Which Cloud Auto-Scaler Should I Use for my Application? Benchmarking Auto-Scale Algorithms

[Poster Paper]

Ahmed Ali-Eldin  
Dept. of Computing Science  
Umeå University  
ahmeda@cs.umu.se

Alexey Ilyushkin  
Delft University of Technology  
Delft, the Netherlands  
a.s.ilyushkin@tudelft.nl

Bogdan Ghit  
Delft University of Technology  
Delft, the Netherlands  
b.i.ghit@tudelft.nl

Nikolas Roman Herbst  
University of Würzburg  
Würzburg, Germany  
nikolas.herbst@uni-wuerzburg.de

Alessandro Papadopoulos  
Dept. of Control  
Lund University  
alessandro.papadopoulos@control.lth.se

Alexandru Iosup  
Dept. of Control  
Lund University  
A.Iosup@tudelft.nl

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored.

ICPE’16 March 12-18, 2016, Delft, Netherlands  
© 2016 Copyright held by the owner/authors.

DOI: http://dx.doi.org/10.1145/2851553.2858677

Rapid elasticity is one of the essential characteristics of cloud computing identified by NIST [17]. Elasticity allows resources to be provisioned and released to scale rapidly outwards and inwards according to demand. Tens – if not hundreds – of algorithms have been proposed in the literature to automatically achieve elastic provisioning [15, 23, 14, 21, 13, 20, 6, 12, 16, 10]. These algorithms are typically referred to as elasticity algorithms, dynamic provisioning techniques, or autoscalers.

While trying to solve the same problem, sometimes with differing assumptions, many of these algorithms are either compared to static provisioning or to a predefined QoS target, e.g., predefined response time target, with very little – or no – comparison to previously published work. This reduces the ability of an application owner or a cloud operator to choose and deploy a suitable algorithm from the literature. Many of these algorithms have been tested with one single – real or synthetic – workload in a specific use-case [13, 14, 10]. While all published algorithms are shown to work in the specific use-case they were designed for with the, typically short, workloads tested with, it is seldom the case that the real scenarios will be any thing close to the test cases for which the algorithms are shown to work. Bursts occur in workloads occasionally. Workload dynamics change over time and the load-mix of an application significantly affects how provisioning should be done [21].

This work aims to validate and better understand the literature on automated rapid elasticity algorithms under different conditions. We compare 10 auto-scalers from the state-of-the-art, namely, [23, 7, 14, 6, 5, 12, 18, 4, 13, 11, 10, 8]. We obtained the code for five of the auto-scalers from their designers and reimplemented five.

Since cloud applications are heterogeneous in nature with different resource requirements and workload characteristics [19], for this study, we choose a set of representative applications. Each of these applications stress different resources. The set of applications chosen include complex web services, scientific workflows, big data processing, simple web services, and video streaming. Due to lack of space, we describe only two applications in more details.

1. Scientific applications and workflows. Currently, workflows are widely used to drive complex computations. A workflow (WF) or a Directed Acyclic Graph (DAG) consists of a set of tasks (nodes) which have precedence constraints among them. Any task can start the execution when all of its input dependencies are satisfied. The whole workflow in our setup is considered as a job. The popularity of workflows brings the diversity of their structures, sizes, and resource requirements. For our experiments we use three well known scientific workflows, namely, Montage, LIGO, and SIPHT. Additionally, we consider two generic workflow structures: a star (bag-of-tasks) and a chain. To schedule and execute workflows we use the KOALA scheduler [9] and OpenNebula private cloud which are deployed on the DAS-4 infrastructure 1. The execution environment for workflow tasks consist of a single head VM and multiple worker VMs.

2. Complex Web-Services. Wikipedia, the free online encyclopedia, is one of the top 10 accessed websites on the Internet [2]. The Wikimedia foundation has open sourced MediaWiki, a custom-made, free and open-source wiki software platform written in PHP and JavaScript. MediaWiki runs using a LAMP stack (Linux, Apache, MySQL, and PHP) or similar installations. This setup is similar to many cloud applications [22] including, e.g., Facebook which uses a modern version of the LAMP stack, and YouTube [1]. In our experiments, we replicate the German Wikipedia on DAS. We have chosen the German Wikipedia since it is one of the most popular Wikis in terms of number of users, and in terms of number of articles [3]. The Wikipedia server architecture deployed in their datacenters is multi-tiered with each tier deployed on a different set of servers, but in our setup, we reduce the number of layers to reduce the complexity. We have chosen to run a bare minimum setup, i.e., with the MediaWiki software, a load-balancer and a MySQL database.

1www.cs.vu.nl/das4
In addition to these general metrics, we look at application specific metrics for the applications such as response time, throughput and request drop rate for the web services, queue length for the scientific workflows and the big data workload, and jitter in the video streaming workload. The aim of our work is to provide researchers and industries with a better understanding of the current state-of-the-art.

1. REFERENCES


2This project is done as a collaboration between different partners in the SPEC Cloud Research Group.