VE444: Networks

Yifei Zhu, assistant professor
University of Michigan-Shanghai Jiao Tong University

Acknowledgment:
This lecture's slides are modified from the slides shared by Prof. Matei Zaharia, Stanford, CS349D Cloud Computing

What is Cloud Computing

- Informal: computing with large datacenters

What is Cloud Computing

- ~~Informal: computing with large datacenters~~
- Our focus: computing as a utility
  - Outsourced to a third party or internal org

Types of Cloud Services

- Infrastructure as a Service (IaaS): VMs, disks
- Platform as a Service (PaaS): Web, MapReduce
- Software as a Service (SaaS): Email, GitHub

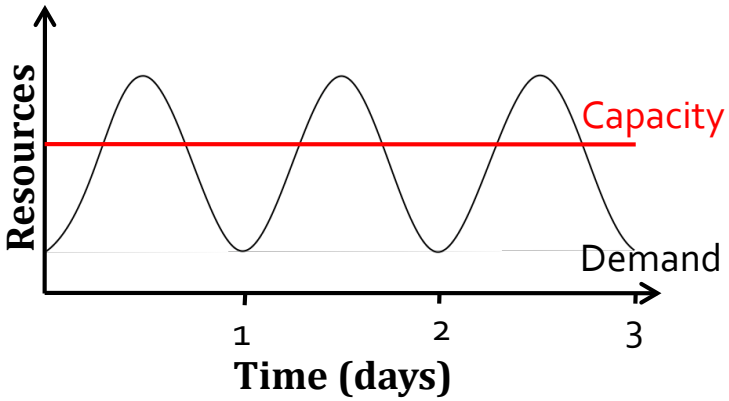
Public vs private clouds: Shared across arbitrary orgs/customers vs internal to one organization

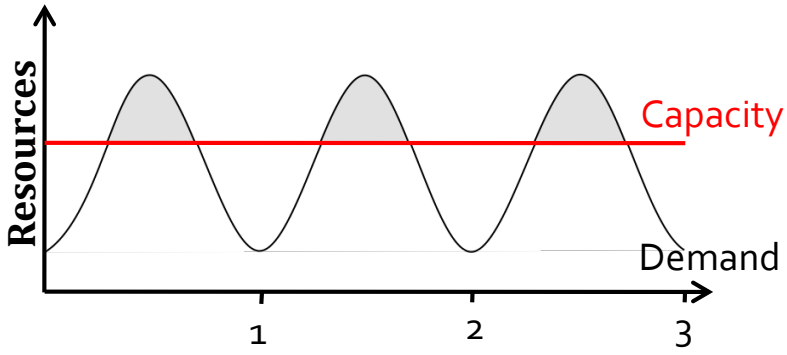
Cloud 101-definition

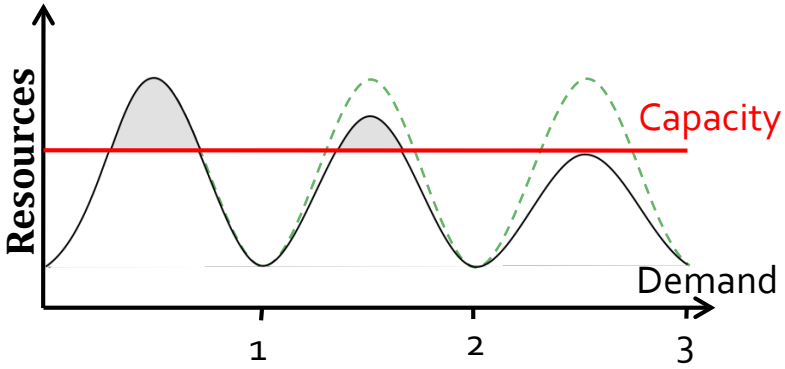
- Definition from ***NIST*** ***(National Institute of Standards and Technology)***
  - Cloud computing is a model for enabling convenient, **on-demand network access** to a **shared pool** of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be **rapidly provisioned and released** with minimal management effort or service provider interaction.

