Four-fold Auto-scaling on a Contemporary Deployment Platform using Docker Containers

Reference:

P. Hoenisch, I. Weber, S. Schulte, L. Zhu, and A. Fekete, “Four- Containers fold Autoscaling on a Contemporary Deployment Platform using Docker (accepted for publication)”, in 13th International Conference on Service Oriented Computing (ICSOC 2015), Goa, India, 2015, pp. NN-NN.

Abstract:

With the advent of Docker, it becomes popular to bundle Web applications (apps) and their libraries into lightweight linux containers and offer them to a wide public by deploying them in the cloud. Compared to previous approaches, like deploying apps in cloud-provided virtual machines (VMs), the use of containers allows faster start-up and less overhead. However, having containers inside VMs makes the decision about elastic scaling more flexible but also more complex. In this contemporary approach to service provisioning, four dimensions of scaling have to be considered: VMs and containers can be adjusted horizontally (changes in the number of instances) and vertically (changes in the computational resources available to instances). In this paper, we address this four-fold auto-scaling by formulating the scaling decision as a multi-objective optimization problem. We evaluate our approach with realistic apps, and show that using our approach we can reduce the average cost per request by about 20-28%.