaranteed in the event of faults.

Traffic splitting helps to reduce congestion and ensure protection in the event of failures. Traffic flows with protection requirements can be split and sent across two or more paths. It is desirable to split large flows when compared to short flows. This is because, large flows are more likely to cause link congestion. Further, they are small in number, implying lesser implementation overhead due to traffic splitting. When the number of flows to which a flow is split increases, the protection grade increases, but the algorithmic complexity and implementation overhead due to traffic splitting also increase. Suppose that a flow which requires a certain protection grade and bandwidth requirement of b units is split into two flows and routed through two link-disjoint paths each with b_1 and b_2 units, respectively. The use of link-disjoint paths guarantees that at least one of the two paths is available in the event of a single link failure. This guarantees the availability of $\min(b_1, b_2)$ units of bandwidth upon a link failure implying a protection grade of $\min(b_1, b_2)/b$. If the flow is split and sent across n link-disjoint paths and $b_{\max}$ is the maximum bandwidth used among these n paths, then the minimum protection grade guaranteed is $[b - b_{\max}]/b$ which happens when a link on the path with $b_{\max}$ units has failed. We can provide 100% protection guarantee by reserving spare bandwidth of $b_{\max}$ units on a backup path. However this requires excessive resources as the total bandwidth used is $b + b_{\max}$ units.

In this paper, we study the problem of VM placement with traffic splitting for reduced congestion and partial traffic protection. We consider flow demands with a specified computing, memory, and bandwidth resource requirements and also a protection grade requirement. We develop an efficient greedy approach based algorithm called Greedy VM placement with Two Path Routing (GVMTPR). The algorithm chooses appropriate servers for placing VMs from a set of candidate servers that can possibly satisfy the server resource requirements. It splits flows into two and route them through two link-disjoint paths so as to reduce congestion while satisfying the bandwidth and protection grade requirements. We use the maximum load on any link as a measure of congestion. We note that the path lengths are short in Data Center networks and there exist many paths of the same length, and therefore the impact of traffic splitting across two paths in less pronounced. Further, our work considers the current network state for multipath routing unlike the Equal Cost Multipath (ECMP) algorithm which chooses a path from among multiple paths without considering the existing utilization of the links. We demonstrate the effectiveness of our proposed algorithm through simulation results.

The rest of the paper is organized as follows. In Section 2 we describe the background and related works. In Section 3, we present the VM placement routing framework and explain various functional units. We formulate the problem in Section 4. We present and discuss the proposed algorithm in Section 5. We carry out performance study through simulations in Section 6. Finally, we make concluding remarks in Section 7.

2. Background and related works

In this section, we briefly review previous works related to this work, namely Data Center networking, then explain the various VM placement Heuristics.

2.1. Data Center networks and congestion control

The goal of Data Center networks is to interconnect a large number of servers with networking switches and high-speed links to offer high aggregate bandwidth. Conventional Data Centers are typically organized in a three-tiered network hierarchy. Servers are organized in racks and each server in a rack is connected to a Top-of-Rack (ToR) switch. ToR switches are connected to aggregation switches. Aggregation switches are connected to a few core switches in the top-tier. Such a hierarchical design faces multiple problems in scaling-up with a number of servers like over-subscription of the network links with increasing number of servers, limited inter-sever capacity, fragmentation of resources and poor resource utilization. To address these issues, new DCN architectures can use either of the following two basic topologies: (a) switch-centric topology such as VL2 [1], Portland [2] (b) server-centric topology such as BCube [3]. While the switches participate in routing and forwarding in the case of switch-centric topology, servers with network interfaces also participate in addition to the switches in the case of server-centric topology. These topologies take the form of multi-rooted trees with one or multiple paths between hosts. Data Center network architectures like Tree in [4–6], Portland in [2], VL2 in [1], and DCell in [7], that follow three-tier architecture have been proposed. Recently, new multi-level recursive network architecture, BCube, has been proposed in [3]. These architectures together with their own routing algorithms have been designed independently with different goals in mind. A