Introduction to Cloud Computing

Introduction

1. What is the origin of cloud computing?
2. What are the technological challenges of cloud computing?
3. What is the future of cloud computing?

Where Did Cloud Computing Start?

1. Cloud data centers host upwards of 50K servers
2. Amazon: House 50-80K servers with a power consumption of between 25 and 30 megawatts
   - 2015: 1.5-5.6 million servers across all data centers
3. Clouds distributed operating system at Georgia Tech
   - 1986-1993
   - Kishore Ramachandran was a PI
   - Yousef Khalidi was the primary student and went on to run Azure product development
4. Cloud computing started as distributed systems research in the 80s and 90s
   - Georgia Tech, University of Washington, IBM, Emory, UCB
5. This led to grid computing in the 90s
   - Engineers and scientists began to pool resources to solve problems
6. Grid computing led to NSF HPC data centers in the mid 90s
7. Resurrection of virtualization (late 90s)
   - Originally pioneered by IBM in the 60s
   - Led to companies such as VMWare
8. Shrinking margins on selling boxes in the mid 00s
   - IBM pioneered the "services computing" model
   - Computing as a utility

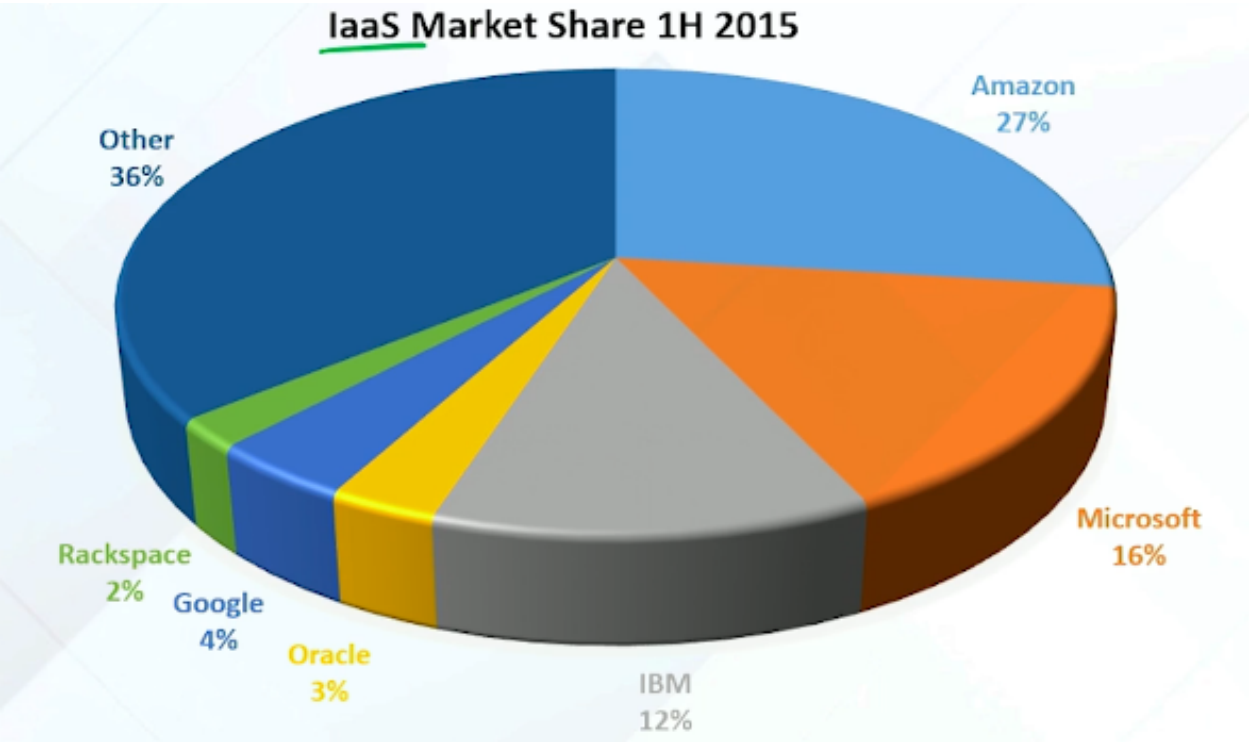
Microsoft Azure Global Presence

What is Cloud Computing?

