Computer Physics Communications 185 (2014) 1355–1363

Contents lists available at ScienceDirect

Computer Physics Communications

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

Uncertainty quantification through the Monte Carlo method in a cloud computing setting

Americo Cunha Jr. [a], Rafael Nasser [b], Rubens Sampaio [a], Hélio Lopes [b,*], Karin Breitman [b]

[a] Department of Mechanical Engineering, PUC–Rio Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro - RJ, 22453-900, Brazil

[b] Department of Informatics, PUC–Rio Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro - RJ, 22453-900, Brazil

ARTICLE INFO

Article history:
Received 16 April 2013
Received in revised form
28 November 2013
Accepted 14 January 2014
Available online 23 January 2014

Keywords:
Uncertainty quantification
Cloud computing
Monte Carlo method
Parallel algorithm
MapReduce

ABSTRACT

The Monte Carlo (MC) method is the most common technique used for uncertainty quantification, due to its simplicity and good statistical results. However, its computational cost is extremely high, and, in many cases, prohibitive. Fortunately, the MC algorithm is easily parallelizable, which allows its use in simulations where the computation of a single realization is very costly. This work presents a methodology for the parallelization of the MC method, in the context of cloud computing. This strategy is based on the MapReduce paradigm, and allows an efficient distribution of tasks in the cloud. This methodology is illustrated on a problem of structural dynamics that is subject to uncertainties. The results show that the technique is capable of producing good results concerning statistical moments of low order. It is shown that even a simple problem may require many realizations for convergence of histograms, which makes the cloud computing strategy very attractive (due to its high scalability capacity and low-cost). Additionally, the results regarding the time of processing and storage space usage allow one to qualify this new methodology as a solution for simulations that require a number of MC realizations beyond the standard.

1. Introduction

Most of the predictions that are necessary for decision making in engineering, economics, actuarial sciences, and so on, are made based on computer models. These models are based on assumptions that may or may not be in accordance with reality. Thus, a model can have uncertainties on its predictions, due to possible wrong assumptions made during its conception. This source of variability on the response of a model is called *model uncertainty* [1]. In addition to modeling errors, the response of a model is also subject to variabilities due to uncertainties on model parameters, which may be due to measurement errors, imperfections in the manufacturing process, and other factors. This second source of randomness on the models response is called *data uncertainty* [1].

One way to take into account these uncertainties is to use the theory of probability, to describe the uncertain parameters as random variables, random processes, and/or random fields. This approach allows one to obtain a model where it is possible to quantify the variability of the response. For instance, the reader can see from [2,3] where techniques of stochastic modeling are applied to describe the dynamics of a drillstring. Other applications in structural dynamics can be seen in [4,5]. It is also worth mentioning the contributions of [6], in the context of hydraulic fracturing, [7], for estimation of financial reserves, and in [8] for the analysis of structures built by heterogeneous hyperelastic materials. For a deeper insight into stochastic modeling, with an emphasis in structural dynamics, the reader is encouraged to read [1,9,10].

To compute the propagation of uncertainties of the random parameters through the model, the most used technique in the literature is the Monte Carlo (MC) method [11]. This technique generates several realizations (samples) of the random parameters according to their distributions (stochastic model). Each of these realizations defines a deterministic problem, which is solved (processing) using a deterministic technique, generating an amount of data. Then, all of these data are combined through statistics to access the response of the random system under analysis [12–14]. A general overview of the MC algorithm can be seen in Fig. 1.

The MC method does not require that one implements a new computer code to simulate a stochastic model. If a deterministic code to simulate a similar deterministic model is available, the stochastic simulation can be performed by running the deterministic program several times, changing only the parameters that are randomly generated. This nonintrusive characteristic is a great advantage of MC when compared with other methods for uncertainty quantification, such as generalized Polynomial Chaos (gPC), [15], which demands a new code for each new random system that one wants to simulate. Additionally, if the MC simulation is performed

* Corresponding author. Tel.: +55 21 3527 1500.
E-mail addresses: lopes@inf.puc-rio.br, hcvlopes@gmail.com (H. Lopes).

0010-4655/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cpc.2014.01.006

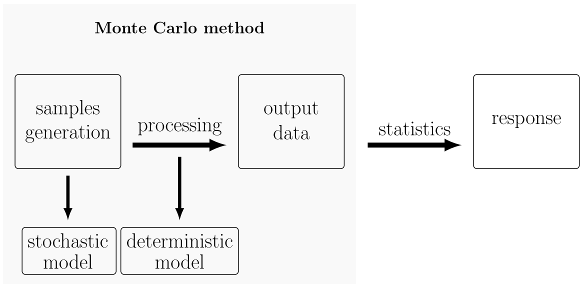

Fig. 1. General overview of the Monte Carlo algorithm.

for a large number of samples, it completely describes the statistical behavior of the random system. Unfortunately, MC is a very time-consuming method, which makes unfeasible its use for complex simulations, when the processing time of a single realization is very large or the number of realizations to an accurate result is huge [12–14].

