Science of Cloud Computing: University of Minnesota Perspective

Jon Weissman and Abhishek Chandra
Department of Computer Science and Engineering
University of Minnesota Twin-Cities
{jon,chandra}@cs.umn.edu

Topic Areas: 1. Cloud Architectures and Systems; 8. Cloud Self-Monitoring and Autonomic Control

A Current Cloud Research

Our current research in cloud computing covers two areas: accelerating applications using the cloud (*proxy* project), and optimizing multi-data-center clouds for distributed MapReduce (*DMapReduce* project). In the proxy project[1], we are developing a proxy network that can accelerate applications that span one or more clouds by optimizing data transfer and performing in-network computations. Proxy nodes can cache, route, and process data that flows between clouds or between the cloud and the end-user. We have demonstrated the benefits of proxies for scientific workflow applications that use data stored in one or more clouds.

In the DMapReduce project, we are developing runtime techniques that can enable a MapReduce application in Hadoop to process data-sets located in different clouds or wide-area sites. Our approach can decide whether it is best to: (i) move all data into one data-center and then perform MapReduce, (ii) create a global MapReduce cluster that spans all locations, or (iii) enable separate MapReduce operations to proceed in parallel in each data-center, with results merged at the end. We have DMapReduce running across multiple Amazon EC2 geographic regions.

[1] DC: Small: One Thousand Points of Light: Accelerating Data-Intensive Applications By Proxy, NSF IIS 0916425

B Proposed Cloud Research

Our future-looking research ideas in cloud computing are two-fold, and we propose to pursue the following projects if funded by NSF.

Nebulas: A novel distributed cloud architecture: As our first research problem, we plan to attack the bottlenecks inherent in the current centralized cloud model due to the wide-area distribution of users and data with respect to the cloud. For example, moving large amounts of data into the cloud for short-term processing, may encounter network bandwidth bottlenecks. Similarly, if a remote user is interacting with a cloud service, then again, they may suffer from network bottlenecks between themselves and the centralized cloud. We propose a new distributed cloud architecture, called *Nebula*, comprised of edge-nodes, that will enable a greater degree of locality awareness for data and user location than does the current cloud. For example, in a Nebula cloud, data can be processed close to where it is stored or generated avoiding expensive network transmission. Similarly, a Nebula node can provide a cloud service in close network proximity to an interacting user. New cloud models such as this will broaden the applicability of clouds to other application areas, particular those that rely on dynamic, distributed data, and interacting users.

Accelerating mobile computing through the cloud: In our second research thrust, we plan to explore a unique opportunity that lies at the intersection of mobile and cloud computing. Mobile devices such as smart phones, PDAs, and tablets, offer the promise of anywhere, anytime computing and communication. Indeed, users increasingly expect their mobile device to support the same activities as their desktop counterparts: seamless multitasking, uninterrupted data access, social networking, game playing, and real work, while on-the-go. To support mobility, these devices are typically constrained in terms of their compute power, energy, and network bandwidth. Supporting the full array of desired applications and behavior in terms of performance, fidelity, and reliability, yet retaining the flexibility of the mobile cannot always be met using local resources alone. To achieve this vision, additional resources must be easily accessible on-demand. The public cloud has the promise to provide such a platform where storage and computational resources can be harnessed on-demand.

In this project, we propose to enhance the mobile user experience along the lines of performance, fidelity, and reliability by leveraging the cloud. We focus on three distinct scenarios: (1) resource-intensive data-driven applications in areas such as speech and image processing, (2) pattern-driven optimization of context-based user activities, and (3) implicit cross-user data and code sharing. We advocate a systems approach where actions can be performed transparently on behalf of the user based on their preferences with respect to response time, battery life, and reliability. A novel aspect to (3) is that we can utilize cross-user profiles and social network data to optimize sharing opportunities in the cloud.