Application for NSF PI meeting on "The Science of Cloud Computing"

Leana Golubchik
Department of Computer Science, University of Southern California
leana@usc.edu

January 17, 2011

Two topics that best fit my research interests in Cloud Computing:

8. Cloud Self-Monitoring and Autonomic Control
10. Green Clouds

Summary of Current Activities

My recent research activities related to Cloud Computing have been broadly in the following two areas: (1) workload characterization, and (2) quality-of-service aware power cost management.

Workload characterization. Characterizing workload is an essential part of provisioning and managing resources in the Cloud. Categorizing requests and their resource demands often requires significant monitoring infrastructure. Our work thus far focused on automatically differentiating and categorizing requests without requiring sophisticated monitoring techniques. Using machine learning-based techniques, our method requires only aggregate measures (e.g., total number of requests and the total CPU and network demands) and does not assume prior knowledge of request categories or their individual resource demands. Our work explores the feasibility of such an approach to show its potential for being lightweight, generic, and easily deployable.

QoS-aware power cost management. In designing and developing large-scale distributed computing infrastructure both, energy costs and quality-of-service (QoS) are critical. However, the trade-off between these metrics is not always simple to characterize. In our work, we explore the tradeoff between operating energy cost and system performance characteristics. The high level goal of our approach is to distribute the system workload across all the geographically distributed locations, so as to reduce the overall cost of computation, without significantly reducing the quality of service experienced by users. We focus on *distributed* solutions tailored to two types of workloads – delay sensitive workloads and delay tolerant workloads – to achieve power cost savings while maintaining a certain level of QoS. Our extensive evaluation, using synthetic workload derived from real world measurements and real world power prices, illustrates that our solutions (a) do result in significant power cost savings at the expense of small increase in delay and (b) are robust under power price dynamics.

Future Research Problems

Large-scale data centers are becoming a critical part of today's computing infrastructure. Their scale requires intelligent resource management techniques, without which it would not be possible to sustain their growth while achieving desired performance characteristics at low costs. In such environments data placement, management of energy costs, ability to predict level of service delivered to users are all critical issues. These concerns lead to difficult and interesting problems in resource management, capacity planning, detection and anticipation of workload changes, and so on. Consequently, below is a brief description of the research directions we plan to pursue in Cloud Computing – we would consider these directions as falling broadly under a high level goal of *flexible cloud computing*.

That is, an important goal would be to improve the flexibility of cloud computing (where meeting performance and reliability requirements while maintaining low cost are central concerns) by enabling automatic workload management (i.e., placement and reconfiguration). Briefly, a few examples of research directions/problems in this context are as follows. One example direction in this context would be to focus on having the flexibility to combine and place different types of workload across the cloud, rather than the current (typical) situation of having physically partitioned resources that accommodate different types of workload/services. In this context, important goals would include enabling high server utilization (without degrading performance), greater opportunities for improved power usage (i.e., deploying a greater number of servers within the same power budget), resilience to failures, as well as graceful degradation under failure or other sources of changes in resource availability. Another example direction would be placement of data (needed by various cloud computing workloads) in such a manner as to provide maximum flexibility in moving workload (from server to server) in order to either improve performance (e.g., if the current server is overloaded) or reduce costs (e.g., if the current server is under-loaded, and movement of workload somewhere else would allow shutting down of that server).

In focusing on problems related to flexible cloud computing, one would need to consider and solve a number of research challenges; such challenges include the following: (1) determining appropriate considerations in and approaches to balancing performance (of various workloads) and cost objectives, (2) determining appropriate (and preferably non-intrusive) techniques for characterizing current and predicting future workloads that would enable appropriate workload placement and movement techniques; and (3) determining appropriate considerations for properly balancing performance and robustness requirements (in addition to cost).

In addressing these challenges and focusing on corresponding research problems, we envision research directions in the following broad categories: (a) *workload placement/scheduling* – i.e., allocation of resources to current workloads such that the required performance and power/cost characteristics are satisfied; (b) *data placement* – i.e., a priori placement of data (needed by the various workloads) so as to maximize the cloud's flexibility to place and (potentially) move workload (at some future time), in order to achieve given performance, reliability, availability, and cost requirements; (c) *workload prediction* – i.e., techniques for predicting and detecting changes in workload (due to a variety of system dynamics) that would allow triggering of workload re-scheduling (re-configuration) in order to meet performance, reliability, and cost requirements; and (d) *workload reconfiguration* – i.e., efficient workload reconfiguration techniques needed under system dynamics and changes in workload characteristics in order to maintain (or improve) performance, resilience, and cost characteristics.