Future Generation Computer Systems 29 (2013) 704–712

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

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

An economic and energy-aware analysis of the viability of outsourcing cluster computing to a cloud

Carlos de Alfonso *, Miguel Caballer, Fernando Alvarruiz, Germán Moltó

Instituto de Instrumentación para Imagen Molecular (I3M), Centro mixto CSIC – Universitat Politècnica de València – CIEMAT, camino de Vera s/n, 46022 Valencia, Spain

ARTICLE INFO

Article history:
Received 27 March 2012
Received in revised form
17 August 2012
Accepted 27 August 2012
Available online 8 September 2012

Keywords:
Cloud computing
Cluster computing
Green computing
Cost analysis

ABSTRACT

This paper compares the total cost of ownership of a physical cluster with the cost of a virtual cloud-based cluster. For that purpose, cost models for both a physical cluster and a cluster on a cloud have been developed. The model for the physical cluster takes into account previous works and incorporates a more detailed study of the costs related to energy consumption and the usage of energy-saving strategies. The model for the cluster on a cloud considers pricing options offered by Amazon EC2, such as reserving instances on a long-term basis, and also considers using tools for powering nodes on and off on demand, in order to avoid the costs associated to keeping idle nodes running. Using these cost models, a comparison is made of physical clusters with cloud clusters of a similar size and performance. The results show that cloud clusters are an interesting option for start-ups and other organizations with a high degree of uncertainty with respect to the computational requirements, while physical clusters are still more economically viable for organizations with a high usage rate.

1. Introduction

One of the main problems faced when deploying a cluster of PCs relates to the high total cost of ownership (TCO). This cost involves not only the purchase and installation of the equipment (computational nodes, network components, cables, hard disks, etc.), but also the operating costs. The latter includes the salaries of the personnel in charge of the installation and maintenance, the electricity consumed, and the costs related to rent appropriate housing and its associated cooling systems. The problem is that the usage patterns of these machines are highly dynamic, where peak loads are often restricted to the context of specific experiments or deadlines. In addition to this, the prices of clusters of PCs is rapidly decreasing (due to technology obsolescence), thus reducing the value of the initial investment in hardware.

As an alternative, researchers might access the resources at the computing centers of national or international institutions. This is the case of the *Spanish Supercomputing Network*, which aggregates several supercomputing centers in Spain. The access to this equipment is supervised by an access committee that grants limited resource access according to the scientific merit of the proposals. Another example is the not-for-profit organization *Partnership for Advanced Computing in Europe* (PRACE), which provides access to a world-class computing and data management infrastructure [1].

Another alternative is to use *cloud computing*, a model for enabling convenient on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [2]. This is the case of Infrastructure as a Service (IaaS), which performs on-demand resource provision of computational resources, storage capacity, network access, etc. This is achieved by means of virtual machines that run on the cloud provider's computing center. In the case of public cloud providers, a pay-per-use pricing model is typically employed, where users are only charged for the resources that they have consumed.

Virtualization was not considered a feasible approach for high performance computing (HPC), due to the penalties involved mainly in the I/O. However, many applications running on PC clusters are CPU-bound, and thus their performance is hardly affected by virtualization [3]. In addition, certain public cloud providers, such as Amazon, offer low-latency links among instances, thus leveraging the idea of using virtual clusters on the cloud. Having a virtual cluster infrastructure on a public cloud provider has a large number of advantages for the end user, since no hardware costs are involved. However, the sustainability of this infrastructure in the long term might represent a high cost, since the pay-as-you-go model offered by the cloud providers implies that a running virtual machine costs money regardless of it being used for computations or not.

Due to the increase of the use of virtualization and cloud technologies, some initiatives to create HPC clusters over cloud infrastructures are emerging. One of the first approaches, described

* Corresponding author. Tel.: +34 963877007x88254.
E-mail address: caralla@upv.es (C. de Alfonso).

0167-739X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2012.08.014

in [4], enabled the launch of a fully functional Hadoop cluster over Amazon EC2 using a set of simple scripts.[1] Other tools can create HPC clusters in the cloud using some kind of local resource management system (LRMS) to manage the jobs. StarCluster[2] uses this approach to create a cluster in the Amazon EC2 infrastructure, with a set of predefined installed applications (Sun Grid Engine, OpenMPI, NFS, etc.) to enable launching parallel jobs to the queue system.

Recently, Cycle Computing used cloud infrastructures to create a 30,000-core HPC cluster using Amazon EC2 standard instances.[3] The cluster ran for about seven hours, with 3,809 compute instances and a total of 26.7 TB of RAM and 2 PB (petabytes) of disk space, with a 10 Gigabit Ethernet network. Amazon itself actually built a supercomputer on its own cloud that made it onto the list of the world's top 500 supercomputers. With 7,000 cores, that specific Amazon cluster ranked number 232 in the world in November 2010 with speeds of 41.82 teraflops, falling to number 451 in June of 2012.[4] It is estimated that the whole Amazon EC2 infrastructure can be ranked number 42 among the world's top 500 supercomputers.[5]

These tools and services turn the cloud into a technologically feasible option for the deployment of clusters of PCs. However, it is important to assess the economic viability of outsourcing the deployment of a cluster on the cloud, compared to the purchase of a physical cluster. For that, this article analyzes and compares the cost of having a physical HPC cluster with that of a similar infrastructure on a public cloud provider.

