Chapter 2

Network-Aware Virtual Machine Placement and Migration in Cloud Data Centers

Md Hasanul Ferdaus
Monash University, Australia

Manzur Murshed
Federation University, Australia

Rodrigo N. Calheiros
The University of Melbourne, Australia

Rajkumar Buyya
The University of Melbourne, Australia

ABSTRACT

With the pragmatic realization of computing as a utility, Cloud Computing has recently emerged as a highly successful alternative IT paradigm. Cloud providers are deploying large-scale data centers across the globe to meet the Cloud customers’ compute, storage, and network resource demands. Efficiency and scalability of these data centers, as well as the performance of the hosted applications, highly depend on the allocations of the data center resources. Very recently, network-aware Virtual Machine (VM) placement and migration is developing as a very promising technique for the optimization of compute-network resource utilization, energy consumption, and network traffic minimization. This chapter presents the relevant background information and a detailed taxonomy that characterizes and classifies the various components of VM placement and migration techniques, as well as an elaborate survey and comparative analysis of the state of the art techniques. Besides highlighting the various aspects and insights of the network-aware VM placement and migration strategies and algorithms proposed by the research community, the survey further identifies the benefits and limitations of the existing techniques and discusses the future research directions.

1. INTRODUCTION

Cloud Computing is a recently emerged computing paradigm that promises virtually unlimited compute, communication, and storage resources where customers are provisioned these resources according to their demands following a pay-per-use business model. In order to meet the increasing consumer demands, Cloud providers are deploying large-scale data centers across the world, consisting of hundreds of thousands of servers. Cloud applications deployed in these data centers such as...
web applications, parallel processing applications, and scientific workflows are primarily composite applications comprised of multiple compute (e.g., Virtual Machines or VMs) and storage components (e.g., storage blocks) that exhibit strong communication correlations among them. Traditional research works on network communication and bandwidth optimization mainly focused on rich connectivity at the edges of the network and dynamic routing protocols to balance the traffic load. With the increasing trend towards more communication intensive applications in the Cloud data centers, the inter-VM network bandwidth consumption is growing rapidly. This situation is aggravated by the sharp rise in the size of the data that are handled, processed, and transferred by the Cloud applications. Furthermore, the overall application performance highly depends on the underlying network resources and services. As a consequence, the network conditions have direct impact on the Service Level Agreements (SLAs) and revenues earned by the Cloud providers.

Recent advancement in virtualization technologies emerges as a very promising tool to address the above mentioned issues and challenges. Normally, VM management decisions are made by using various capacity planning tools such as VMware Capacity Planner (“VMware Capacity Planner”, 2014) and their objectives are set to consolidate VMs for higher utilization of compute resources (e.g., CPU and memory) and minimization of power consumption, while ignoring the network resource consumption and possible prospects of optimization. As a result, this often leads to situations where VM pairs with high mutual traffic loads are placed on physical servers with large network cost between them. Such VM placement decisions not only put stress on the network links, but also have adverse effects on the application performance. Several recent measurement studies in operational data centers reveal the fact that there exists low correlation between the average pairwise traffic rates between the VMs and the end-to-end network costs of the hosting servers (Meng, Pappas, & Zhang, 2010). Also because of the heterogeneity of the deployed workloads, traffic distribution of individual VMs exhibit highly uneven patterns. Moreover, there exists stable per-VM traffic at large timescale: VM pairs with relatively heavier traffic tend to exhibit the higher rates whereas VMs pairs with relatively low traffic tend to exhibit the lower rates. Such observational insights of the traffic conditions in data centers have opened up new research challenges and potentials. One such emerging research area is the network-aware VM placement and migration that covers various online and offline VM placement decisions, scheduling, and migration mechanisms with diverse objectives such as network traffic reduction, bandwidth optimization, data center energy consumption minimization, network-aware VM consolidation, and traffic-aware load balancing.

Optimization of VM placement and migration decisions has been proven to be practical and effective in the arena of physical server resource utilization and energy consumption reduction, and a plethora of research contributions have already been made addressing such problems. Until recently, a handful of research attempts are made to address the VM placement and migration problem focusing on inter-server network distance, run-time inter-VM traffic characteristics, server load and resource constraints, compute and network resource demands of VMs, data storage locations, and so on. These works not only differ in the addressed system assumptions and modeling techniques, but also vary considerably in the proposed solution approaches and the conducted performance evaluation techniques and environments. As a consequence, there is a rapidly growing need for elaborate taxonomy, survey, and comparative analysis of the existing works in this emerging research area. In order to analyze and assess these works in a uniform fashion, this chapter presents an overview of the aspects of Cloud data center management as background information, followed by various state-of-the-art