



CS 6111

Cloud Computing

Introduction

Welcome to CS6111: Cloud Computing!



Welcome!

- Instructor: Haiying Shen
- Associate Professor at Dept. of Computer Science
- Research interests:
 - Cloud computing and datacenters, Big data, Cyber-physical systems, Distributed systems, Machine learning applications
 - Email: hs6ms@virginia.edu
- Office: Rice Hall 303
- Office Hours: 3:15 - 4:15 MoWe; other times by appointment, [Zoom](#), Passcode: 312954
- Website: <http://www.cs.virginia.edu/~hs6ms>

- TAs:

- Alireza Namazi, mez4em@virginia.edu

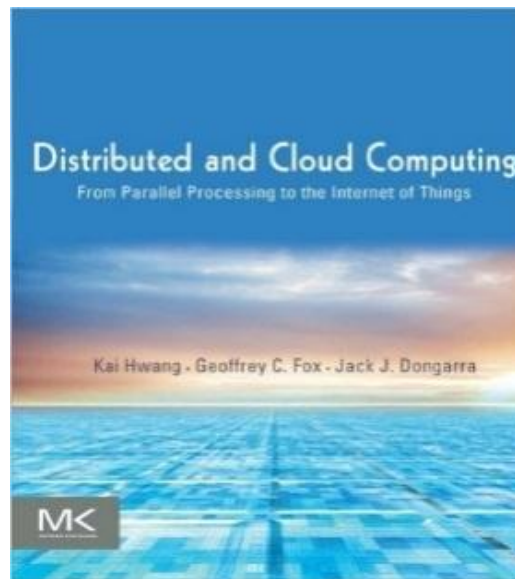
Office Hour: 3:15 - 4:15 TuTh; other times by appointment, [Zoom](#), Passcode: 312954

Everything will be in canvas

- <https://canvas.virginia.edu/>

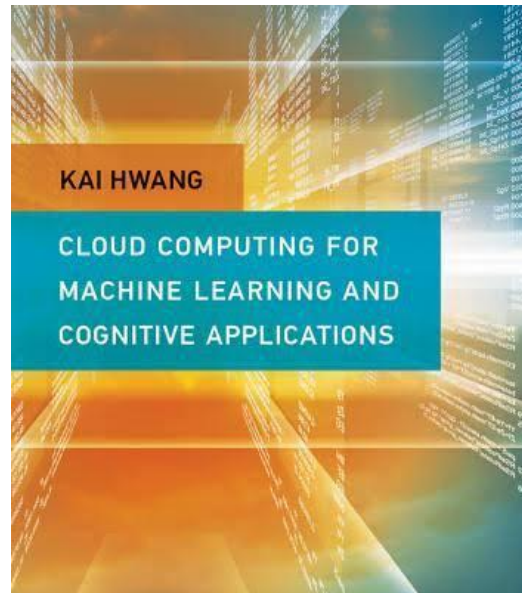
Reference book

- K. Hwang, G. Fox and J. Dongarra,
Distributed and Cloud Computing: From
Parallel Processing to the Internet of Things



Reference book

- K. Hwang, Cloud Computing for Machine Learning and Cognitive Applications, The MIT Press, 2017. (ASIN: B073RX8B2Y)



Course Materials

- Recent papers from top conferences like SysML, ATC, CoNext, NSDI, OSDI, SigComm
- Papers from different categories are listed
- Considering the recent developments in field of natural language processing (e.g. ChatGPT), LLMs will be a special focus.
- Cloud Systems for Machine Learning