Meanwhile, the MC method algorithm can easily be parallelized because each realization can be done separately and then aggregated to calculate the statistics. The parallelization of the MC algorithm allows one to obtain significant gains in terms of processing time, which can enable the use of the MC method in complex simulations. In this context, cloud computing can be a very fruitful tool to enable the use of the MC method to access complex stochastic models because it is a natural environment for implementation of parallelization strategies. Moreover, in theory, cloud computing offers almost infinite scalability in terms of storage space, memory and processing capability, with a financial cost significantly lower than the one that is necessary to acquire a traditional cluster with the same capacity.

In this spirit, this work presents a methodology for implementing the MC method in a cloud computing setting, which is inspired by the MapReduce paradigm [16]. This approach consists of splitting, among several instances of the cloud environment, the MC calculation, processing each one of these tasks in parallel, and finally merging the results into a single instance to compute the statistics. As an example, the methodology is applied to a simple problem of stochastic structural dynamics. The use of cloud in not new in the context of engineering and sciences [17]. We would like to mention the work of Ari and Muhtaroglu [18] that proposes a cloud computing service for finite element analysis, the work of Jorissen et al. [19] that proposes a scientific cloud computing platform that offers high performance computation capability for materials simulations, and the work of Wang et al. [20] that discusses the Cumulus cloud based project with its applications to scientific computing, just to cite a few.

This paper is organized as follows. Section 2 makes a brief presentation of the cloud computing concept. Section 3 presents a parallelization strategy for the MC method in the context of cloud computing. Section 4 describes the case of study in which the proposed methodology is exemplified. Section 5 presents and discusses the statistics done with the data and the convergence of the results. Finally, Section 6 presents the conclusions and highlights the main contribution of this work.

2. Cloud computing

Traditionally, the term cloud is a metaphor about the way the Internet is usually represented in network diagrams. In these diagrams, the icon of the cloud represents all the technologies that make the Internet work, ignoring the infrastructure and complexity that it includes. Likewise, the term cloud has been used as an abstraction for a combination of various computer technologies and paradigms, e.g., virtualization, utility computing, grid computing, service-oriented architecture and others, which together provide computational resources on demand, such as storage, database, bandwidth and processing [21].

Therefore, cloud computing can be understood as a style of computing where information technology capabilities are elastic, scalable, and are provided as services to the users via the Internet [22,23]. In this style of computing, the computational resources are provided for the users on demand, as a pay-as-you-go business model, where they only need to pay for the resources that were effectively used. Due to its great potential for solving practical problems of computing, it is recognized as one of the top five emerging technologies that will have a major impact on the quality of science and society over the next 20 years [24].

The reader can see from [25] a detailed comparison between three cloud providers (Amazon EC2, Microsoft Azure and Rackspace) and a traditional cluster of machines. These experiments were done using the well-known NAS parallel benchmarks as an example of general scientific application. That article demonstrates that the cloud can have a higher performance and cost efficiency than a traditional cluster.

Furthermore, a traditional cluster require huge investments in hardware and in their maintenance and one cannot "turn off resources contracts", while they are unnecessary, to save money. Then, traditional clusters are almost prohibitive for scientific research without large financial resources.

In a cloud computing environment one pays only per hour of use of one virtual machine. Other costs of this platform are the shared/redundant storage and data transfers. For the data transfer, all inbound data transfers (i.e., data going to the cloud) are free and the price for outbound data transfers (i.e., data going out of the cloud) is a small cost that depends on the volume of data transferred.

Given these characteristics, it is easy to imagine a situation where computational resources can be turned on and off according to demand, providing unprecedented savings compared with acquisition and maintenance of a traditional cluster. In addition, if it is possible to parallelize the execution, the total duration of the process can be minimized using more virtual machines of the cloud.

3. Parallelization of the Monte Carlo method in the cloud

The strategy to run the MC algorithm in parallel, as proposed in this work is influenced by the MapReduce paradigm [16], which was originally presented to support the processing of large collections of data in parallel and distributed environments. This paradigm consists in two phases: the first (Map) divides the computational job into several partitions and each partition is executed in parallel by different machines; the second phase (Reduce) collects the partial results returned by each machine, aggregates partial results and computes a response to the computational job.

We propose a MapReduce strategy for parallel execution in the cloud of the MC method that is composed of three steps: *split*, *process*, and *merge*. The split and the process steps correspond to the Map, while the merge corresponds to the Reduce step. This strategy of parallelization was implemented in a cloud computing setting called McCloud [26,27], which runs on the Microsoft Windows Azure platform (http://www.windowsazure.com). A general overview of the strategy can be seen in Fig. 2, and a detailed description of each step is made below.