Cloud Computing Architecture and Application Programming
DISC'09 Tutorial, half day.
Sept. 22nd 2009

Roger Barga, Jose Bernabeu-Auban, Dennis Gannon and Christophe Poulain, Microsoft Corporation.

Contact: barga@microsoft.com

Background

Over the past decade, scientific and engineering research via computing has emerged as the third pillar of the scientific process, complementing theory and experiment. Several studies have highlighted the importance of computational science as a critical enabler of scientific discovery and competitiveness in the physical and biological sciences, medicine and health care, and design and manufacturing. The ability to create rich, detailed models of natural and artificial phenomena and to process large volumes of experimental data, itself created by a new generation of scientific instruments that are themselves powered by computing, makes computing a universal intellectual amplifier, advancing all of science and engineering and powering the knowledge economy. This revolution has been enabled by the availability of inexpensive, powerful processors; low cost, large capacity storage devices; and flexible software tools, each driven by a vibrant consumer and industry marketplace.

The explosive growth of research computing systems has created major management, technical and fiscal challenges for both funding agencies and research universities. Typically, faculty members acquire research computing systems, usually small to medium (32-256 nodes) clusters, via research grants and contracts and departmental funds. This distributed acquisition of research computing and inadequate plans for long-term sustainability and technology refresh, mean that universities and funding agencies that support university research, are now struggling to create and maintain compute and data centers to house these systems and to operate and maintain them reliably, in energy-efficient, environmentally friendly ways. Moreover, university budget constraints make efficiency ever more necessary. A growing challenge is satisfying the ever rising demand for research computing and data management – the enabler of scientific discovery. Fortuitously, the emergence of cloud computing– software and services hosted by networks of commercial data centers and accessible over the Internet – offers a solution to this conundrum.

Cloud Computing

The explosive growth and rapid development of cloud services are driven by technology and business economics. Consolidating computing and storage in very large data centers creates economies of scale in facility design and construction, equipment acquisition and operations and maintenance that are not possible when the elements are distributed. However, the benefits of cloud services extend far beyond economies of scale.

First, optimized and consolidated facilities reduce total energy consumption, and they can be designed to exploit environmentally friendly and renewable energy sources. Second, cloud computing enables a "pay only for use" strategy where users bear no cost unless they use the cloud services, and then pay only for the number of service units consumed. Third, groups can deploy and expand services rapidly – in minutes, rather than the weeks or months needed to procure and install local infrastructure – to meet rising demand or to address time-critical needs. Finally, the elasticity of cloud services means that time and computing are interchangeable – the user cost to use 10,000 processors for one hour is the same as using ten processors for 1,000 hours. This is a transformative equivalence; even individuals and small companies can exploit computing resources at a scale heretofore accessible only to large companies and governments.

By outsourcing computing, data management and business intelligence services to cloud software plus services[1] providers, businesses are increasing operational efficiencies and decreasing costs, allowing them to focus on their core competencies. Similar opportunities exist in academic and research computing, but these opportunities are not being exploited.

DISC'09 Tutorial Description

The goal of this tutorial is to demonstrate how clouds can augment traditional supercomputing by expanding access to data and tools to a broader community of users than are currently served by the conventional HPC centers. Supercomputers provide the capability to conduct massive simulation and analysis computations for a few users at a time. They are not designed for on-demand access by hundreds or thousands of simultaneous users. In addition, supercomputers are not configurable by their users. Thus supercomputer applications must be modified and, in some cases, refactored as hardware and systems software is upgraded. Clouds offer the ability for each user to customize the execution environment, and to archive that customization for future use independently of the infrastructure's lifecycle.

Currently, a number of publically accessible computational platforms provide instant access to cloud-hosted services such as web search, maps, photo galleries and social networks. There are now hundreds of cloud-based services we use in our everyday life and we are starting to see some of them also touch our scientific lives. For example, Google and Live maps have been used to gain insight from geo-distributed sensor data and the Sloan Digital Sky Survey and the SkyServer have provided scientific data and tools to thousands of astronomy users. We are now at an important inflection point in the capability of cloud computing to serve the research community. Not only has the total capacity of the commercial data centers exceeded that of supercomputing centers, we now have the software infrastructure in place to allow anybody to build scalable scientific services for broad classes of users, without having to deploy, maintain and upgrade dedicated and expensive compute and data servers. This tutorial will introduce the attendees to this new technology.

This tutorial will be of value to those interested in exposing data and services to a broader audience of users without incurring the costs of acquiring and maintaining scalable infrastructure. This tutorial will introduce the attendees to the key concepts and technologies used to build and deploy scientific data analysis applications on cloud platforms. The tutorial begins with general concepts of data center architecture including the use of virtualization; the role of low power, multicore and packaging; and web service architectures. We will look at the cloud storage models with a detailed look at the Azure XStore and a brief look at Google's BigTable and GFS.

We will then focus on models of application programming. We will describe both commercial and open source tools for "map reduce" computation including Hadoop and Dryad and workflow tools for orchestrating remote data services. Following this we will examine cloud application frameworks by looking at Google's App Engine and Microsoft Azure. Throughout the tutorial we will use scientific examples to illustrate the potential applications. The tutorial concludes with a view of the future for the cloud in science.

[1] Software plus services refers to cloud services hosted in data centers, but augmented with local client software.