ABSTRACT

HSU, CHIN-JUNG. Improving Performance of Data-Intensive Computing for the Cloud. (Under the direction of Dr. Vincent W. Freeh.)

Data-intensive computing is extremely important because the ever-increasing data can generate information that improves and even transforms our daily life. In parallel with this demand, cloud computing shifts the computing paradigm of hosting infrastructures because users are able to pay for what they actually need and to choose resource configurations that fit diverse service requirements. Unfortunately, hosting data-intensive applications on the cloud encounters performance challenges such as reliability and efficiency due to dynamic workload, changing configurations and performance interference in the shared cloud infrastructure. Furthermore, it is not trivial to choose the most cost-effective cloud configurations because of diverse application characteristics and numerous resource choices. It is also not feasible to benchmark applications against different resource choices exhaustively.

In this dissertation, we aim to make data-intensive computing more reliable and efficient and cost-effective running on the cloud. To this end, we create robust and consistent performance models for providing reliable storage services, and design data replication and placement schemes for optimizing system throughput and query latency. We also explore the cost and performance trade-off in search for the most cost-effective way to configure and deploy data-intensive applications.

First, we implement *Inside-Out*, an automatic model building tool that creates accurate performance models for distributed storage services. Inside-Out is a black-box approach. It builds high-level performance models by applying machine learning techniques to low-level system performance metrics collected from individual components of the distributed SDS system. Inside-Out uses a two-level learning method that combines two machine learning models to automatically filter irrelevant features, boost prediction accuracy, and yield consistent prediction.

Second, we present *Rainbow*, a fine-grained, workload-aware data replication and placement scheme, for efficient cloud elasticity. This work examines the trade-off between replication factors, partition granularity, and placement strategy. It shows that coarse-grain, workload-aware replication is able to improve performance over a näive uniform data placement. Dividing the dataset into small sets, fine-grain replication, improves performance because it better matches the anticipated workload and well tolerates small mis-predictions. We propose two fine-grained placement schemes to maximize load balancing and to exploit cache locality.

Last, we are developing an approach that is able to determine the most cost-effective cloud configurations fitting diverse cost and performance requirements. We also investigate the best process to handle very large dataset on the cloud.