Lecture 23: December 4

Lecturer: Prashant Shenoy *TA: Sean Barker & Demetre Lavigne*

23.1 Cloud Computing

Cloud Computing is a growing field where a data center rents resources to customers by providing them virtual machines. These services will rent you access to a server for as little as two cents an hour. A key component of cloud computing is that you only pay for what you use (bandwidth, storage, etc). Cloud Computing services are able to do this by exploiting the fact that multiple virtual machines can be run on a single physical machine, allowing them to divide each physical server up between several customers. **Economy of Scale** also contributes to cloud computing because larger data centers are cheaper to buy and run than smaller ones. There are some problems such as privacy and security since multiple customers are sharing resources. Another problem is scalability; it is often the case that an application that works for 10,000s of servers will need to be retooled when it has to scale to 100,000s of servers (or millions of servers). Conversely, when you reach a scale of having millions of servers, what programming model will allow them to be used effectively?

23.2 Cloud Services

There are three main areas of cloud computing services: **Software as a Service** (SaaS), **Platform as a Service** (PaaS), and **Infrastructure as a Service** (IaaS).

23.2.1 Platform as a Service

A PaaS is a software platform that allows you to write your own applications but the service provider takes care of scalability. For the service provider to handle scaling a user's application they provide an API which the application must conform to. The API allows the user's application to use more or less resources (scale) as needed. Typically the amount of state that a PaaS application can use is limited because more application state makes it more difficult to scale. PaaS service providers will determine how to scale an application based on some kind of *service level agreement* (SLA). An example SLA might be that every request to a user's application must be serviced within one second. Examples of PaaS include: Google App Engine (python) and Microsoft Azure. Google's App Engine doesn't use virtualization, but it uses multi-threading and increases or decreases the number of "worker threads" as demand changes. PaaS provides a simple way to deploy applications that require scalability but is limited in how much control the user has.

23.2.2 Infrastructure as a Service

IaaS allows a user to rent raw infrastructure (servers, bandwidth, storage, etc) and they can use that infrastructure however they want to. Servce providers that offer IaaS use virtualization to share each server with multiple customers. This model allows customers to simply request resources via a web interface (for

example) and they are allotted instantly and automatically. Pricing can be very competitive, in fact bidding on resources is sometimes used to adjust prices when utilization is low and the service provider needs more customers. An example of IaaS is Amazon's EC2.

23.2.3 Software as a Service

SaaS is hosted applications that are managed by some service provider. With SaaS, a user doesn't write an application; the service provider has applications and the user can pay to use them. This simplifies management for the user in that they don't have to own/run servers or maintain application code. Examples of SaaS include: Google Apps (Gmail, Docs, etc) and Salesforce.

23.3 Private Cloud Computing

All of the examples given so far have been public cloud computing services. With public cloud computing anyone can pay to use the service. Some enterprises cannot use these kind of services due to regulatory concerns (customer data being on shared machines, for example). Others are not comfortable with sharing resources with potential competitors or they might have privacy concerns (e.g. industrial secrets). A **private** cloud addresses these concerns. The enterprise itself uses cloud computing concepts within their own private data center. The functionality is essentially the same, but the costs are typically higher. OpenStack is an example open source private cloud IaaS.

23.3.1 Hybrid Cloud

Hybrid clouds combines a private cloud with a public one. This is motivated by when a private cloud has sufficient resources most of the time but sometimes becomes overcommitted. It is possible for the enterprise to buy more resources, but this may not be efficient since they would have to support a larger infrastructure than they need most of the time. The hybrid approach allows them to keep the same infrastructure and just add capacity from the public cloud when needed.