Computer Networks 91 (2015) 528–547

Contents lists available at ScienceDirect

Computer Networks

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

Survey Paper

A survey on data center networking for cloud computing

Bin Wang [a,*], Zhengwei Qi [a], Ruhui Ma [a], Haibing Guan [a], Athanasios V. Vasilakos [b]

[a] School of Software, Shanghai Jiao Tong University, Shanghai, China
[b] Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Sweden

ARTICLE INFO

Article history:
Received 30 October 2014
Revised 2 June 2015
Accepted 27 August 2015
Available online 10 September 2015

Keywords:
Cloud computing
Data center networking
Virtualization
Software defined networking

ABSTRACT

Data Center Networks (DCNs) are an essential infrastructure that impact the success of cloud computing. A scalable and efficient data center is crucial in both the construction and operation of stable cloud services. In recent years, the growing importance of data center networking has drawn much attention to related issues including connective simplification and service stability. However, existing DCNs lack the necessary agility for multi-tenant demands in the cloud, creating poor responsiveness and limited scalability. In this paper, we present an overview of data center networks for cloud computing and evaluate construction prototypes based on these issues. We provide, specifically, detailed descriptions of several important aspects: the physical architecture, virtualized infrastructure, and DCN routing. Each section of this work discusses and evaluates resolution approaches, and presents the use cases for cloud computing service. In our attempt to build insight relevant to future research, we also present some open research issues. Based on experience gained in both research and industrial trials, the future of data center networking must include careful consideration of the interactions between the important aspects mentioned above.

1. Introduction

Cloud services have increased rapidly in number and scale across many application areas. Recent reports showed that the cloud service market reached a size of $ 148.9 billion by 2014 [1]. As of today, a simple web search request may touch 1000+ servers, while a large computing request sometimes consumes thousands of machines [2]. Cloud computing requires the underlying network to be fast, carry large amounts of traffic, and be scalable. Building such complex environments using traditional networks is time-consuming and expensive.

Data center is a set of servers, storage and network devices, power systems, cooling systems, etc [3]. Data centers are intended for large-scale service applications such as online businesses, Smart Grid [3,4] and scientific computation [2]. Data Center Network (DCN) includes data center and provides the connections to the data center, which is described by its network topology, routing/switching equipment, and the protocols it uses [3,4]. DCN offers many features to help organize cloud computing for the following reasons:

- DCN permits the connection of thousands of data center servers in an efficient way, so that cloud computing could simply expand its service by following the DCN topology.
- DCN offers traffic reliability and efficiency to massive machine-to-machine communications in which activities from cloud computing emerge as the workloads distributed on data center servers.
- DCN supports various virtualization techniques that help DCN to create Virtual Machine (VM) [5], virtual network, and virtual function. DCN should possess the scalability to provide isolation and migration to massive numbers of virtual instances.

* Corresponding author. Tel.: +8618516199503.
E-mail addresses: binqbu2002@sjtu.edu.cn (B. Wang), qizhwei@sjtu.edu.cn (Z. Qi), ruhuima@sjtu.edu.cn (R. Ma), hbguan@sjtu.edu.cn (H. Guan), th.vasilakos@gmail.com (A.V. Vasilakos).

