Journal of Network and Computer Applications 56 (2015) 166–187

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Performance modeling and analysis of virtualized multi-tier applications under dynamic workloads

Keyvan RahimiZadeh [a,*], Morteza AnaLoui [a], Peyman Kabiri [a], Bahman Javadi [b]

[a] School of Computer Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran
[b] School of Computing, Engineering and Mathematics, University of Western Sydney, Parramatta 2751, NSW, Australia

ARTICLE INFO

Article history:
Received 30 January 2015
Received in revised form
23 May 2015
Accepted 27 June 2015
Available online 21 July 2015

Keywords:
Multi-tier application
Virtualization
Queuing network
Performance evaluation
Workload collocation

ABSTRACT

Virtual machine technology facilitates implementation of modern Internet services, especially multi-tier applications. Server virtualization aims to reduce the cost of service provisioning and improve fault tolerance, portability and security of virtualized services by sharing the resources amongst consolidated servers. Although designing applications based on multi-tier architecture brings benefits to service providers in terms of service flexibility, functionality, and reusability of software in comparison to traditional architecture, deploying multi-tier applications in virtualization environments faces many challenges. In addition to virtualization overhead and virtual machines interference, dynamic and unpredictable nature of submitted workloads makes the performance evaluation of virtualized multi-tier applications very complex task. In this article, we propose a novel analytical model based on queuing network to evaluate virtualized multi-tier applications. The efficiency of the proposed model is evaluated by using a series of in-depth experiments in different configurations to study the performance of multi-tier applications. The conducted experiments examine consolidation of RUBiS and Wikipedia tiers in different scenarios under both bursty and non-bursty workloads. We present the results based on quantitative and qualitative analyses that are useful for virtualized multi-tier applications deployment. The performance modeling and experimental results are in excellent agreement and confirm the efficiency of the proposed model to evaluate and analyze the performance of virtualized multi-tier applications.

1. Introduction

Modern multi-tiered applications need to deploy multiple services to achieve more flexibility of services and reusability of software in comparison to providing services based on traditional Web architecture. Instead of deploying entire application through traditional architecture, service providers can add more tiers or adjust tiers configuration by separating application functionality in separated tiers. This approach enables service providers to flex and reuse applications in a distributed fashion. For instance, Microsoft Azure Cloud Service as given in Turkarslan (2015) and Amazon Elastic Compute Cloud (EC2) as given in Tavis and Fitzsimons (2015) facilitate provisioning of multi-tier Web applications. Although traditional server implementation aims to guarantee the Service Level Agreement (SLA) by establishing such services on separate physical servers, this may result in underutilized resources. Multi-tier architecture has fundamental performance advantage over the traditional Web-based architecture in implementing such applications. For instance, this architecture supports more users with high workload intensity by increasing the capacity and the number of servers on each tier. In addition to consuming too much space and power, high expenditures of system management and maintenance are the common cost of service deployment in traditional server architecture. Therefore, service providers have motivated to employ virtualization technology for achieving business goals without sacrificing overall service performance.

In recent years, there has been an increasing interest in large datacenters to develop enterprise applications by means of virtualization technology. Virtualization technology is capable of reducing the cost of deploying huge number of servers previously running on multiple physical hosts by consolidating them into smaller number of physical hosts. This pioneer technology comforts deployment of services and helps to heighten Quality of Service (QoS) in terms of reasonable response time, high availability, and scalability in datacenters, grid, and Cloud computing environments (Toosi et al., 2011; Manvi and Krishna Shyam, 2014). Furthermore, it provides capabilities such as virtual machine (VM)

* Corresponding author at: School of Computer Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran. Tel.: +98 917 145 0850.
E-mail address: rahimizadeh@iust.ac.ir (K. RahimiZadeh).

http://dx.doi.org/10.1016/j.jnca.2015.06.020

migration to reduce downtime for migrated VMs (Jin et al., 2011; Quesnel et al., 2013), balance the submitted workloads amongst hosts, improve service fault tolerance and reduce the overall power consumption in datacenters (Buyya et al., 2010; Wasim Ahmad et al., 2015). Currently, giant Cloud providers such as Amazon Web Services, Google, Microsoft, Yahoo, Apple, Cisco, Citrix, IBM, etc. benefit from virtualization technology to improve speed, agility, and elasticity of Internet services.

Although server virtualization for multi-tier applications has several advantages, there are a number of challenges regarding implementation of such applications in virtualized environments. The first challenge is that hosted application, which is running on virtualized operating systems, is not informed about its execution on the virtualized environment. Hypervisor intercepts virtualized operating system instructions and conveys them to the physical resources. This mediation incurs overhead and might degrade performance of running applications.

