1. Introduction

As computer systems evolve, the volume of data to be processed increases significantly, either as a consequence of the expanding amount of available information, or due to the possibility of performing highly complex operations that were not feasible in the past. Nevertheless, tasks that depend on the manipulation of large amounts of information are still performed at large computational cost, i.e., either the processing time will be large, or they will require intensive use of computer resources.

In this scenario, the efficient use of available computational resources is paramount, and creates a demand for systems that can optimize the use of resources in relation to the amount of data to be processed. This problem becomes increasingly critical when the volume of information to be processed is variable, i.e., there is a seasonal variation of demand. Such demand variations are caused by a variety of factors, such as an unanticipated burst of client requests, a time-critical simulation, or high volumes of simultaneous video uploads, e.g. as a consequence of a public contest. In these cases, there are moments when the demand is very low (resources are almost idle) while, conversely, at other moments, the processing demand exceeds the resources capacity.

Moreover, from an economical perspective, seasonal demands do not justify a massive investment in infrastructure, just to provide enough computing power for peak situations. In this light, the ability to build adaptive systems, capable of using on demand resources provided by Cloud Computing infrastructures [1,2,3], is very attractive.

1.1 Context

One of the areas where the volume of information to be processed is increasing at a very fast pace is Video Production and Distribution [38]. The increasing popularity of videos on the Internet, allied to the advances in network technology seen in the last decade, is drastically changing TV as we know it. In the past decades there was a very clear distinction among those who produced, distributed, and consumed video contents. Traditionally, video footage was produced by TV channels and by independent companies, distributed to local

broadcasters, to then reach general audiences. Roles and responsibilities were clear down the line [31].

A few years ago the "family around the TV set" scenario was greatly challenged, as cable multiplied the number of choices, and hardware prices allowed middle class families to own as many TV sets as there were members in the family. If cable only stretched the model, the Internet has the potential to completely reinvent TV as we know it.

Firstly, because it allows users to see what they want, when they want – suppressing the need for additional hardware. Digital Video Recorders (notably the TiVo) popularized the concept of letting the consumer chose a more convenient time to watch a program. However, with the increase in bandwidth availability for the last mile [39], cable and ASDL [40], it makes much more sense to stream the content directly from the internet than recording it for later use.

Secondly, and much more important, because it removes the barrier that separates producers, distributers and consumers. In the Internet, anyone can produce and distribute high quality content. As a result, there is much more available, making the competition for audiences much tougher, and changing viewing habits in an irreversible fashion.

Thirdly, the Internet allows the mix and match of multi source content. It is becoming commonplace for networks to mix their own footage with User Generated Content (UGC) – to provide a more realistic experience. Recent Haiti earthquake coverage featured as much home made mobile phone videos than proprietary professional footage.

In this scenario, video production and distribution are no longer the privilege of a few. Users were given a front seat, forcing giants to explore new business models. What we see, in fact, is that every day the evolution of technology is challenging established business models, being driven by consumers that require a more complete, flexible and interactive experience.

From a technical point of view, there are also big challenges, which include the ability to process, index, store and distribute very large amounts of data.

Fifty years ago, video was produced in a single format, mostly because its consumption was restricted to some specific screens. Today, however, it is often

necessary to generate different versions (encoding and compression) of the same information piece, so one can have access to it on dozens of different devices, such as PC, mobile phones, game consoles, media centers, tablets, PDAs, eBook readers, not to mention router type devices that allow the use of regular TV sets to display Internet content, such as Apple TV, Google TV, BoxeeBox, among others. Such devices offer different hardware capabilities, which often means a different compatibility with media formats and compression profiles. This is a critical problem, as video processing is notably computationally expensive as it is data intensive, time, and resource consuming.

In addiction to the proliferation of different devices, as video technology evolves, it is not uncommon that the market elects preferred platforms for Internet Video distribution. A decade ago RealNetworks's technology RealVideo and RealMedia took the lead, followed by MS Window Media, Adobe's Flash with the future looking into HTML 5 [41]. Each of these technologies defines a set of associated formats, codecs and transport protocols. Switching video platforms will likely require transcoding contents to new formats, codecs, as well as server migration. For a medium sized content provider, with hundreds of thousands of video files, transcoding all this content, to support this technology evolution, becomes an unprecedented challenge.

Associated with this need for multiple formats, with the increasing of bandwidth availability in last mile, there is also an increasing demand for high quality content, or, specifically, high definition videos, that requires much more computational resources to be produced, since a HD (high definition) video may have six times more information than a SD (standard definition) video. Considering that video consumption in the Internet is growing, as well as the demand for higher quality content (high definition, 3D), dealing efficiently with transcoding processes is considered very strategic.

In summary, the challenge for video production and distribution companies is to process unprecedented volumes of high definition video, and distribute it across several different devices and several different media platforms, which means a significant increasing in the encoding process, and, as consequence, in the time and cost required to perform this task.

1.2 Goals

In this thesis we propose an architecture for processing large volumes of video. Our proposal takes advantage of the elasticity provided by Cloud Computing infrastructures. We define elasticity as the ability of a software to provision computational resources needed to perform a particular task on demand. The scope of this work is limited to the range of software applications that have the following characteristics:

• there is a need for processing large volumes of data, i.e., applications where the amount of information to be processed to obtain the desired result is so large that the time required to complete this task using a traditional approach is not acceptable;

• there is a seasonal demand for processing, i.e., there are moments where the amount of information to be processed exceeds the capacity of available resources, while at other times the same resources may be idle;

• extraordinary and *ad hoc* situations, where the infrastructure will need to scale up or down repeated times, e.g., transcoding all legacy content from one compression technology to a different one;

• there is a gain in reducing the maximum possible time necessary to obtain the desired result, e.g., immediacy;

1.3 Main Contributions

The main contributions of this research are as follows:

- To propose a general architecture for large scale distributed video processing system, which is capable to address these problems;
- Develop an algorithm that allows parallel and distributed task processing, which takes advantage of the elasticity provided by Cloud platforms, adapting computational resources according to current demands;
- To implement different prototypes to demonstrate the feasibility of the proposed approach;

- To identify and provide a precise formulation of problems characterized by the seasonal demand for large volumes of video processing;