Abstract

Multi-Layered Cloud Applications Auto-Scaling Performance Analysis

Anshul Jindal; Vladimir Podolskiy; Prof. Dr. Michael Gerndt

Informatik 10 - Lehrstuhl für Rechnertechnik und Rechnerorganisation; Fakultät für Informatik
Technische Universität München

Introduction: Cloud Computing[1] is a type of Internet-based computing that provides shared computer processing resources and data to computers and other devices on demand. With the introduction of virtualization[2] software, a physical computing device can be separated into one or more "virtual" devices, each of which can be easily used and managed to perform computing tasks. Provision and orchestration of physical and virtual resource is crucial for both Quality of Service (QoS) guarantee and cost management in cloud computing environment. Auto-scaling mechanism for these resources is essential in the lifecycle management of cloud applications. Presently several cloud auto-scaling solutions exists in the market (AWS Autoscaling[3], Google Kubernetes[4,5]). However, the application developer or the deployment administrator does not know which auto-scaling solution will be best suited for his/her application. As a result, they deploy with some already known deployment solutions and later get to know when already some resources and money has been wasted that it is not the best solution for their application. Our tool will automatically estimate and analyze the different configurations of existing cloud auto-scaling solutions in respect to performance and costs metrics, and presents the user with the best suited configuration for the deployment of application along with the pros and cons of other configurations.

Method: The method to be adopted for building the tool is as follows:

1. Identify all major market-level auto-scaling solutions (Amazon AWS auto-scale, Google Kubernetes, Microsoft Azure etc.) on all the levels of virtualization.
2. Develop the model of auto-scaling solutions where each model will include:
    a. Metrics used as markers for auto-scaling decisions (e.g. CPU load, memory consumption etc.).
    b. Tuning parameters that are used by the auto-scaling solution.
    c. Auto-scaling solutions' impact on cost.
3. Identify metrics of quality for auto-scaling solutions, like:
    a. Time taken to scale in/out the application under variable load.
    b. CPU and memory consumption.
    c. Success, Failure, and Restart rate.
4. Develop tool to estimate the quality of the multilayered application.
   The input for the tool:
    a. Application (to get the metrics)
        i. CPU Intensive Application
        ii. I/O Intensive Application
        iii. Network Application
    b. Different predefined configurations or an user defined customized configuration of auto-scaling solutions.
    c. Big-Data to test the auto-scaling.
    d. A dynamic load generation law (should be bound to different scenarios, e.g. scale in\scale out).

---

[1] "What is Cloud Computing?". Amazon Web Services. Retrieved 2017-05-09
[2] "Virtualization in education" (PDF). IBM. October 2007. 2017-05-09.
[3] "Auto Scaling". Amazon Web Services. Retrieved 2017-05-09.
[4] "What is Kubernetes?". Kubernetes. Retrieved 2017-05-09
[5] "Google Made Its Secret Blueprint Public to Boost Its Cloud". Retrieved 2017-05-09.

e. Metrics of quality for auto-scaling solutions to estimate.

Output from the tool:

a. Performance and cost of auto-scaling versus configurations of auto-scaling solutions with the current applications. For each configuration of auto-scalers:
    i. Performance values for the desired quality metrics.
    ii. Costs of auto-scaling
    iii. Pros and cons for the solution
b. Identify the best suited configuration or an ideal auto-scaling solution for the deployment of application.

Conclusion: There is a significant increase in the deployment[6] of applications on the cloud. Auto-scaling is one of the potential technologies that help Cloud operators support maximal number of users with high QOS while keeping the resource consumption at a low level. However, without proper analysis of different auto-scaling solutions the one used by them may not satisfy requirements such as reliability, stability and resource efficiency which are main challenges faced by equipment vendors and service providers. To address these challenges, our tool will automatically estimate and analyze the different cloud auto-scaling solutions, and presents the user with the best suited configuration for the deployment of application along with the pros and cons of other solutions. Hence the user can have the option to go with the best solution.

6 "Cloud Computing Trends: 2016 State of the Cloud Survey", Retrieved 2017-05-09