Cloud Economics: For Users

- In traditional computing model, two common problems :
- Underestimate system utilization which result in under provision

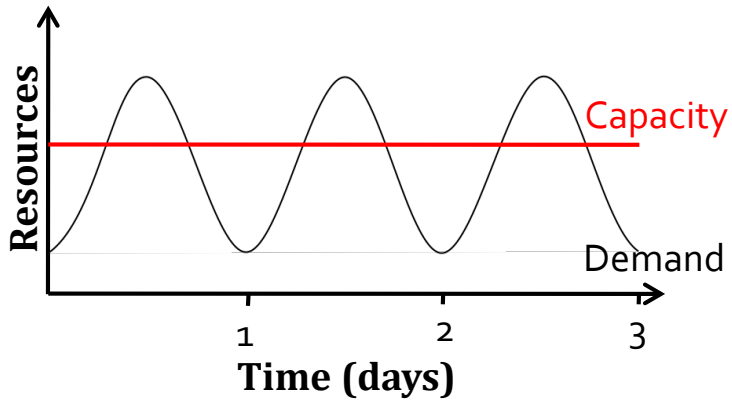






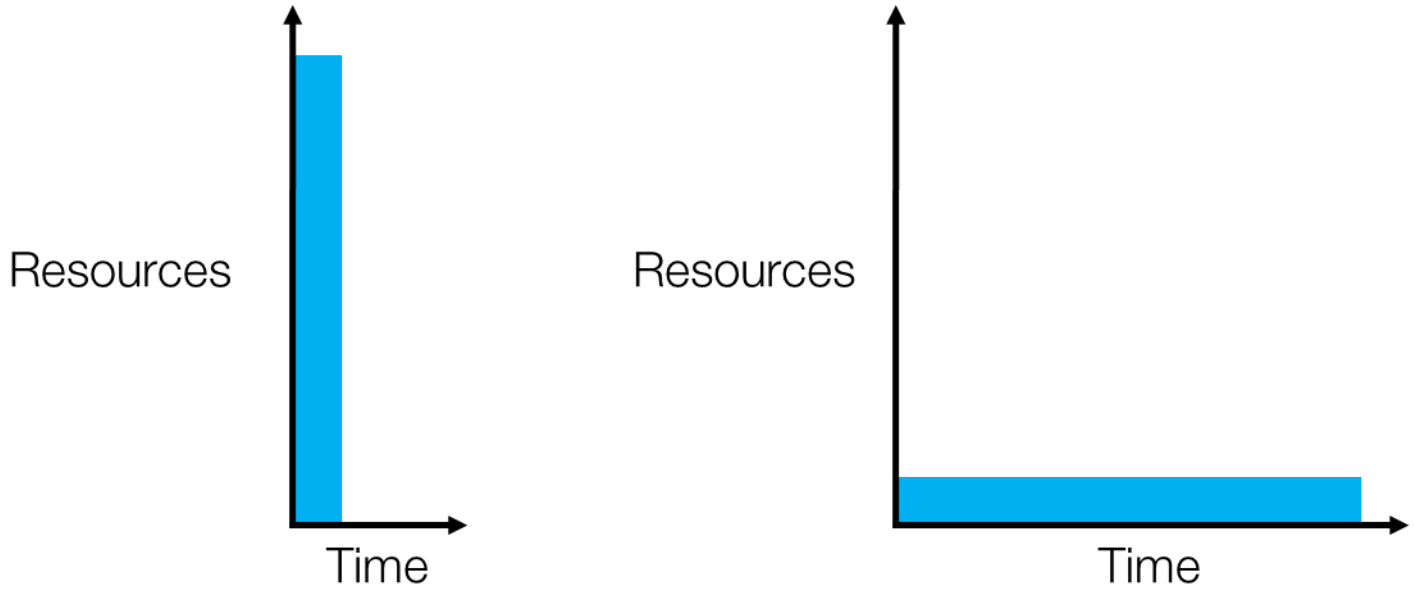
Cloud Economics: For Users

- Pay-as-you-go (usage-based) pricing:
  - Most services charge per minute, per byte, etc
  - No minimum or up-front fee
  - Helpful when apps have variable utilization



Cloud Economics: For Users

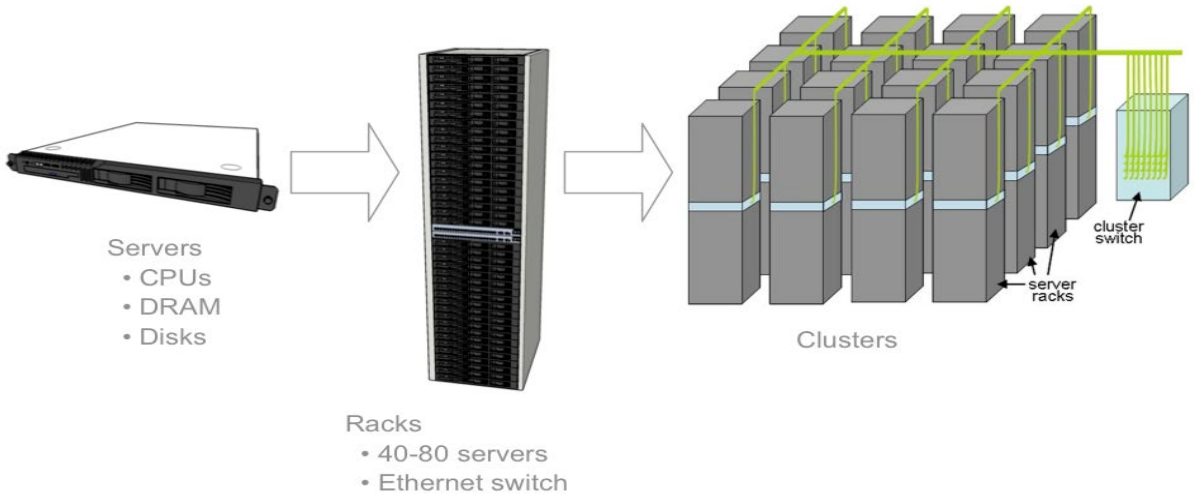
- **Elasticity:**
  - Using 1000 servers for 1 hour costs the same as 1 server for 1000 hours
  - Same price to get a result faster!



