Future Generation Computer Systems 49 (2015) 68–76

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

High reliable real-time bandwidth scheduling for virtual machines with hidden Markov predicting in telehealth platform

Jihe Wang [a], Meikang Qiu [b], Bing Guo [a,*]

[a] Computer Science College, Sichuan University, Chengdu, SC, 610064, China

[b] Pace University, Seidenberg School of Computer Science and Information Systems, Manhattan, NY 10038, USA

HIGHLIGHTS

- We design a service-oriented architecture (SOA) for telehealth service in cloud.
- We propose a data coherence protocol among distributed Dcs to measure the bandwidth demand.
- A HMM-based bandwidth predicting algorithm is designed to help cloud broker manage the network resource effectively.
- A bandwidth allocating module is implemented for bandwidth allocating and recycling.
- The reliability and sensitivity of the model are tested, and the performance is much better than other four traditional predicting approaches.

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form
1 August 2014
Accepted 15 August 2014
Available online 16 September 2014

Keywords:
Reliable
Telehealth
Bandwidth
HMM
Prediction
Data center

ABSTRACT

Reliable and high-performance resource scheduling for *Virtual Machines* (VMs) in cloud can guarantee the efficiency of remote rescue with telehealth system. When a local disaster, e.g. earthquake and tsunami, happens in a densely populated area, the surging health care demand leads to the increasing workload in *Data Centers* (DCs) by storing and delivering a mass of patients' information and real-time physiology signals. However, the current self-adaptive scheduling methods cannot provide a high-accuracy recognizing of the two conditions: *urgency* or *normal*, which would procrastinate the system into a high-performance status, while the best rescue time is lost. In this paper, we propose a *Primary Node*-based architecture for typical telehealth service on cloud, which takes into account both storage and delivery efficiency. We also design a novel algorithm to predicting and allocating the future bandwidth of all VMs in the telehealth service context. This method is able to dynamically adjust each parameter of a *Hidden Markov Model* (HMM) through collecting the historical information of the bandwidth workload. After we predict the future bandwidth consumption of VMs, a high-performance scheduling method is used to adjust the bandwidth to each VM for health care service. The simulation results prove that this algorithm provides a high-accurate prediction, which guides the allocating module to make decision before the request burst comes. Nevertheless, our algorithm improves the reliability of telehealth services for storing and delivering patients' information among DCs.

1. Introduction

After the major earthquake disasters in Haiti and Japan, the high demand for international collaboration on rapid emergency response and disaster management put the study of emergency treatment into a new perspective [1]. During one emergency rescue, there is a large amount of demand for the remote help from health care service system, which includes personal health information system, health risk assessment system, guidance system, etc. As a cloud-based information and expert system, telehealth system can provide reliable and personalized services to each patient. On the *SaaS* (Software-as-a-service) level, the telehealth system is able to refer the historical and current records to decide how to deal with patient's emergent situation, especially for those chronic patient. On the *PaaS* (Platform-as-a-Service) level, the resources offered by *Data Centers* (DCs), such as CPU time, bandwidth, and storage space, could be rented to service providers, which deploy their services on DC [2]. In order to reduce the power consumption of DCs, machines would work in idle status to save

* Corresponding author. Tel.: +86 13980664852.
E-mail addresses: wangjihe46@gmail.com (J. Wang), mqiu@pace.edu (M. Qiu), guobing@scu.edu.cn (B. Guo).

http://dx.doi.org/10.1016/j.future.2014.08.006
0167-739X/

electricity fee, while the performance also shrinks. In the low-performance status, the could system has a relative lower bandwidth and resource utilization, less availability, and longer waiting and service time for VMs [3]. Thus, how to predict the service demands precisely and recover the system from the low-performance status quickly are the difficult issues in emergency rescue. Because of heterogeneous software/hardware components and complicated connection among them, the probability of failure improves [4]. Service reliability and availability arouse more attention [5].

Traditional reliability analyzing includes two different types of methodologies: *Measure-based Quantification* and *Model-based Quantification*. The measure-based method can approach higher accuracy. However, it is too expensive to determine a lot of parameters and configurations, which makes it not always possible for all system design [6]. The model-based one uses either *Non-State–Space Model* or *State–Space Model* to analyze the reliability of a complexity system. The non-state–space model uses *Reliability Block Diagrams* (RBDs) [7], *Reliability Graphs* (relgraphs) [8], and *Fault Trees* (FTs) to effectively solve for system reliability, system availability, and system mean time to failure. The large scale problem and the endless analyzing of min-path between course and consequence events result in exponential growth of graph.

The state–space one is able to model complexity interactions between components, such as Markov chains. Many examples of dependences among system components have been observed in practice and captured by *Continuous-Time Markov Chains* (CTMCs) [9].