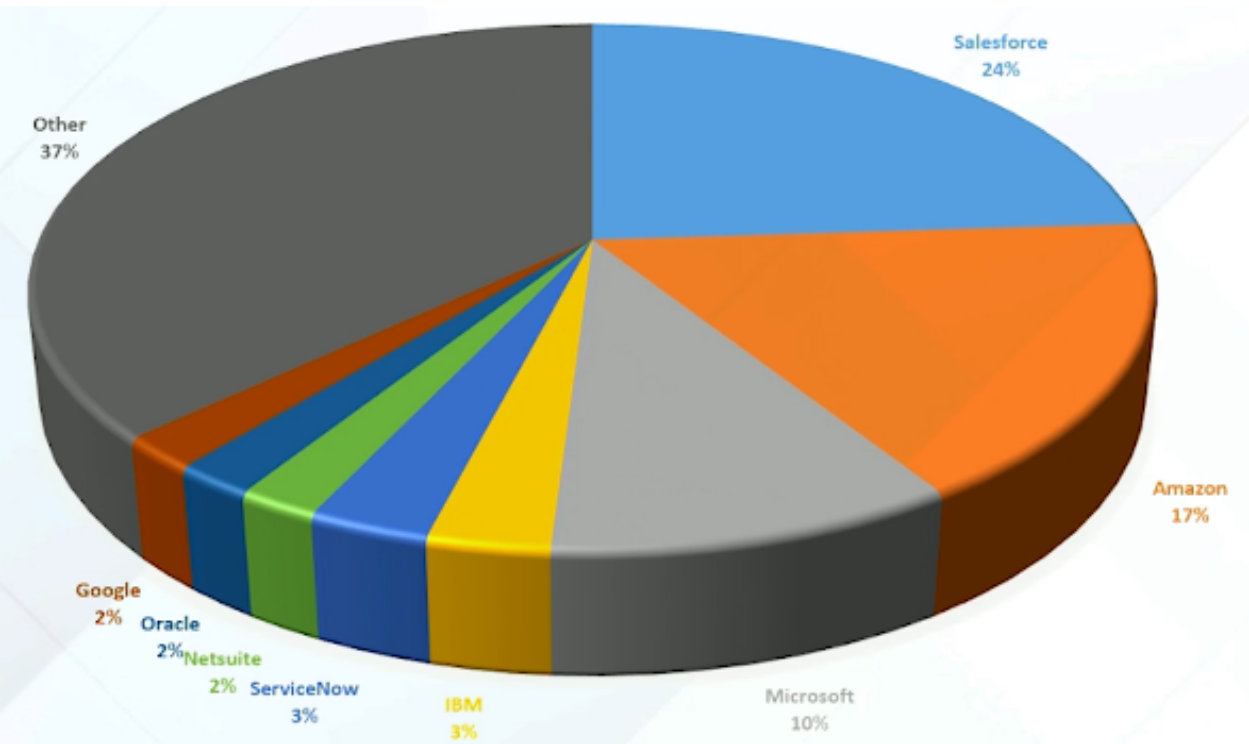
1. Amazon: Cloud computing, by definition, refers to the on-demand delivery of IT resources and applications via the Internet with pay-as-you-go pricing
2. IBM: Cloud computing is the delivery of on-demand computing resources - everything from applications to data centers - over the Internet on a pay-for-use basis
3. Computational resources (CPUs, memory, storage) in data centers available as "utilities" via the Internet
   - Illusion of infinite computational capacity
   - Ability to elastically increase/decrease resources based on need
   - "Pay as you go" model based on resource usage
   - Applications delivered as "services" over the Internet
4. Why Cloud Computing?
   - No capital or operational expenditure for owning/maintaining computational resources
   - Elasticity: Being able to shrink/expand resources based on need
   - Maintenance/upgrades are someone else's problem
   - Availability: No down time for the resources or services
   - "Pay as you go" model
   - Business services can be "out sourced"
     - Concentrate on core competency and let IBM/Amazon/Microsoft deal with the IT services
   - Disaster recovery of assets due to geographic replication of data

