Simulation Modelling Practice and Theory 82 (2018) 147–159

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

# Modeling of correlated resources availability in distributed computing systems

Bahman Javadi*, Kenan M. Matawie

*School of Computing, Engineering and Mathematics, Western Sydney University, Australia*

ARTICLE INFO

*Article history:*
Received 8 September 2017
Revised 27 December 2017
Accepted 28 December 2017

*Keywords:*
Volunteer computing
Resource availability
Statistical modeling
Markov model
Time series

ABSTRACT

Volunteer computing systems are large-scale distributed systems with large number of heterogeneous and unreliable Internet-connected hosts. Volunteer computing resources are suitable mainly to run High-Throughput Computing (HTC) applications due to their unavailability rate and frequent churn. Although they provide Peta-scale computing power for many scientific projects across the globe, efficient usage of this platform for different types of applications still has not been investigated in depth. So, characterizing, analyzing and modeling such resources availability in volunteer computing is becoming essential and important for efficient application scheduling. In this paper, we focus on statistical modeling of volunteer resources, which exhibit non-random pattern in their availability time. The proposed models take into account the autocorrelation structure in individual and subset of hosts whose availability has temporal correlation. We applied our methodology on real traces from the SETI@home project with more than 230,000 hosts. We showed that Markovian arrival process and ARIMA time series can model the availability and unavailability intervals of volunteer resources with a reasonable to excellent level of accuracy.

## 1. Introduction

Volunteer computing systems are large-scale distributed systems with large number of heterogeneous and unreliable Internet-connected hosts. These platforms provide more than 10 PetaFLOPS computing power to more than 70 scientific projects in different areas such as astronomy, physics, mathematics and chemistry [1,2]. No doubt utilization of such resources is essential and growing as it provides immense computational power on the order of PetaFLOPS and storage on the order of PetaBytes at almost zero costs [3]. However, volunteer computing resources are only suitable to run High-Throughput Computing (HTC) applications due to their unavailability rate and frequent churn (several times a day). So characterizing, analyzing and modeling such resources availability in volunteer computing is an essential requirement to broaden the types of applications that can be executed in this system, which is the main goal of this paper.

The overall objective of the modeling is to use the free resources of volunteer computing systems for execution of scientific applications in form of many task computing workloads. Many task computing (MTC) is a new paradigm to bridge the gap between High-Throughput Computing (HTC) and High-Performance Computing (HPC) [4]. The structure of MTC applications can be considered as graphs of discrete tasks. These tasks can be different in terms of size, communication patterns and intensity. In most cases, the data dependencies among tasks are handled through file sharing that is a feasible way of

* Corresponding author.
*E-mail address:* b.javadi@westernsydney.edu.au (B. Javadi).

https://doi.org/10.1016/j.simpat.2017.12.017
1569-190X/© 2017 Elsevier B.V. All rights reserved.

tasks communication in volunteer computing platforms. However, in contrast to HTC applications, MTC applications have relatively short tasks (i.e., seconds to minutes long), so they need fast response time. Therefore, resources availability and their temporal structure is crucial and important for efficient scheduling of these applications.

In previous work, an analysis and methodology was proposed to form subsets of hosts with similar statistical properties that can be modeled with similar distribution functions [5]. This paper explained that about 21% of hosts exhibit *random* availability, which can be modeled with a few distinct distributions from different families. It was also shown how to apply the proposed models for stochastic scheduling of Bag-of Tasks (BoT) applications in a resource brokering context [6]. In this paper, we focus on statistical modeling of volunteer resources which exhibit *non-random* pattern in their availability time. To do this, we extended the existing methodology to further characterize, analyze and consider autocorrelation structure in modeling the subset of hosts whose availability has temporal correlation. Moreover, we are interested to find a host model as well as a system model to help the system to do more efficient task scheduling for MTC applications. In the host model, the behavior of a single host will be modeled, while for the system model, the behavior of all non-iid hosts will be modeled collectively. These two models can be adapted by the scheduler to optimize the specific criteria based on system and user requirements.[1]

We applied the proposed methodology on real traces from the SETI@home project with more than 230,000 hosts. We selected various statistical models that have the ability to fit traces with temporal dependencies at both host and system levels. We conducted the model fitting and analyzed the model complexity and accuracy through state reduction. We propose a queuing simulation technique to evaluate quality of the modeling. The results show that Markovian arrival process and ARIMA time series can model the availability and unavailability intervals of volunteer resources with a good degree of accuracy.

The rest of this paper is organized as follows. Related work is described in Section 2. In Section 3, we present the detail of modeling workflow and real traces. Section 4 includes how statistical models are selected. The model fitting and analysis of the model parameters are presented in Section 5. In Section 6, the model evaluation through simulation experiments is discussed. Conclusions and future work are presented in Section 7.

## 2. Related work

This section describes the related work in modeling and analysis of availability in volunteer computing systems. There are several research on collecting of real availability traces in volunteer computing platforms. Most of these studies are focused on host availability [8–10], which is different from CPU availability considered in this paper. CPU availability is defined as the time when a host's CPU is available to run the application as a volunteer resource. In other words, host availability might be a misleading metric as a host can be available but not its CPU.

Moreover, some papers only focused on volunteer resources in the enterprise or university [11,12] while we use real traces that includes hosts in the enterprise, university and home. Some studies such as [13,14] used availability traces of hundreds of hosts over a limited time period (e.g., a few weeks). In contrast, we study real traces of hundreds of thousands hosts over the period of 1.5 years.

There are many related work for availability modeling of volunteer systems, but most of them if not all did not take into account the temporal dependency of resource availability [15–17]. For instance, in [17], authors used the average availability as a distance metric to find cluster of hosts with the similar level of availability. So, the availability intervals were ignored as the goal was to find the correlated hosts. It has been shown that effective resource selection and scheduling is strongly depended on temporal structure of availability in such platforms [18,19]. Hence, we propose statistical models considering the autocorrelation structure in subset of hosts whose availability has temporal correlation.

There are very limited work for availability modeling at the host level and most of them are related to the CPU load modeling [20,21]. In [22], a forecasting approach based on vector autoregressive models and a tendency-based technique is proposed. In this approach, for each host, three different prediction will be examined and selected automatically and the prediction for the next hour will be provided. In contrast, we are looking at the modeling of the CPU availability and unavailability for the whole lifetime of the host using time series approaches.

In previous work, modeling and methodology to form subsets of hosts with purely random availability was introduced [5,6]. Clustering technique was also used to form groups of hosts that can be modeled with similar distribution functions. It was revealed that cluster formation by static criteria such as host location, time zone and CPU speeds can not have the same results as clustering by availability distribution. In other words, there is no correlation between host location, time zone and CPU speeds of host with the length of availability intervals. In contrast, we consider statistical modeling of volunteer resources, which exhibit non-random pattern with temporal correlation in their availability time.

## 3. Modeling methodology

In this section, we present the details of real traces as well as the modeling workflow used in this paper. The list of all abbreviations used in the paper is shown in Table 4.

[1] This paper is the extended version of [7].