LLM



Machine Learning (ML) Applications



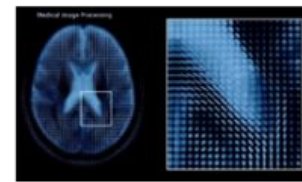
Self-driving



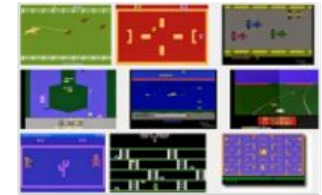
Surveillance detection



Translation



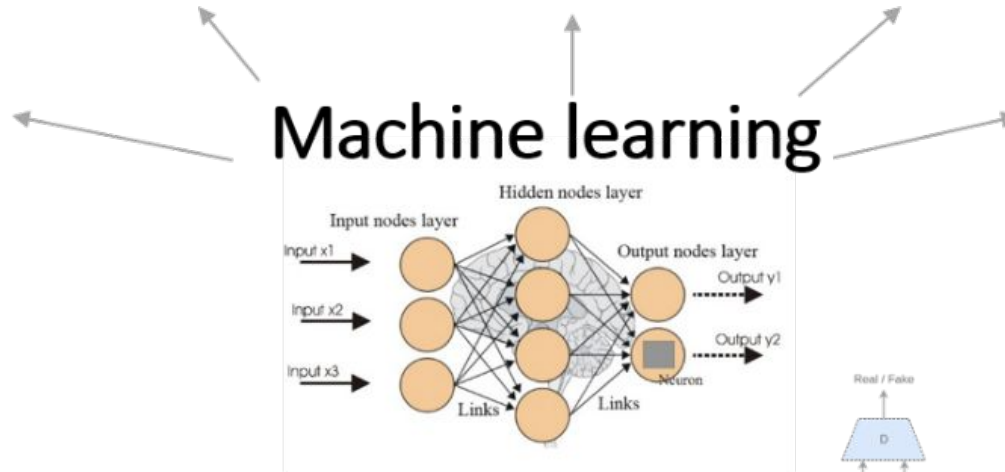
Medical diagnostics



Game



Personal assistant



Art

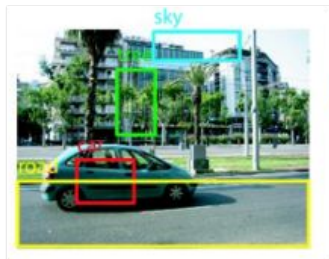


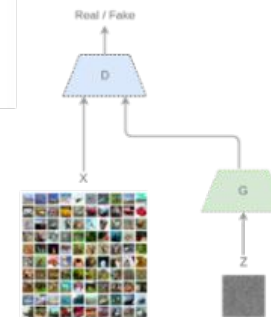
Image recognition



