A View of Cloud Computing

Summary by Nikhil Buduma

Cloud Computing Definition

Refers to:

1) applications delivered over the Internet
2) Hardware and systems software in data centers that provide these services

The composite of SaaS (Software as a Service) + utility computing

What Cloud Computing Brings to the Table

1) The appearance of infinite computing resources available on demand
2) Elimination of up-front commitment/cost by users
3) Ability to pay for compute use on a short-term basis and release use as needed
4) Economies of scale due to very large data centers
5) Higher utilization by multiplexing workloads from different organizations
6) Simplify operation and increase utilization via resource virtualization

Classes of Utility Computing

Cloud Apps: 1) computation, 2) storage, and 3) communication

Offerings exist at various levels of abstraction

EC2 - looks like physical hardware and can control nearly the entire software stack

AppEngine - targeted exclusively at web applications and appropriate scaling/availability mechanisms

Azure - intermediate between AppEngine and EC2 that runs on the .NET stack

Economics of Elasticity

Pay-as-you-go model enable overcoming the following scenarios:

1) provisioning static servers for peak load → underutilization (both diurnal and seasonal volatilities)
2) sometimes cannot predict, let alone provision in advance
3) cost associativity for batch analytics

Obstacles

Business Continuity and Service Availability

Adoption can sometimes be hindered by wariness of availability (outages, going out of business, regulatory action)

Potentially mitigated by using multiple cloud providers

Data Lock-In

There’s still low interoperability between platforms due to proprietary storage API’s

Attractive to cloud computing providers but users are vulnerable to price increases, security issues, or providers going out of business

Standardized API would mitigate problems and also potentially open up hybrid computing approaches

Data Confidentiality and Auditability

Trust is a major issue and in the face of HIPAA and other legislation, sometimes a legal requirement

Internal security also a significant issue

Solutions: Encryption, VLANs, Firewalls

Data Transfer Bottlenecks

Data transfer costs are high and add up, which means providers have to think about placement/traffic at every level

Solutions: shipping disks and higher b/w switches

Performance Unpredictability

VMs share CPUs and main memory surprisingly well, but network and disk I/O is more problematic.

HPC need every thread of a program to be running simultaneously and there currently isn’t any good way to do that

solutions: flash storage to improve disk I/O, gang scheduling for HPC

Scalable Storage

Less clear how to apply cloud principles to persistent storage

Open problem: Access the cloud advantage of arbitrarily scaling up and scaling down demand

Bugs in Large-Scale Distributed Systems

difficult to debug large distributed applications because buggy conditions aren’t easily reproducible

solution: capture new information by leveraging the fact that we’re running on a VM and not physical hardware

Scaling Quickly

AWS charges by the hour for the number of instances occupied even if the machine is idle

Idle servers also use a lot of power that could otherwise be saved

Solution: use machine learning to allow dynamic scaling

Reputation Fate Sharing

One customer’s bad behavior can affect others using the same cloud

Legal action at a data center (e.g. FBI raid) can hurt other tenants

Solution: Offer reputation-guarding services like those for email

Software Licensing

Current licenses usually restrict the computers on which the software can run

Users pay for annually for software and then a maintenance fee → doesn’t jive well with the pay-as-you-go model

Solution: Pay-for-use licenses