Future Generation Computer Systems 29 (2013) 1786–1794

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

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

Cost minimization for computational applications on hybrid cloud infrastructures

Maciej Malawski [a,b,*], Kamil Figiela [b], Jarek Nabrzyski [a]

[a] University of Notre Dame, Center for Research Computing, Notre Dame, IN 46556, USA

[b] AGH University of Science and Technology, Department of Computer Science, Mickiewicza 30, 30-059 Kraków, Poland

ARTICLE INFO

Article history:
Received 1 August 2012
Received in revised form
4 January 2013
Accepted 16 January 2013
Available online 23 January 2013

Keywords:
Distributed systems
Cloud computing
Constrained optimization

ABSTRACT

We address the problem of task planning on multiple clouds formulated as a mixed integer nonlinear programming problem (MINLP). Its specification with AMPL modeling language allows us to apply solvers such as Bonmin and Cbc. Our model assumes multiple heterogeneous compute and storage cloud providers, such as Amazon, Rackspace, GoGrid, ElasticHosts and a private cloud, parameterized by costs and performance, including constraints on maximum number of resources at each cloud. The optimization objective is the total cost, under deadline constraint. We compute the relation between deadline and cost for a sample set of data- and compute-intensive tasks, representing bioinformatics experiments. Our results illustrate typical problems when making decisions on deployment planning on clouds and how they can be addressed using optimization techniques.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In contrast to already well established computing and storage resources (clusters, grids) for the research community, clouds in the form of infrastructure-as-a-service (IaaS) platforms (pioneered by Amazon EC2) provide on-demand resource provisioning with a pay-per-use model. These capabilities together with the benefits introduced by virtualization, make clouds attractive to the scientific community [1]. In addition to public clouds such as Amazon EC2 or Rackspace, private and community cloud installations have been deployed for the purpose of scientific projects, e.g. FutureGrid[1] or campus-based private cloud at Notre Dame.[2] As a result, multiple deployment scenarios differing in costs and performance, coupled together with new provisioning models offered by clouds make the problem of resource allocation and capacity planning for scientific applications a challenge.

The motivation for this research comes from our previous work [2,3], in which we run experiments with compute-intensive bioinformatics application on a hybrid cloud consisting of Amazon EC2 and a private cloud. The application is composed of a set of components (deployed as virtual machines) that communicate using a queue (Amazon SQS) and process data that is stored on a cloud storage (Amazon S3). The results of these experiments indicate that clouds do not introduce significant delays in terms of virtualization overhead and deployment times. However, multiple options for placement of application components and input/output data, which differ in their performance and costs, lead to non-trivial resource allocation decisions. For example, when data is stored on the public cloud, the data transfer costs between storage and a private cloud may become large enough to make it more economical to pay for compute resources from the public cloud than to transfer the data to a private cloud where computing is cheaper.

In this paper, we address the resource allocation problem by applying the optimization techniques using AMPL modeling language [4], which provides access to a wide range of ready to use solvers. Our model assumes multiple heterogeneous compute and storage cloud providers, such as Amazon, Rackspace, ElasticHosts and a private cloud, parameterized by costs and performance. We also assume that the number of resources of a given type in each cloud may be limited, which is often the case not only for private clouds, but also for larger commercial ones. The optimization objective is the total cost, under deadline constraint. To illustrate how these optimization tools can be useful for planning decisions, we analyze the relations between deadline and cost for different task and data sizes, which are close to our experiments with bioinformatics applications.

The main contributions of the paper are the following:

- We formulate the problem of minimization of cost of running computational application on hybrid cloud infrastructure as a mixed integer nonlinear programming problem and its specification with AMPL modeling language.

* Corresponding author at: AGH University of Science and Technology, Department of Computer Science, Mickiewicza 30, 30-059 Kraków, Poland. Tel.: +48 12 3283353.

E-mail addresses: malawski@agh.edu.pl (M. Malawski), naber@nd.edu (J. Nabrzyski).

[1] http://futuregrid.org.

[2] http://www.cse.nd.edu/~ccl/operations/opennebula/.

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2013.01.004

– We evaluate the model on scenarios involving limited and unlimited public and private cloud resources, for compute-intensive and data-intensive tasks, and for a wide range of deadline parameters.
– We discuss the results and lessons learned from the model and its evaluation.

The paper is organized as follows: after discussing the related work in Section 2, we introduce the details and assumptions of our application and infrastructure model in Section 3. Then, in Section 4 we formulate the problem using AMPL by specifying the variables, parameters, constraints and optimization goals. Section 5 presents the results we obtained by applying the model to the scenarios involving multiple public and private clouds, overlapping computation and data transfers, and identifying special cases. In Section 6 we provide a sensitivity analysis of our model and show how such analysis can be useful for potential users or computing service resellers. In Section 7 we estimate how our model behaves if the task sizes are not uniform and change dynamically. The conclusions and future work are given in Section 8.