http://dx.doi.org/10.1016/j.comnet.2015.08.040
1389-1286/

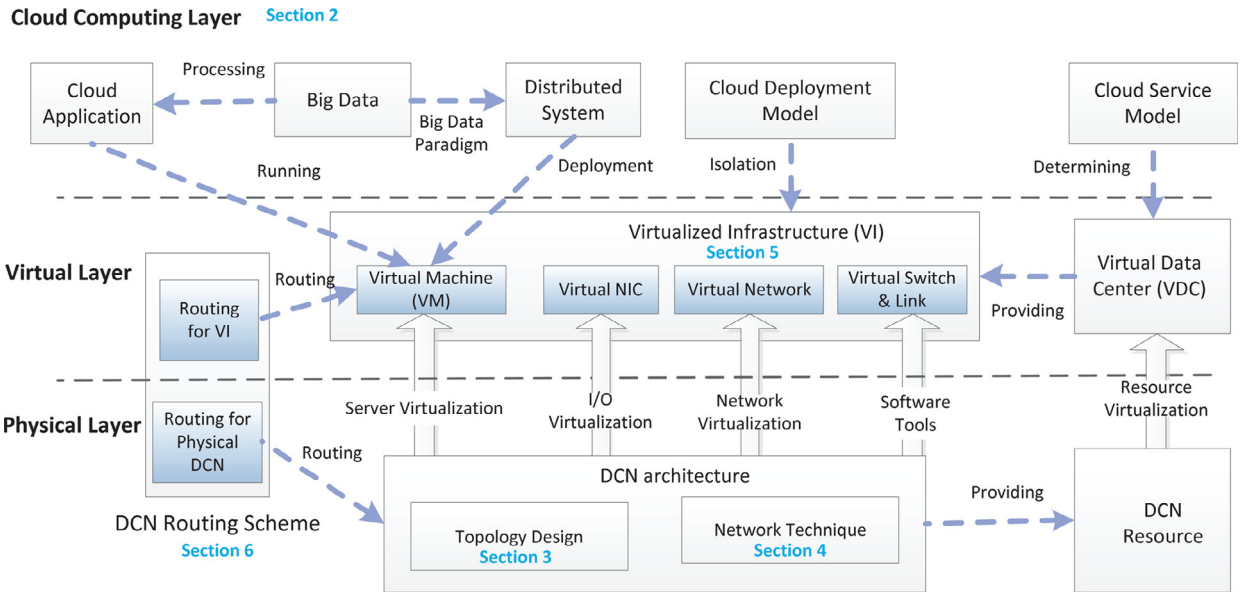

Fig. 1. Cloud computing network architecture.

- Existing research on DCNs has generated solutions for many use cases such as green computing, data center backup, some of which can also address the challenges of cloud service.

1.1. Cloud computing and DCN

Fig. 1 depicts cloud computing architecture that illustrates the separation of cloud computing layer, virtual layer and physical layer. The physical layer is mainly responsible for offering hardware resources such as CPU, memory, and bandwidth, and for enforcing routing and traffic management in a real-world DCN. Activities from the other two layers become workloads, where the high-level activities involve information being tunneled and transmitted by various physical agents, e.g., server, switch, or middlebox. Virtual layer offers virtual instances such as VM and virtual network to the cloud users. In this layer, a virtual resource is allocated according to various policies, which serve the cloud service model. The last layer is the cloud computing layer, which arranges all the cloud activities on top of the virtual layer. In this paper, cloud activities mainly include cloud service model, cloud deployment model, big data [6], and cloud distributed system. All these activities make great demands on the other layers offered by DCN in terms of traffic efficiency, on-demand allocation, and reliability.

1.2. DCN performance in cloud computing service

We will briefly discuss the performance that remains for existing DCN techniques. The key findings of this paper are:

- Most DCN architectures are not optimally designed for use with cloud computing. For example, the tree-based DCNs, such as Fat-tree or Three-tier [7–9], are the most deployed architectures in the enterprise cloud. However, these architectures lack scalability in terms of load balancing and rate control, which make them poor at managing traffic from thousands of servers and VMs. On the other hand, the newer architectures for cloud computing are not effective at addressing or topology self-learning, and as such are currently not used as the DCN architectures for cloud computing.
- Scalability is the most constraining problem for DCN solutions that support cloud computing. In particular, it constrains both centralized and distributed approaches to routing, control plane management, resource allocation, etc. We believe that this problem could be solved with the collaboration of different network elements. For example, it is impossible for a centralized load balancing approach to collect and control every switch in a large-scale DCN, but a solution could reduce computational workloads in the centralized controller if some flows are decided locally in a commodity switch.
- DCN architectures are divergent in topology. This makes it difficult to manage general routing, virtualization, and load balancing in the higher layers. For example, many load balancers are fast at rerouting local flow [10]. However, since they are not aware of the switch locations or loads elsewhere, some DCN architectures with redundant path design will undermine the load balancing performance with a forwarding loop. On the other hand, DCN-specific solutions lack interfaces for Internet access. For example, some DCNs customize their routing protocols in a way that might not provide the interface for the Border Gateway Protocol (BGP) [11] at layer 3.

1.3. Objectives and organization

This paper presents a comprehensive overview of the existing data center networks for cloud computing. The goal of this paper is to understand the challenges and research