Types of Clouds

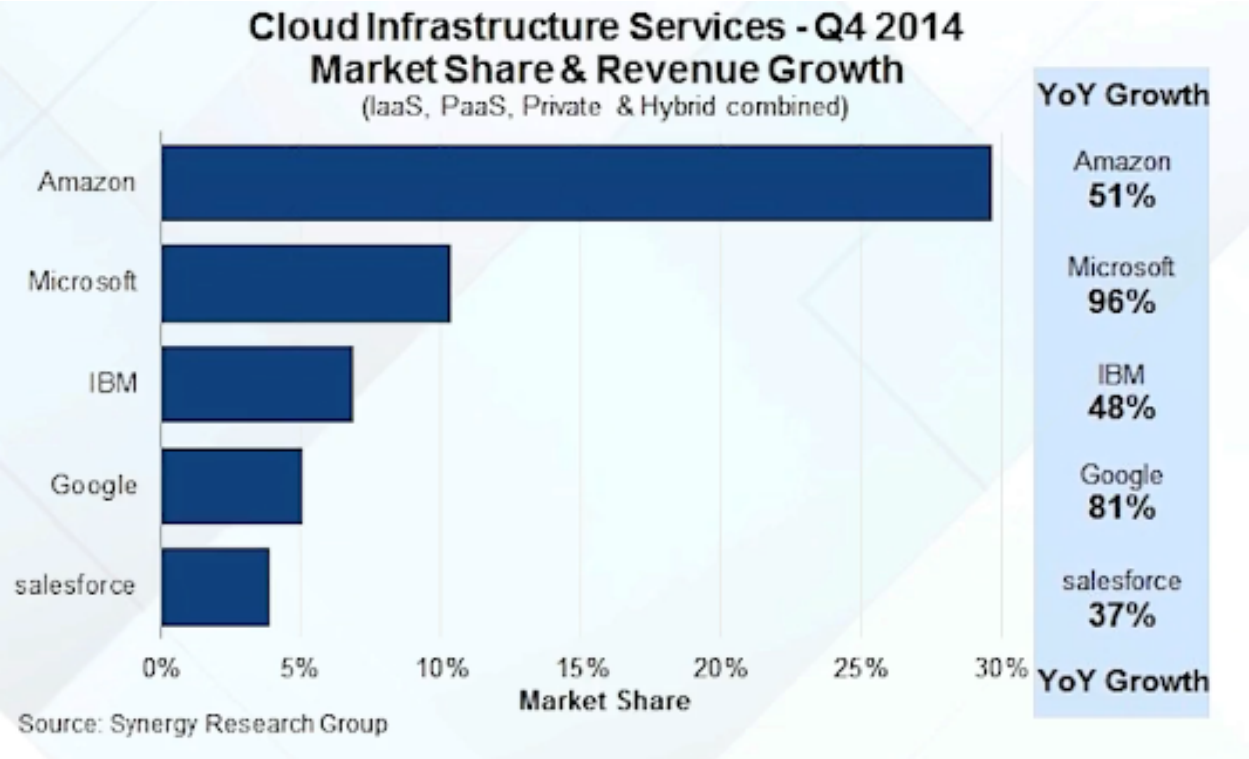
1. Different types of clouds
   - Public
     - Resources are shared among users
     - Users given the illusion that the resources are "theirs"
     - Virtualization at all levels including the network traffic guarantees perfect isolation at resource level and performance level
     - e.g., Amazon EC2
   - Private
     - Resources are physically dedicated to the individual user
     - Often the service provider may have the data center on user's premises
     - VMWare offers such services as part of their business model
   - Hybrid
     - Combines private and public
     - Keep sensitive business logic and mission-critical data in private cloud
     - Keep more mundane services (trend analysis, test and development, business projections, etc) in public cloud
     - Cloud bursting: Private cloud connects to the public cloud when demand exceeds a threshold
2. Cloud service models
   - Infrastructure as a service (IaaS)
     - Service provider offers to rent resources: CPUs, memory, network bandwidth, storage (Amazon EC2)
     - Use them as you would use your own cluster in your basement
   - Platform as a service (PaaS)
     - In addition to renting resources, service provider offers APIs for programming the resources and developing applications that run on these resources (Microsoft Azure)
     - Reduces the pain point for the cloud developer in developing, performance-tuning, and scaling large-scale cloud applications
   - Software as a service (SaaS)
     - Service provider offers services to increase end-user productivity (Gmail, Dropbox, YouTube, games)
     - User does not see physical resources in the cloud



