Introduction to Cloud Computing

Princeton Splash, Spring 2019

by Robert Liu (rrliu [at] princeton [dot] edu)

Key Questions

- Why does cloud computing exist? How did it evolve from the Internet?
- What insights do we gain from analyzing cloud services as infrastructure?
- How does running applications in the cloud benefit software developers?
- What are the key technologies that power cloud services?

Case Studies: Simple is Hard

- Amazon: how does the world's largest e-commerce platform provide 24/7/365 service?
- Google Docs: what kinds of problems does real-time multi-user document editing bring?
- Google Search: what systems do you need to index and search the entire Web?
- Facebook: how do they deliver the same content to billions of users worldwide?

Basic Definitions

- Server: a computer or program that provides functionality for a *client*
- Datacenter: a building that contains and supports multiple servers, connected together
- Internet: that wonderful world-wide web we all use, with billions of users, servers, sites
- Computer: anything with a web connection, from a PC to a security camera
- Database: A system that stores and retrieves data for end users

Concepts to be Covered

- What happens when you load up a modern web site?
- How can you build a site that won't crash when it gets popular?
- ALPS: High **availability**, low **latency**, **partition**-tolerant, **scalable** systems
- CAP theorem: why you can only pick two of: {consistency, availability, partition-tolerance}
- Reliability: what does "multiple nines" of uptime mean for system designers?
- Read-write amplification and tail latency: speed is essential!
- Concurrency: try to avoid letting users stomp on each others' toes and data
- Parallelism: multiple computers working together can do difficult things more quickly!
- Caching: using computers' memories to give quick answers to common questions

Recommended Reading

- James Somers, The Friendship That Made Google Huge: a profile of Jeff Dean and Sanjay Ghemawat, Google Senior Fellows who built industry-leading systems.
- DeCandia et. al., Dynamo: Amazon's Highly-Available Key-Value Store: just the first two pages gives a strong sense of the problems in modern web systems design.
- Nygren, Sitaraman, and Sun, The Akamai Network: A Platform for High-Performance Internet Applications: a description of the industry-leading content-delivery network's architecture, which also gives an impression of the Internet's problems.