Auto Elastic: Automatic Resource Elasticity for High Performance Applications in the Cloud

Elasticity is undoubtedly one of the most striking characteristics of cloud computing. Especially in the area of high performance computing (HPC), elasticity can be used to execute irregular and CPU-intensive applications. However, the on-the-fly increase/decrease in resources is more widespread in Web systems, which have their own IaaS-level load balancer. Considering the HPC area, current approaches usually focus on batch jobs or assumptions such as previous knowledge of application phases, source code rewriting or the stop-reconfigure-and-go approach for elasticity. In this context, this article presents Auto Elastic, a PaaS-level elasticity model for HPC in the cloud. Its differential approach consists of providing elasticity for high performance applications without user intervention or source code modification. The scientific contributions of Auto Elastic are twofold: (i) an Aging-based approach to resource allocation and deallocation actions to avoid unnecessary virtual machine (VM) reconfigurations (thrashing) and (ii) asynchronism in creating and terminating VMs in such a way that the application does not need to wait for completing these procedures. The prototype evaluation using Open Nebula middleware showed performance gains of up to 26 percent in the execution time of an application with the Auto Elastic manager. Moreover, we obtained low intrusiveness for Auto Elastic when reconfigurations do not occur.