Speech recognition



Natural language

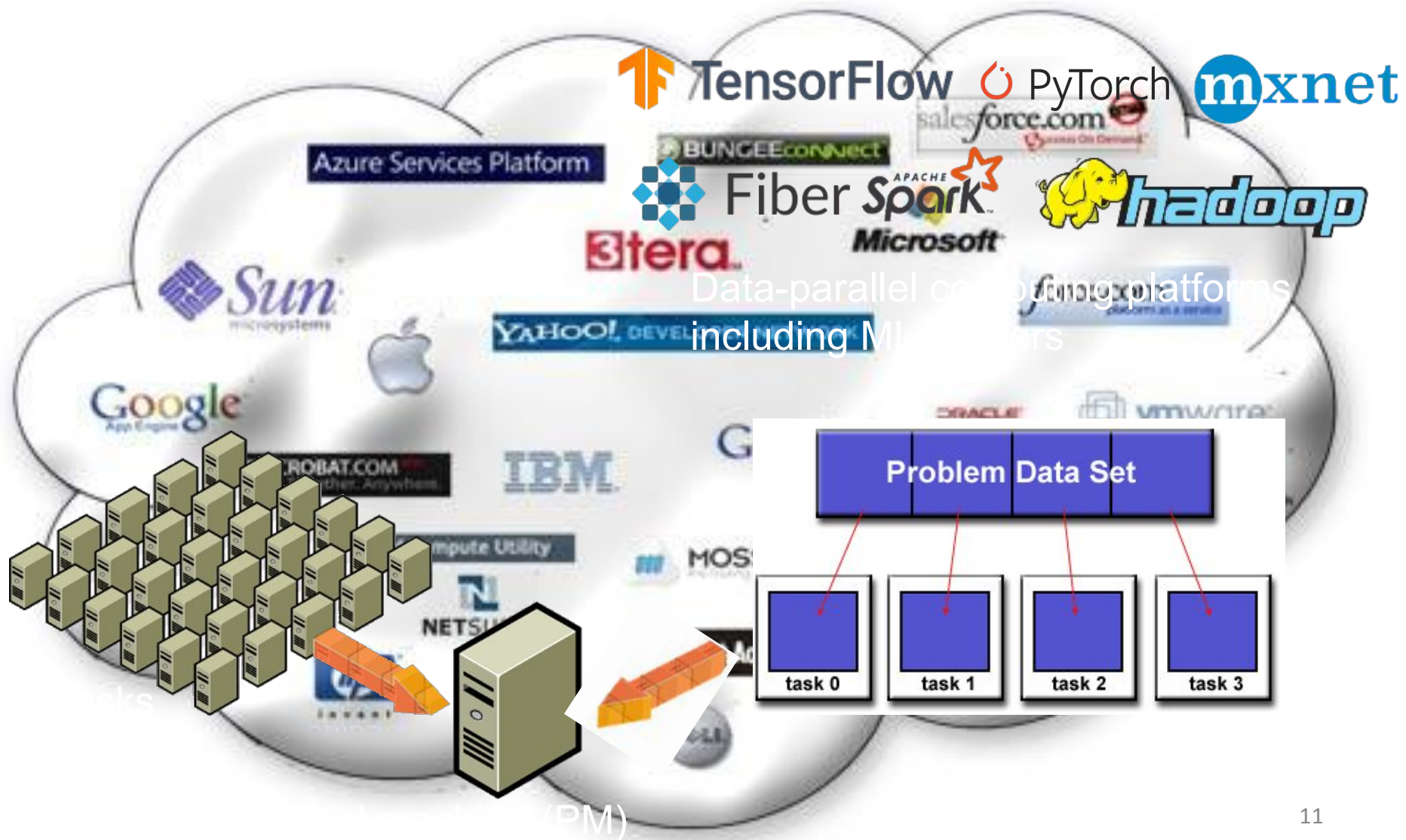


Generative model



Reinforcement learning

Computing Platforms



Job Scheduling

Goal:

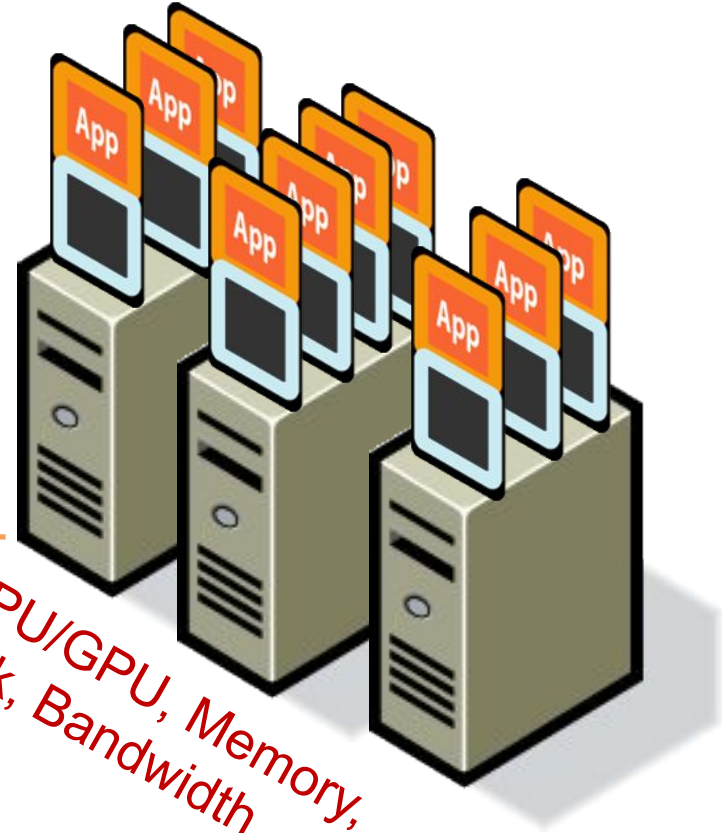
Improve job performance

Reduce system cost



Resource management:

CPU/GPU, Memory,
Disk, Bandwidth



TensorFlow



PyTorch

mxnet



Fiber



APACHE
spark



hadoop

Job Scheduling

Goal:

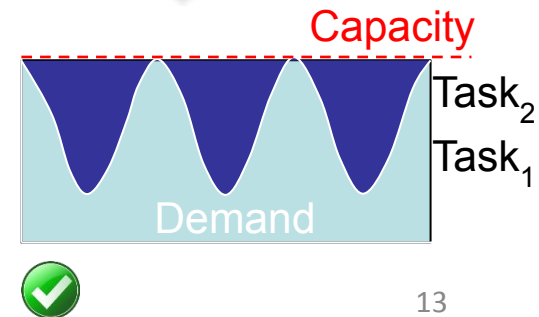
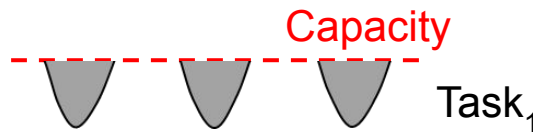
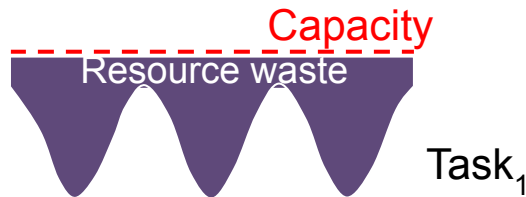
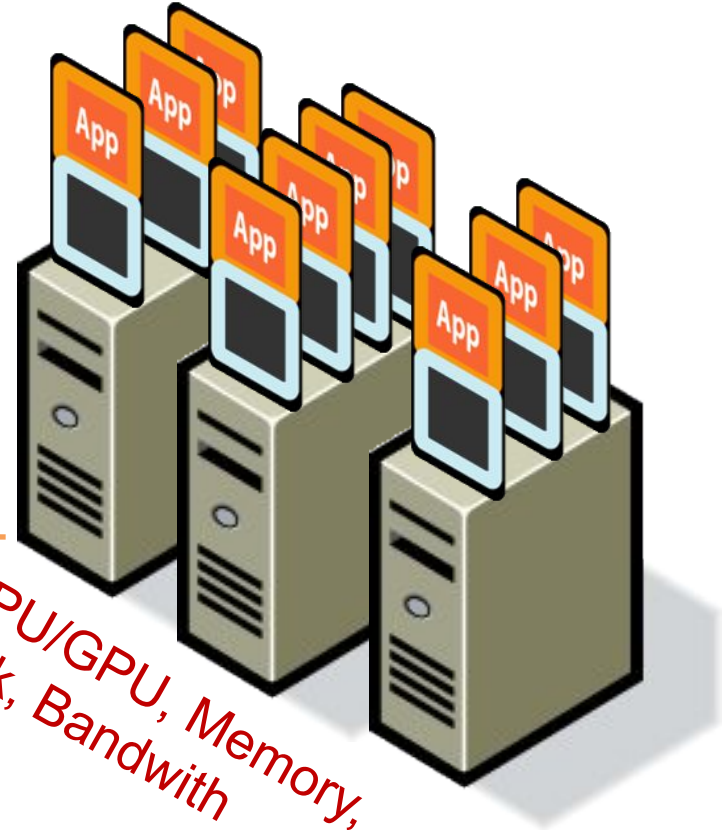
Improve job performance

Reduce system cost



Resource management:

CPU/GPU, Memory,
Disk, Bandwidth



Job Scheduling

Task scheduling

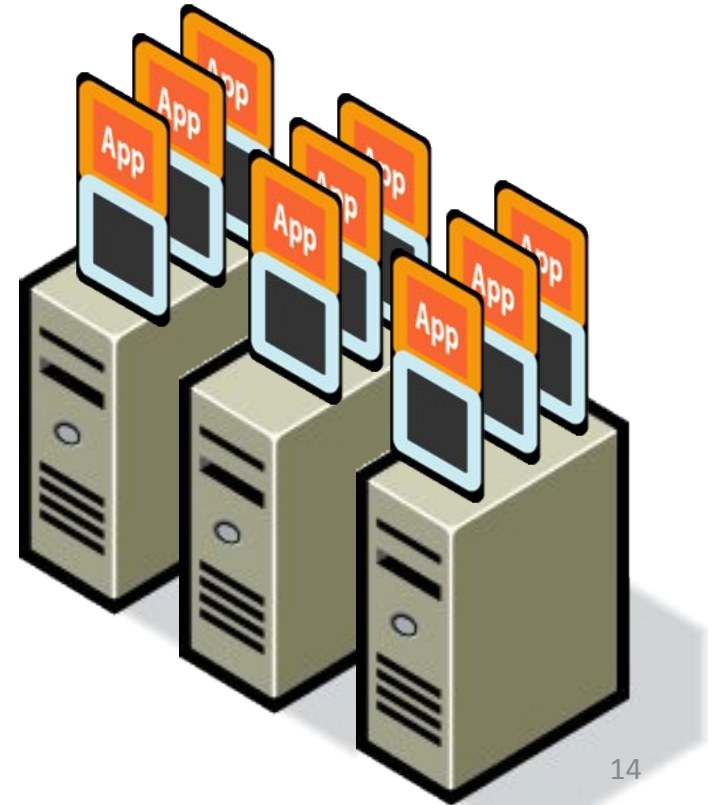


Load balancing
by **task migration**

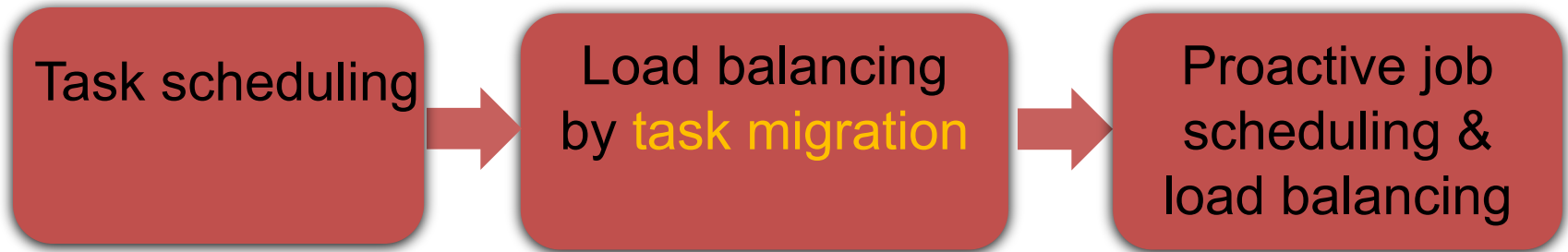


Proactive job
scheduling &
load balancing

Task rescheduling



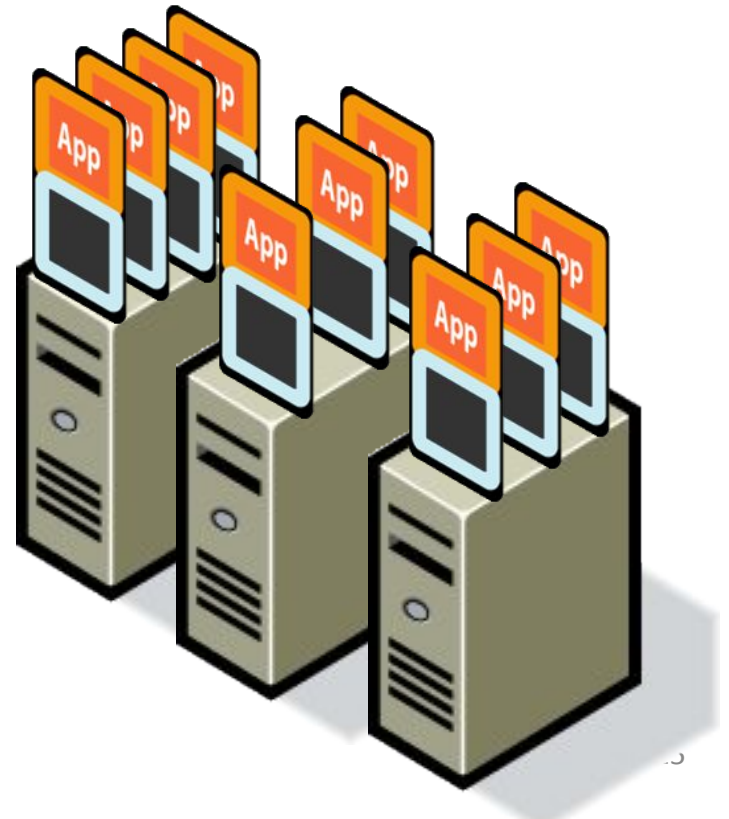
Job Scheduling



Over-loaded PMs



Under-loaded PMs



Paper Discussion Session

(3 papers from the previous 2 classes)



Jury

Vote based on the understanding of the paper and the thoughts beyond

Each student in the winner group gains 1 credit in the final grade

Program committee



Authors



Project

- About project

What is Cloud Computing?