Second, operational interferences of VMs may diminish isolation and reduce the performance of the VMs (Chen et al., 2007b; Cherkasova et al., 2009; Zhu and Tung, 2012); however, the hypervisor aims to provide an isolated operational environment for the hosted VMs. Additionally, hypervisors implementation complexities and interactions between virtualized system layers including application, operating system (OS), hypervisor, and physical resources may drastically permute VMs isolation. Nevertheless, hypervisor layer provides well-defined interfaces to facilitate communications and hide implementation complexities of VMs. Hence, designing a proper hypervisor and a resource management scheme for consolidated applications are a vital task (Younge et al., 2011; Ku et al., 2014).

Third, virtualization vendors have promised that most applications could run in virtualized environments where application suppliers are able to certify the products as virtualization-ready products. However, some applications inherently are not good candidates to run in such environments. Performance of workloads with high resource demands, especially in *I/O* bound workloads, may deteriorate when they reside in a shared environment (Pu et al., 2013). In order to consolidate applications from traditional servers to virtualized platforms, it is critical to predict the performance of hosted applications in virtualized environments. In this way, efficient resource configuration and workload placement regarding workload idiosyncrasies and hypervisor capabilities can mitigate performance loss.

Fourth, dynamic and unpredictable nature of submitted workloads (e.g. workload burstiness), hardens performance evaluation and prediction of virtualized multi-tier applications (VMTAs). Accordingly, it is difficult to guarantee SLA wherein resources are statically allocated to multi-tier applications in such circumstances. Eventually, performance evaluation and modeling of multi-tier applications is more difficult than single-tier applications because tiers are not identical and their resource demands are different. Furthermore, the logical and operational dependencies between tiers can shift performance bottleneck of one tier to other tiers (e.g. successor tier to corresponding preceding tier(s)), which consequently deteriorates overall application performance.

Given the aforementioned issues, it is crucial to investigate a substantial performance evaluation, resource configuration, and workload scheduling to alleviate performance loss of multi-tier applications in virtualized datacenters. Finding a solution that comprehensively covers all the mentioned issues is troublesome.

In this paper, an analytical model based on queuing network (QN) is proposed to evaluate the performance of VMTAs considering interactions between tiers. We also investigate performance metrics of hypervisor and hosted tiers individually and analyze the overall application performance. Following this, we validate the proposed model in different configurations to capture the performance of two well-known applications. The conducted experiments examine consolidation of variety of workloads including realistic RUBiS and Wikipedia traces. We use the former multi-tier application as an open source multi-tier *e*-commerce application that represents core functionality of *e*Bay auction site and the latter as a well-known Web-based encyclopedia workload. We study system under test (SUT) when it hosts combination of RUBiS and Wikipedia tiers as *I/O* and CPU extensive workloads, respectively. Additionally, this paper investigates the effect of workload burstiness on overall performance of considered applications in virtualized environment.

The central theme of this paper is to model and evaluate the performance of the disparate VMTAs subject to proximity of their tiers in different configuration scenarios. We propose a novel closed-form analytical model to predict the request response time of collocated VMTAs as a function of workload characteristics (request rate and demands on resources), tiers functionalities, and proximity of VMTAs' tiers. Our proposal introduces in-detail modeling and in-depth evaluation of VMTAs' tiers proximity. We present the results based on quantitative and qualitative analyses that are useful for deployment of VMTAs and placement of corresponding tiers in a virtualized environment. In summary, the main contributions of this paper are as follows:

- We propose an analytical model based on queuing network to estimate the aggregated QoS metrics of multi-tier applications in a virtualized datacenter.
- We introduce a methodology to measure the VMTA characteristics such as workload rate and demands on individual tiers, the transition probabilities of requests between tiers, and applications performance metrics.
- We conduct a series of in-depth experiments in different scenarios to evaluate the performance of consolidated VMTAs' tiers in the presence of gradual and bursty workloads.

The reminder of this paper is structured as follows. Section 2 briefly reviews virtualization technology and the common architecture of multi-tier applications. We follow this by surveying the related work in Section 3 and describe proposed QN modeling to evaluate VMTAs in Section 4. Measurement methodology, workload description, and implementation considerations to evaluate VMTAs are introduced in Section 5. It will then go on to present analytical and experimental results in Section 6. Section 7 concludes the paper with review of contributions, explains how closely this work has reached its goals, and brings forth areas for further researches and future works.

2. Background

This section exposes a brief overview of virtualization technology and multi-tier application architecture.

2.1. Virtualization technology architecture

In recent years, virtualization technologies are receiving renewed attention in both industry and academia. Virtualization technology aims to provide an environment where VMs can efficiently operate and effectively share the system resources to meet hosted applications QoS with minimized operational cost. Each VM hosts a whole software stack including operating system, middleware, and applications. In fact, virtualization appends an abstraction layer to the computer system architecture for multiplexing hardware resources amongst hosted VMs. This layer is Virtual Machine Monitor (VMM), also called hypervisor[1], emulates

[1] We are using the words hypervisor and VMM, interchangeably.