Feedback Autonomic Provisioning for Guaranteeing Performance in Map Reduce Systems

Companies have fast growing amounts of data to process and store, a data explosion is happening next to us. Currently one of the most common approaches to treat these vast data quantities are based on the Map Reduce parallel programming paradigm. While its use is widespread in the industry, ensuring performance constraints, while at the same time minimizing costs, still provides considerable challenges. We propose a coarse grained control theoretical approach, based on techniques that have already proved their usefulness in the control community. We introduce the first algorithm to create dynamic models for Big Data Map Reduce systems, running a concurrent workload. Furthermore, we identify two important control use cases: relaxed performance - minimal resource and strict performance. For the first case we develop two feedback control mechanism. A classical feedback controller and an even based feedback that minimises the number of cluster reconfigurations as well. Moreover, to address strict performance requirements a feed forward predictive controller that efficiently suppresses the effects of large workload size variations is developed. All the controllers are validated online in a benchmark running in a real 60 node Map Reduce cluster, using a data intensive Business Intelligence workload. Our experiments demonstrate the success of the control strategies employed in assuring service time constraints.