Resource Elasticity

Subjects: Computer Science, Interdisciplinary Applications
Contributor: Gabriel Souto Fischer

Elasticity as the ability of a system to be able to add or remove computational resources during the use and processing of data, in order to adapt to workload variations in real time . These resources can be CPU cores, RAM memory, GPU, Virtual Machine instances (known as VM's), among others. Elasticity is a dynamic property of Cloud Computing. There are two types of elasticity in the literature: horizontal and vertical. Vertical elasticity consists of the system's ability to increase or decrease the characteristics of computational resources, such as CPU time, cores, memory and network. Horizontal elasticity consists of the system's ability to add or remove instances of computational resources associated with the application that uses them. This entry will describe the main classifications for elastic systems.

Keywords: Elasticity ; Cloud

1. Introduction

Al-Dhuraibi et al.[1] define the concept of elasticity as the ability of a system to be able to add or remove computational resources during the use and processing of data, in order to adapt to workload variations in real time. These resources can be CPU cores, RAM memory, GPU, Virtual Machine instances (known as VM's), among others. According to research by \citetexto{Rostirolla:2017:8027084}, the concept of elasticity can be extended to the management of energy consumption in smart cities, so that a city's resources can be turned on and off automatically, as needed of users. Based on this, it can be inferred that the concept of elasticity, despite being proposed for the dynamic allocation of resources in cloud computing, can be extended to other application areas. For Al-Dhuraibi et al.[1] and Rostirolla et al.[2] scalability is the ability of a system to support growing workloads, making use of a wide range of additional resources.

Al-Dhuraibi et al.[1] say that elasticity is built on the concept of scalability, and can be considered an automation of this process, however, it aims to optimize the resources as quickly as possible at a given moment in any system. Another term closely associated with elasticity is efficiency, which attempts to adequately define how a resource can be efficiently used in the process of scaling a system. Therefore, the smaller the amount of resources used to perform a given task in an acceptable time, the better the efficiency of the scalable system. Elastic systems can be classified in several ways, depending on their attributes, which can be:

- Configuration
- Scope
- Objective
- Method
- Architecture
- Provider
- Operation Mode

2. Configuration

According to Al-Dhuraibi et al.[1], the configuration of an elastic system refers to the specific allocation of CPU, memory, network and storage. In this context, the configuration represents the initial reservation of resources for the system. During the first purchase of the elastic service, the user generally chooses from a list of resource packages, depending on their needs.

According to Al-Dhuraibi et al.[1], the configuration can have two modes of operation:

- Rigid or
- Configurable.

In rigid mode, resources are offered and provisioned at constant capacities. In this mode, virtual machine instances have predefined resource limits (CPU, memory, among others). The problem with rigid configuration is that allocated resources rarely meet demand, so there is always a lower or excessive provisioning than is actually necessary. In configurable mode, the customer is able to choose the specific resources of each of the virtual machines.

3. Scope

For Al-Dhuraibi et al.[1], elasticity actions can be applied at two levels, **infrastructure** or **application/platform**. Elasticity actions are responsible for executing the decisions made by the elasticity control system, in order to resize resources according to the need and the established strategy. Elasticity actions can be performed at the infrastructure level, where the elasticity controller monitors the system and makes decisions. Cloud infrastructures are based on virtualization technology, which can be virtual machines or containers. Furthermore, regarding the application/platform, they may be of a layer or multiple layers.

4. Purpose

Resource elasticity can have different purposes, such as improving performance, increasing resource capacity, saving energy, reducing costs, and ensuring availability. Regarding elasticity objectives, there are different perspectives: Cloud IaaS (Infrastructure as a Service) providers try to maximize profit by minimizing resources while offering a good Quality of Service (QoS), PaaS providers (Platform as a Service) seek to minimize the cost they pay to the cloud and customers seek to increase their Quality of Experience (QoE) and minimize your payments. Elasticity solutions cannot fulfill elasticity purposes from different perspectives at the same time, each solution typically deals with one perspective. However, some solutions try to find an optimal way to balance some of the contradictory objectives[1].