- Cloud computing: large groups of remote servers networked to allow centralized data storage and online access to computer services or resources



- Why cloud computing?
 - Classical computing
 - Buy & Own, Install, Manage
 - High capital expenditure
 - Cloud computing
 - Subscribe & Use
 - Pay as you go



What is Cloud Computing?



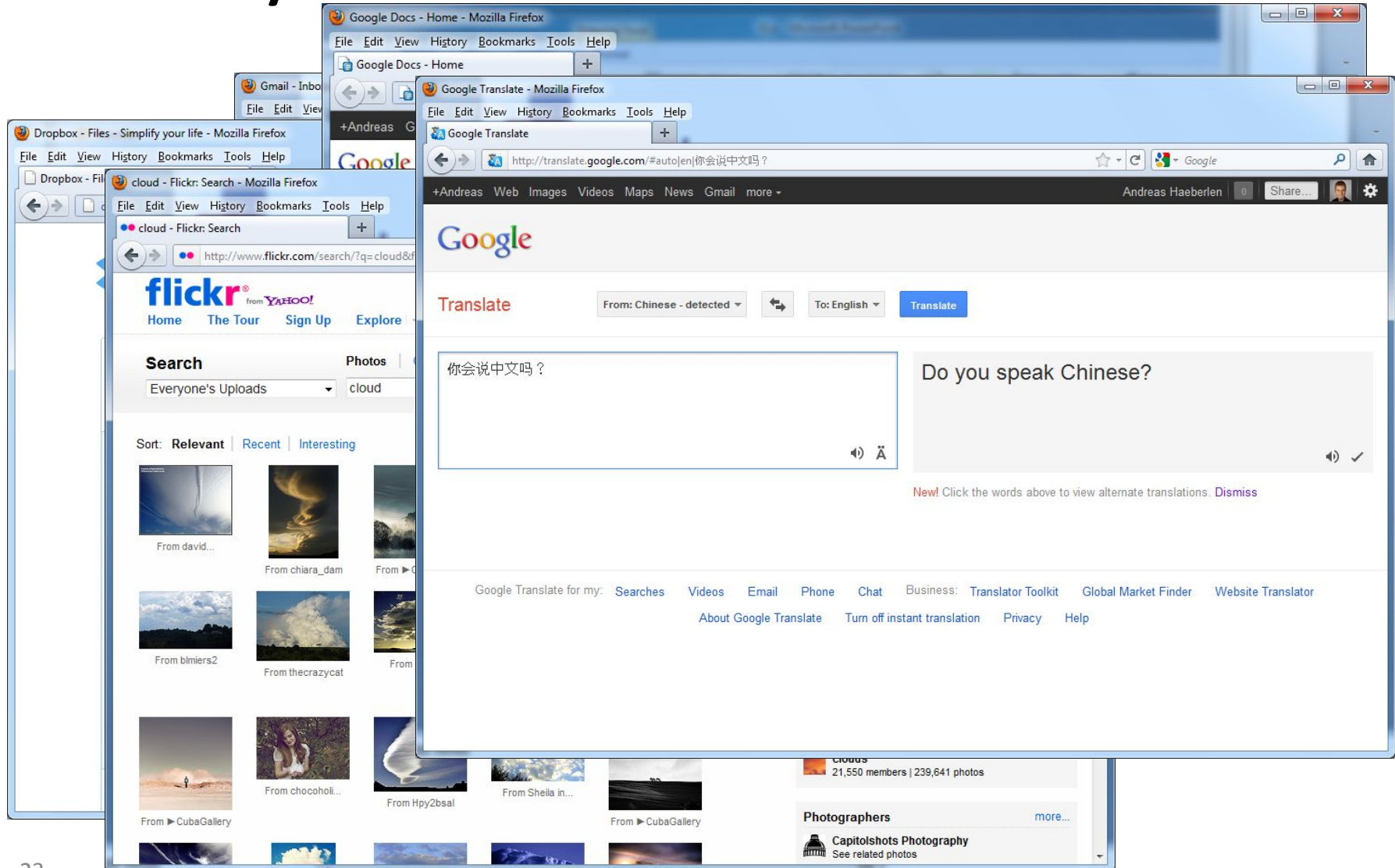
Do You Use Cloud?