Cloud Economics: For Providers

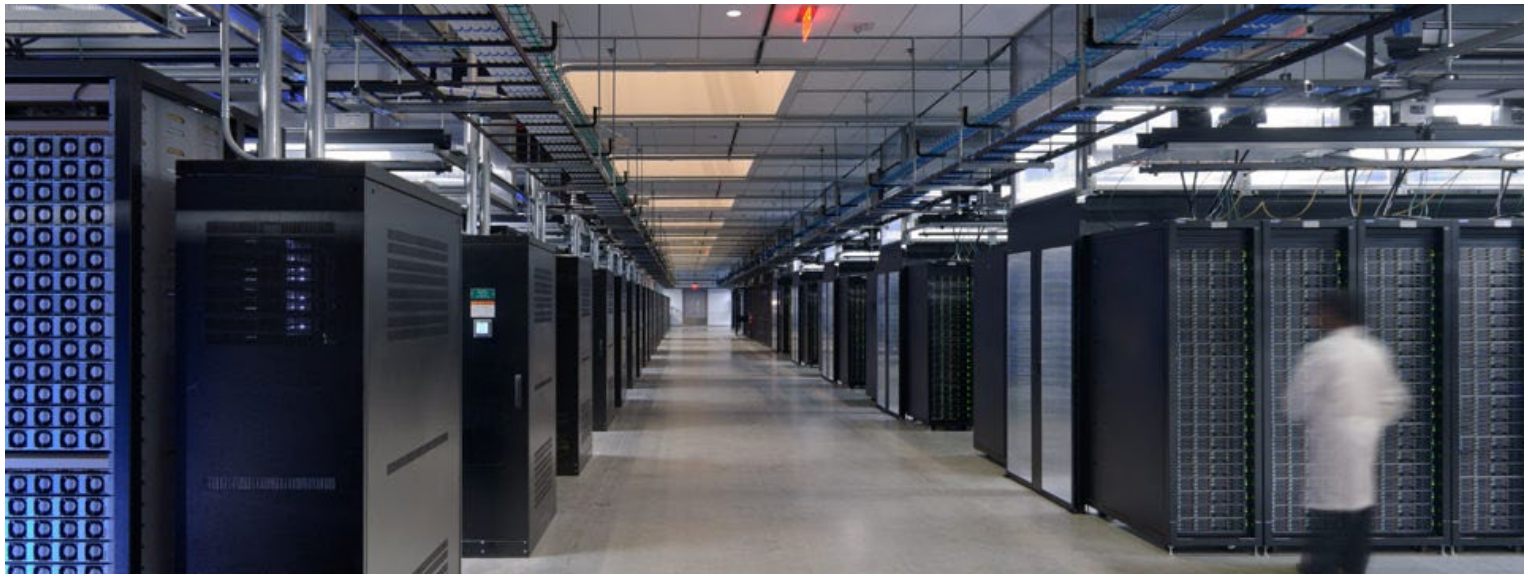
- **Economies of scale:**
  - Purchasing, powering & managing machines at scale gives lower per-unit costs than customers'
  - Tradeoff: fast growth vs efficiency
  - Tradeoff: flexibility vs cost

Datacenter Hardware



- Rows of rack-mounted servers
- Datacenter: 50 – 200K of servers, 10 – 100MW
- Often organized as few and mostly independent clusters

Datacenter Example



IaaS clouds:

- On-demand (pay-as-you-go)
  - Static hourly rate x run hours = Subscription
- Reserved instance
  - One-time subscription fee
  - Free/discounted usage fee during the reservation period
- Spot instance
  - Users bid for computing instances
  - No service guarantee