IaaS Marketplace



PaaS Marketplace



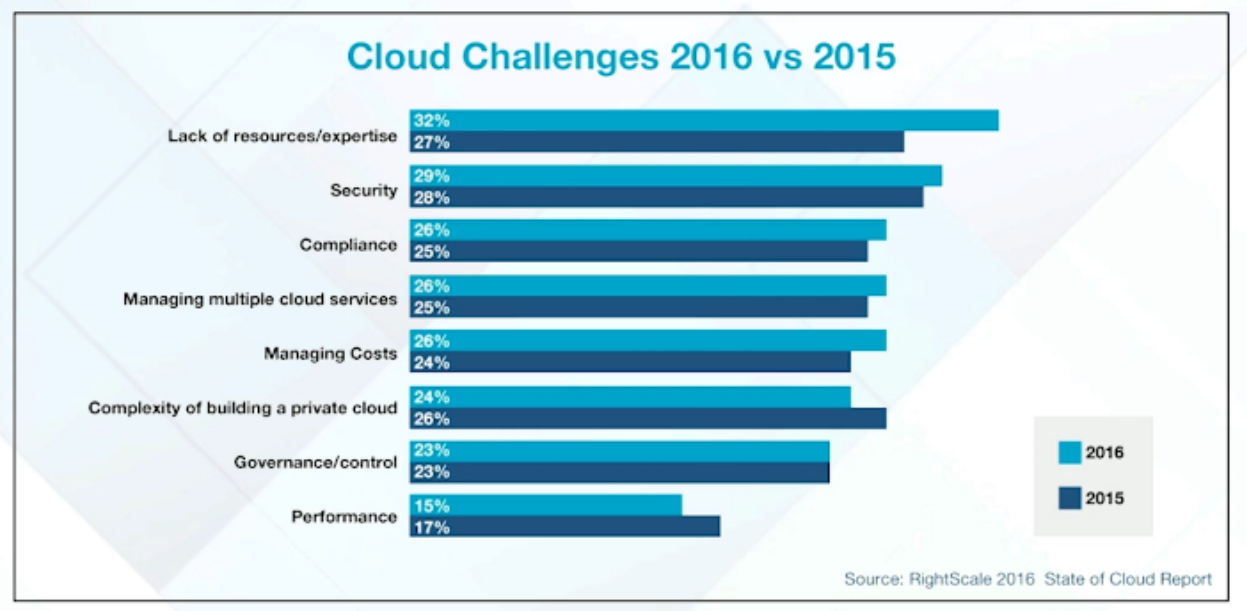
Overall Marketplace

Security Issues and Challenges

1. Challenges with cloud computing
   - Data security
   - Lock-in with a service provider
   - Network latency to the provider
   - Network bandwidth to the provider
   - Dependence on reliable Internet connectivity
2. Security Issues
   - Data breaches
   - Compromised credentials and broken authentication
   - Hacked interfaces and APIs
   - Exploited system vulnerabilities
   - Account hijacking
   - Malicious insiders
   - Parasitic computing
   - Permanent data loss
   - Inadequate diligence
   - Cloud service abuses
   - DoS attacks
   - Shared technology, shared dangers
3. Current Issues being tackled in cloud computing
   - Mobile computing
   - Architecture and virtualization
   - IoT and mobile on the cloud
   - Security and privacy
   - Distributed cloud/edge computing

- Big data
- HPC
- Network (SDN and NFV)

4. Complex issues in cloud computing
   - Optimal scheduling and resource management
   - Communication isolation, NFV, and SDN
   - HPC with loosely coupled networks
   - Real-time computations
   - Energy
     - Amazon estimates 500 MW of power from 23 datacenters in northern VA
   - Incorporating heterogeneous resources (GPUs and other accelerators)
   - Human in the loop and integration with devices
   - Improving resource utilization



Challenges

Future of Cloud Computing

1. What is next in cloud computing?
   - Energy efficient computing
     - New network hardware: Software-defined hardware (software switch), FPGA-based NICs, improved optical amplification
   - "Big data" as a service
   - Rethink security policies -> Better identity management
   - Edge computing support
2. What is current cloud computing good for?
   - Throughput oriented apps
     - Search, mail, reservations, banking, e-commerce
   - Increasingly for streaming videos using CDNs or proprietary networks such as Netflix
     - 90% of Internet traffic is video
   - Interactive apps (human in the loop)
3. Limitations of existing cloud
   - Based on large data centers

    - High latency/poor bandwidth for data-intensive apps
  - API designed for traditional web applications
    - Not suitable for the future Internet apps
    - This is because IoT and many new apps need interactive resposne at computational perception speeds (sense -> process -> actuate)
    - Sensors are geo-distributed, so latency to the cloud is a limitation and uninteresting sensor streams should be quenched at the source
4. Future applications in IoT
  - Common characteristics
    - Dealing with real-world data streams
    - Real-time interaction among mobile devices
    - Wide-area analytics
  - Requirements
    - Dynamic scalability
    - Low-latency communication
    - Efficient in-network processing
5. Future cloud
  - Encompassing geo-distributed edge computing in the context of IoT
  - Distributed programming models and runtime systems
  - Geo-distributed resource allocation
  - Static and dynamic analyses of apps expressed as dataflow graphs with temporal constraints
  - Security and privacy issues for IoT
  - System architecture of edge computing nodes
  - Front-haul networks combining fiber and Wifi
  - Deployments and field study (camera networks for campus security)
  - More issues?

Conclusion

1. Next, will learn technical details of cloud computing starting with networks