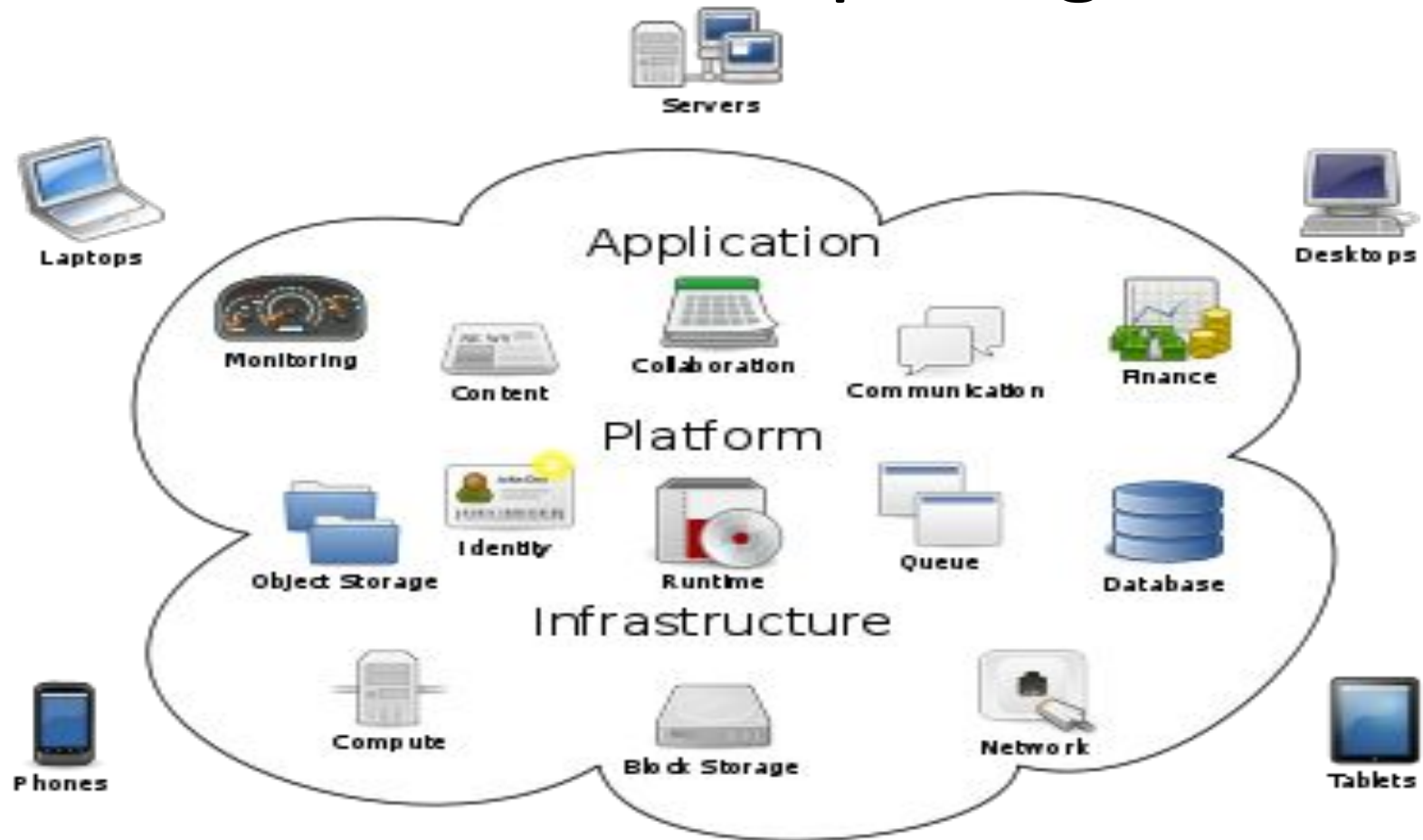
Cloud computing

- Definition:
 - Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software, and information are provided to computers and other devices as a metered service over a network (typically the Internet).
- Architecture:
 - **The Intercloud:** an interconnected global "cloud of clouds"
 - **Cloud engineering:** the application of [engineering](#) disciplines to cloud computing

Have you used 'the cloud' before?



Cloud computing



Cloud Computing

Cloud computing

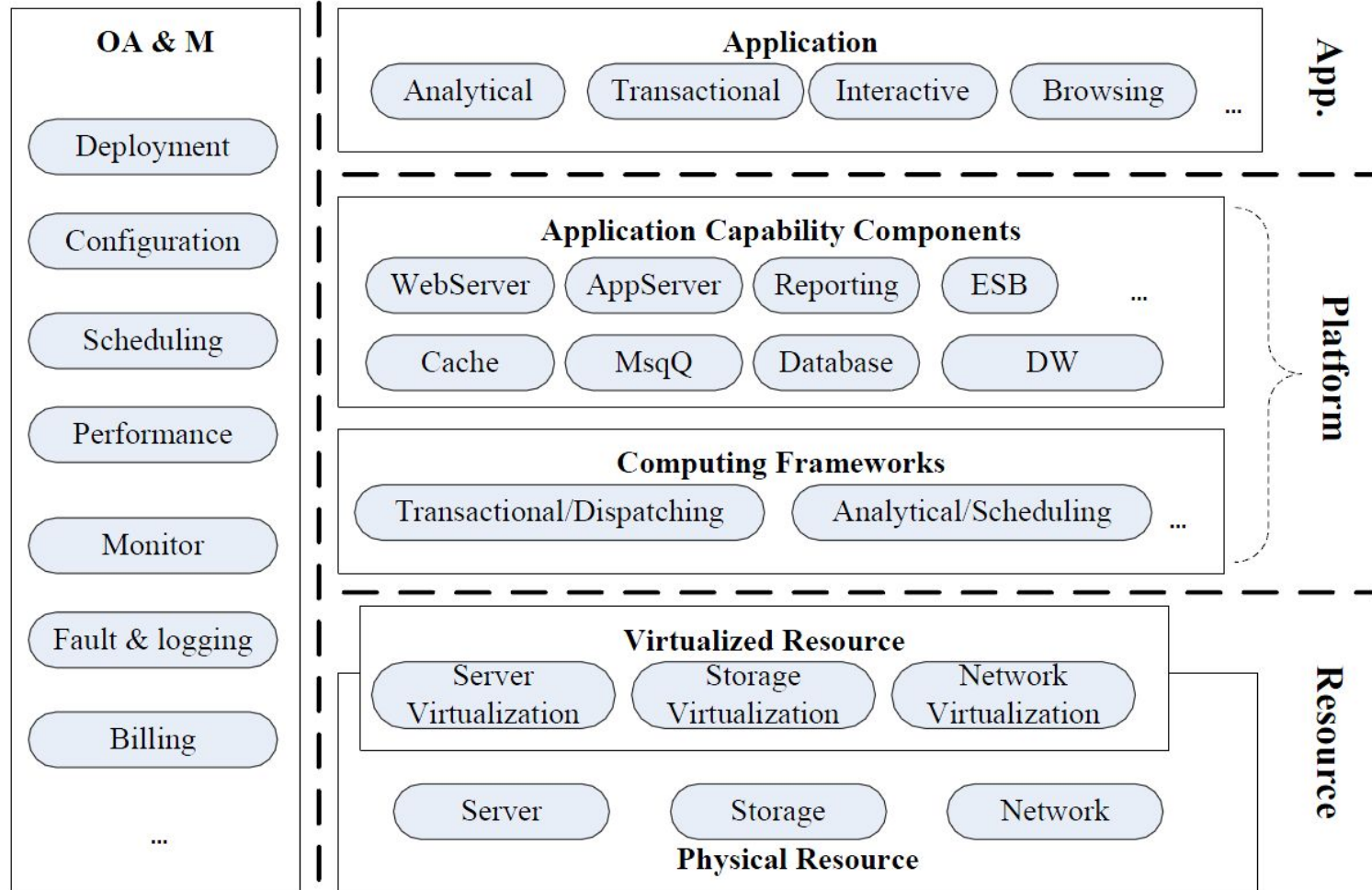


Figure 2: Reference cloud computing architecture

Note

- This course belongs to Systems rather than Application.
- If you would like to learn how to use AWS services, this course is not a right choice.

Note

- For the students in the waiting list



Questions?