Data moving is a random event in the distributed cloud system, which is driven by the applications running on machines, called VMs. For example, in the MapReduce framework [10], the map workers are distributed in a cluster of machines. Each map worker is fed with a chunk of data from the local or remote disk. This computing architecture would lead to the fact that, when there is a mass of workers running parallel on the cloud system to process data, data chunks would be delivered through the cloud internal network. A promising telehealth system should focus on providing reliable and available services to users, where, the data delivering is much more complicated. On one hand, medical record, i.e., *Personal Health Record* (PHR), is always quite sensitive for both patients and hospitals. Before accessing those information, very strict verification are required on all service levels so that patients' information is only delivered to the persons concerned [11]. On the other hand, both hardware and software limitation would reduce the performance of telehealth service. We focus on the bandwidth allocation for virtual machine on cloud, which is one of the most difficult issues in resource scheduling. Unlike the previous studies on component interactions, we would like to provide a high-reliable bandwidth allocating service to VMs, which uses a lightweight measure-based method to evolve each parameter of a Hidden Markov Model (HMM), and uses this model to predict the bandwidth consumption in future. The important contributions of this paper are:

- a high-reliable *SOA* (Service-oriented Architecture) is designed for telehealth service in cloud, which includes a *Primary Node*-based protocol to implement PHR delivery among distributed data centers;
- we use HMM to build a status-deciding algorithm about the bandwidth allocations for VMs. Based on the low overhead of sampling, this module is able to analyze the future status of the system, which will be an important condition for changing the serving status of the telehealth system;
- based on the algorithm, a smart bandwidth allocating module is implemented in the broker of data center, where, most of the system parameters can be obtained easily. This module automatically adjusts DC to the high-performance status as a heavy workload predicted by the algorithm. The responding time of our model can be reduced dramatically.

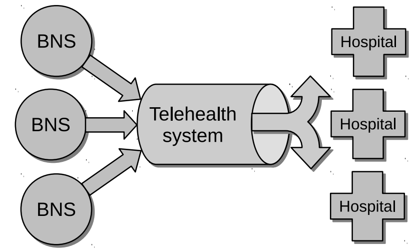

Fig. 1. The overall architecture of the cloud-based telehealth platform.

The rest of the paper is organized as follows. Section 2 presents our design of the reliable architecture for normal telehealth application. Section 3 presents a novel HMM-based bandwidth allocation algorithm, and a smart bandwidth allocating module is also introduced in this section. Section 4 is the simulation results of the algorithm. Sections 5 and 6 are some related work and conclusions, respectively.

2. System design

SOAs are an architectural paradigm for building software applications from a number of loosely coupled distributed services [12]. This paradigm adapts the computing architecture for different kinds of the services in cloud. Since it considers the efficiency and reliability of cloud system at the beginning of system design, telehealth is one of the ideal applications to be implemented in this way. Although extensive fault-tolerant scheduling algorithms have been proposed for real-time tasks in parallel and distributed systems, quality of service (QoS) requirements of tasks have not been taken into account [13]. In this section, we will give an overview of our SOA telehealth cloud system. And then, some primary techniques and system abstract model are presented.

2.1. Overall of telehealth cloud

As shown in Fig. 1, a telehealth system should include at least three parts. In the front-end, a mass of *Body Sensor Networks* (BSNs) subsystem keeps detecting patients' physiological signals and transporting them to cloud. In the middle-end, distributed data centers arouse a large number of VMs to provide services to each patient. At the same time, PHRs are updated in data centers, where, patient's new signals and information records are analyzed and stored in the distributed storage system. In the back-end, authorized hospitals, medical experts, and nurses can look up their patients' current and historical information so that some emergency decisions and treatments can be done.

BSN is a wireless network of wearable computing devices to monitor the health condition of body. A number of intelligent physiological sensors can be integrated into a wearable wireless body area network for computer-assisted rehabilitation or early detection of medical conditions. With the proposed BSN architecture, a number of wireless biosensors including 3-lead ECG, 2-lead ECG strip, and SpO2 sensors have been developed [14]. By adopting the IEEE 802.15.4 standards [15], sufficient bandwidth is available for demanding continuous physiological and context sensing. In our design, sensors keep producing real-time physical sampling data, and report the data to cloud through near hubs (PC or PDA).

In a telehealth system, there are a variety of formats representing patient's medical records, which is often outsourced to be stored at distributed cloud providers. However, there have been wide privacy concerns as personal health information could be exposed to those third party servers and to unauthorized parties. To assure the patient's control over access to their own PHRs, it is a promising method to encrypt the PHRs before outsourcing [16].