5. Method

According to Al-Dhuraibi et al.[1], to deploy elasticity solutions, one of the following methods must be implemented: horizontal scalability, vertical scalability or hybrid scalability. Horizontal and vertical scaling techniques have their advantages and disadvantages. Horizontal elasticity is easy to implement and is widely used in commercial environments. However, it can lead to inefficient use of resources, due to the fact that it provides fixed or static instances, which in most cases are not able to adjust exactly to the demand required by the application. Vertical elasticity allows instances to be resized, however it is not widely used, with few commercial systems that support it. In horizontal scalability, instances are added and removed according to load balancer techniques and in vertical scalability, resources such as memory, CPU, are resized at run time.

Migration can also be considered as a necessary action to further enable vertical scaling when there are not enough resources available, or when it is necessary to transfer an instance to a physical machine with fewer resources in order to improve performance. Before executing the VM migration or replication process, a Resource Allocation Strategy (RAS) is used to decide where the new instance will be allocated or created, on which cloud server. There are also systems that use a structure that combines vertical scaling, adding and removing resources from existing VMs, and horizontal scaling, adding new VMs as needed. RAS can be based on the cost and speed of use of each virtual machine, the CPU usage and cost of the physical machine, user-specified load conditions, among others[1].

6. Architecture

For Al-Dhuraibi et al.[1], the architecture of solutions for resource elasticity management can be centralized or decentralized. In the centralized architecture there is only one elasticity controller, responsible for performing automatic resizing, provisioning and deprovisioning resources. In the decentralized architecture, there are several elasticity controllers, responsible for provisioning resources on different cloud platforms. In this model there is also an arbitrator, considered a key point in the decentralized architecture, responsible for allocating resources to the controllers in the different components of the system.

7. Provider

Resource elasticity solutions can be applied to one or multiple cloud providers. A single cloud provider can be public or private, physically present in one or several regions or data centers. In this context, multi-cloud means that there is more than one cloud provider. Cloud providers can include hybrid clouds that can be private or public. Most solutions and proposals for resource elasticity only support a single cloud provider[1].

8. Operation Mode

Mode of operation refers to the interactions necessary to perform elastic actions in the system. Normally, elasticity actions are performed automatically. Scalability actions can be achieved by manual user intervention, through a manual or programmable mode, where elasticity actions are generally performed through an Application Programming Interface (API). Manual policy is used in some commercial cloud systems where the user is responsible for monitoring the virtual environment and performing all resizing actions. This mode, despite being linked to the concept of scalability, cannot be considered as an elasticity mode, as it violates the concept of automation, necessary for the system to be considered elastic. Therefore, an elastic system has only one mode, automatic mode, where all resizing actions are carried out automatically, and can be classified into two sub-modes:

- Reactive and
- Proactive or Predictive

In reactive elasticity, elasticity actions are triggered according to predefined rules or thresholds, causing the system to react by triggering actions to adapt changes to the system according to the load (workload or resource usage)[1].

According to Al-Dhuraibi et al.[1], in proactive or predictive elasticity, elasticity actions are triggered based on forecasting techniques, anticipating the future needs of the application and triggering elasticity actions based on this predicted anticipation.

References

1. AL-DHURAIBI, Y.; PARAISO, F.; DJARALLAH, N.; MERLE, P. Elasticity in Cloud Computing: state of the art and research challenges. IEEE Transactions on Servicesm Computing, USA, v. PP, n. 99, p. 1–1, 2017.
2. ROSTIROLLA, G.; ROSA RIGHI, R. da; BARBOSA, J. L. V.; COSTA, C. A. da. ElCity: an elastic multilevel energy saving model for smart cities. IEEE Transactions on Sustainable Computing, Piscataway, NJ, USA, v. PP, n. 99, p. 1–1, 2017.