2. Related work

The problem of resource provisioning in IaaS clouds has been recently addressed in [5,6]. They typically consider unpredictable dynamic workloads and optimize the objectives such as cost, runtime or utility function by autoscaling the resource pool at runtime. These approaches, however, do not address the problem of data transfer time and cost, which we consider an important factor.

Integer programming approach has been applied to the optimization of service selection for activities of QoS aware grid workflows [7]. On the other hand, in our model we assume the IaaS cloud infrastructure, while the objective function takes into account costs and delays of data transfers associated with the tasks.

The cost minimization problem on clouds addressed in [8] uses a different model from ours. We impose a deadline constraint and assume that the number of instances available from providers may be limited. To satisfy these constraints, the planner has to choose resources from multiple providers. Our model also assumes that VM instances are billed per hour of usage.

3. Model

3.1. Application model

The goal of this research is to minimize the cost of processing a given number of tasks on a hybrid cloud platform, as illustrated in Fig. 1. We assume that tasks are independent from each other, but they have identical computational cost and require a constant amount of data transfer.

The assumption of homogeneous tasks can be justified by the reason that there are many examples of scientific applications (e.g. scientific workflows or large parameter sweeps) that include a stage of a high number of parallel nearly identical tasks. Such examples can be found e.g. in typical scientific workflows executed using Pegasus Workflow Management system, where e.g. CyberShake or LIGO workflows have a parallel stage of nearly homogeneous tasks [9]. Other examples are Wien2K and ASTRO workflows that consist of iteratively executed parallel stages comprising homogeneous tasks [10]. Due to the high number of parallel branches, these stages accumulate the most significant computing time of the whole application, so optimization of the execution of this stage is crucial. Moreover, if the tasks are not ideally homogeneous, it is possible to approximate them using a uniform set of tasks with the mean computational cost and data

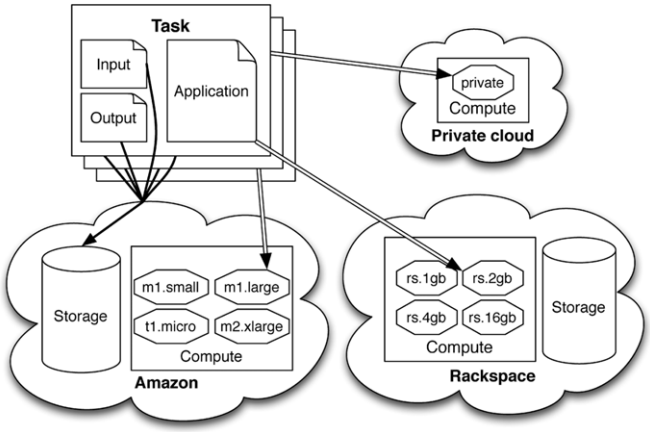

Fig. 1. The model of application and infrastructure.

sizes of the application tasks. Of course, in real execution the actual task performance may vary, so the solution obtained using our optimization method becomes only approximate of the best allocation, and the actual cost may be higher and deadline may be exceeded. In order to estimate the quality of this approximation, we evaluate the impact of dynamic task runtime and non-uniform tasks in Section 7.

We assume that for each task a certain amount of input data needs to be downloaded, and after it finishes, the output results need to be stored. In the case of data-intensive tasks, the transfers may contribute a significant amount of total task runtime.

3.2. Infrastructure model

Two types of cloud services are required to complete tasks: storage and virtual machines. Amazon S3 and Rackspace Cloud Files are considered as examples of storage providers, while Amazon EC2, Rackspace, GoGrid and ElasticHosts represent computational services. In addition, the model includes a private cloud running on own hardware. Each cloud provider offers multiple types of virtual machine instances with different performance and price.

For each provider the number of running virtual machines may be limited. This is mainly the case for private clouds that have a limited capacity, but also the public clouds often impose limits on the number of virtual machines. E.g. Amazon EC2 allows maximum of 20 instances and requires to request a special permission to increase that limit.

Cloud providers charge their users for each running virtual machine on an hourly basis. Additionally, users are charged for remote data transfer while local transfer inside provider's cloud is usually free. These two aspects of pricing policies may have a significant impact on the cost of completing a computational task.

Cloud services are characterized by their pricing and performance. Instance types are described by price per hour, relative performance and data transfer cost. To assess the relative performance of clouds it is possible to run application-specific benchmarks on all of them, or to use publicly available cloud benchmarking services, such as CloudHarmony.[3] CloudHarmony defines performance of cloud instances in the units named CloudHarmony Compute Units (CCU) as similar to Amazon EC2 Compute Unit (ECU), which are approximately equivalent to CPU capacity of a 1.0–1.2 GHz 2007 Opteron or 2007 Xeon processor. Storage platforms include fees for

[3] http://blog.cloudharmony.com/2010/05/what-is-ecu-cpu-benchmarking-in-cloud.html.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات