Outline

- What is a cloud?
- Applications suited for the cloud
- Data center networks
- Service models

Cloud Computing

- Shared pool of compute, storage, and network resources leased to tenants on-demand
- Key characteristics
  - Virtualized
    - Physical servers and storage devices are divided into pieces and each tenant gets an isolated piece
  - Resources are elastic and scalable
    - Tenants can request more resources when they need them
    - E.g., a company uses more virtual machines during the day when employees are using applications
  - Pay-per-use
    - Only pay for the resources you use
    - E.g., pay for a virtual machine per hour of use
  - On-demand
    - Tenants can request and release resources whenever they want
  - Resilient
    - Multiple pools of physical resources that are unlikely to fail simultaneously
    - E.g., multiple data centers around the world to avoid impacts from natural disasters
  - Shared
    - Multiple tenants share the same physical resources
    - Physical servers and storage are usually more isolated than network switches and links

Applications Suited for the Cloud

- Web sites, or web services
- Big data

Data Center Networks

- Cloud data center -- collection of physical compute, storage, and network resources
- High bandwidth and low latency are critical in cloud data center networks
- Typical physical network topology is a tree

  ![Network Topology Diagram]

  - Links higher in the topology are oversubscribed
    - Cannot handle all servers sending at the maximum rate
    - Oversubscription ratio = capacity of links below relative to capacity of links above

- Common traffic patterns
  - North-south -- traffic is exchanged between servers in the data center and hosts outside the data center; must traverse core switches
    - Common with web sites
  - East-west -- traffic is exchanged between servers or storage systems within the data center; typically only traverses ToR and aggregation switches
    - Common with back-end of web sites/web services
    - Common with MapReduce
  - Many-to-one -- many servers exchange traffic with a single server; causes a problem caused TCP incast

- Emerging data center topology -- Fat Trees

  ![Fat Trees Diagram]

  - Helps address large volumes of east-west traffic
  - Provides redundancy
  - Provides full bisection bandwidth
    - Every physical server can send at the maximum capacity of NIC (typically 10G) without causing congestion
    - The bottleneck is the network interface, not a link in the network
    - Does not fix the TCP incast problem
Cloud Service models

- **Infrastructure-as-a-Service (IaaS)**
  - Tenants lease virtual machines, virtual storage, and virtual networks
  - Tenants must manage the operating system, file system, etc.
  - E.g., Amazon EC2, Microsoft Azure, Rackspace, Google Compute Engine

- **Platform-as-a-Service (PaaS)**
  - Tenants lease resources to run applications written in a specific language -- Python, Java, Hadoop/MapReduce
  - Cloud provider manages the operating system, file system, and network
  - E.g., CloudFoundry, Oracle Cloud

- **Software-as-a-Service (SaaS)**
  - Tenants lease machines that run specific software
  - E.g., Salesforce, Concur, Constant Contact, NetSuite

- **Storage-as-a-Service (STaaS)**
  - E.g., Dropbox, Google Drive, SkyDrive

- **Ownership**
  - Public -- anyone can request and use resources
  - Private -- resources are only available to tenants (e.g., departments) within a company or organization
  - Hybrid -- tenants use a combination of public and private cloud resources

Key Challenges

- Large scale networks -- cloud data centers have tens of thousands of servers
- Shared infrastructure -- tenants are competing for bandwidth
- Security -- virtual machines, virtual storage devices, and virtual networks must be isolated so that tenants cannot access each other’s data
- Fixing problems -- many different layers where problems can occur: tenant application, tenant OS, virtual network interface, virtual switch, physical network interface, top of rack switch, aggregation switch, core switch