Profiling-Based Task Scheduling for Factory-Worker Applications in Infrastructure-as-a-Service Clouds

Reference:

R. Zabolotnyi, P. Leitner, and S. Dustdar, "Profiling-Based Task Scheduling for Factory-Worker Applications in Infrastructure-as-a-Service Clouds", in 40th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2014), Verona, Italy, 2014, pages 119-126.

Abstract:

With the recent advances of cloud computing, effective resource usage (e.g., CPU, memory or network) becomes an important question as application developers have to continuously pay for rented resources, even if they are not used effectively. In order to maintain required performance levels, it is currently common to reserve resources for peak resource usage or possible resource usage overlaps, if more than one task is executed on a host. While this is a reasonable approach for long-running applications or web servers, for some applications with disperse resource usage over time, this strategy causes significant over-provisioning and thus resource wastage and financial loss. In this paper we present a profiling-based task scheduling approach for factory-worker applications that schedules tasks within the defined resource limitations (e.g., existing machine memory size or network quota) and distributes the tasks in the cloud environment in order to use resources effectively. The evaluation of our approach approved the efficiency of the proposed algorithm and minimal performance overhead. In case of the evaluated application, the presented scheduling process leads up to 33% resource savings with only 1% of performance loss.