The remainder of the paper is structured as follows. First, Section 2 describes the related work in the literature comparing physical clusters with virtual clusters deployed on the cloud, either economically or in terms of performance. Then, Section 3 dissects the TCO of an HPC cluster, introducing an energy-aware cost model for physical clusters, and a cost model for virtual clusters on the cloud. Section 4 then introduces some simplifying assumptions and considerations in the cost models developed in the previous section, also presenting data related to energy consumption, prices of hardware components, and cloud instance prices. The resulting models are then used in Section 4.2 in order to compare physical clusters with cloud clusters. A discussion of the results is presented in Section 5, followed by concluding remarks in Section 6.

2. Related work

There is recent work in the literature comparing large-scale public cloud infrastructures with PC clusters, especially for HPC. In [5], the authors include a comprehensive evaluation of performance comparing physical HPC clusters to virtual HPC clusters on Amazon EC2, where the larger network latency in the latter introduces a severe performance penalty for parallel applications. In [6], a similar performance comparison is made with workflow applications composed of loosely coupled parallel applications consisting of computational tasks linked via data and control dependencies. Different EC2 instance types were employed to assess the performance of the applications, and a virtualization overhead below 8% was computed.

Other works have studied the cost or benefit of using cloud technologies from different points of view. In [7], the authors study the cost of executing the Montage astronomy application in public cloud environments. In [8], the authors evaluate the cost of expanding a local virtual cluster using a cloud technology provider, in order to reduce the response time of the user requests. In [9], the authors compare the performance and monetary cost benefits of clouds versus desktop grids (or volunteer computing) infrastructures, ranging in size and storage. In [10], the TCO and utilization cost of a cloud infrastructure are analyzed from the point of view of the IaaS service provider. The authors also developed a web tool with which users can introduce the parameters of their clouds and obtain the total cost analysis. Finally, other works such as [11,12] or [13] have tried to compare the cost of owning a datacenter infrastructure versus the pay-per-use costs of cloud deployments.

Those contributions show that it is crucial to evaluate the economic impact of outsourcing an organization's HPC computing infrastructure to an external cloud provider. As opposed to previous works, this paper performs a detailed comparison between physical and virtual HPC clusters from the point of view of the TCO, considering energetic, management, and infrastructural issues, using concepts and estimations from related work, but considering a more detailed analysis. In previous works, the cost of the energy is only estimated, while in this work a detailed model of the energy consumption cost has been defined, where green-aware technologies are a key task to minimize the energetic consumption and the costs in the cloud deployments. The electrical power required to run a cluster and the price of the energy have made owners to take such cost into account when operating the computing infrastructures and to create heuristics to try to reduce its impact [14,15]. This paper also analyzes how the Amazon EC2 reserved instances can be used to reduce the cost when the users can estimate the average usage of the cluster.

3. The total cost of ownership (TCO) of an HPC cluster

The TCO is generally used as a means of addressing the real costs attributed to owning and managing an IT infrastructure in a business. Therefore, the TCO of owning an HPC cluster not only includes the capital cost, but also the cost of operating the IT infrastructure, and other factors [11].

The cost of owning an HPC cluster can be modeled according to the expression (1), where C_F stands for the fixed costs, which only occur once, as opposed to variable costs (C_V), required during the operation of the equipment. C_F can be detailed as in (2).

$$C = C_F + C_V \tag{1}$$

$$C_F = C_P + C_S + C_{CP} + C_A. \tag{2}$$

Concerning the costs related to the purchase and configuration of the equipment (C_P), we have considered the computing nodes, additional components, such as switches, power device units, etc., and auxiliary physical elements (racks, cables, etc.). Besides the costs of purchasing the cluster itself, it is important to consider the costs related to buying or renting the physical space where the cluster will be located, together with appropriate refurbishment (C_S). The expression includes the costs related to the purchase of the cooling system (C_{CP}) and the administrative costs (C_A) involved in the purchase (mortgages, loans, infrastructure documentation, etc.). These costs have already been studied in the literature (see, e.g., [16]). Another concept traditionally considered when calculating the TCO is equipment disposal. This topic has not been included in the equation since, due to initiatives such as the European Recycling Platform,[6] most computer vendors, such Dell[7]

[1] http://wiki.apache.org/hadoop/AmazonEC2.

[2] http://web.mit.edu/stardev/cluster/.

[3] http://arstechnica.com/business/news/2011/09/30000-core-cluster-built-on-amazon-ec2-cloud.

[4] http://www.top500.org/system/details/10661.

[5] http://www.readwriteweb.com/cloud/2011/11/amazon-ec2-now-42-supercompute.php.

[6] http://www.erp-recycling.org/.

[7] http://www1.euro.dell.com/content/topics/topic.aspx/emea/topics/services/recycle?c=es&cs=esbsdt1&l=es&s=bsd.