Cloud Computing
Introduction & Course Logistics

Minchen Yu

SDS@CUHK-SZ

Fall 2024

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

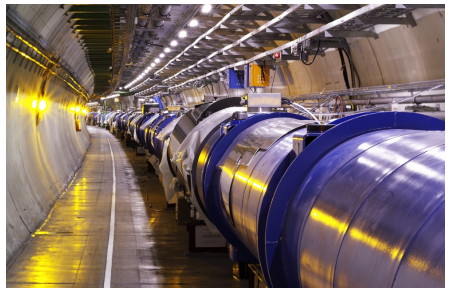
SCHOOL OF
DATA SCIENCE
數據科學學院

About me

- Minchen Yu, Assistant Professor, SDS
  - Email: yuminchen@cuhk.edu.cn
  - Office: DaoYuan 420c
  - Office hour: Wednesday 2-3 PM

- Research interests
  - Cloud computing systems, with particular focus on serverless computing, big data and machine learning systems

Data, data, data!



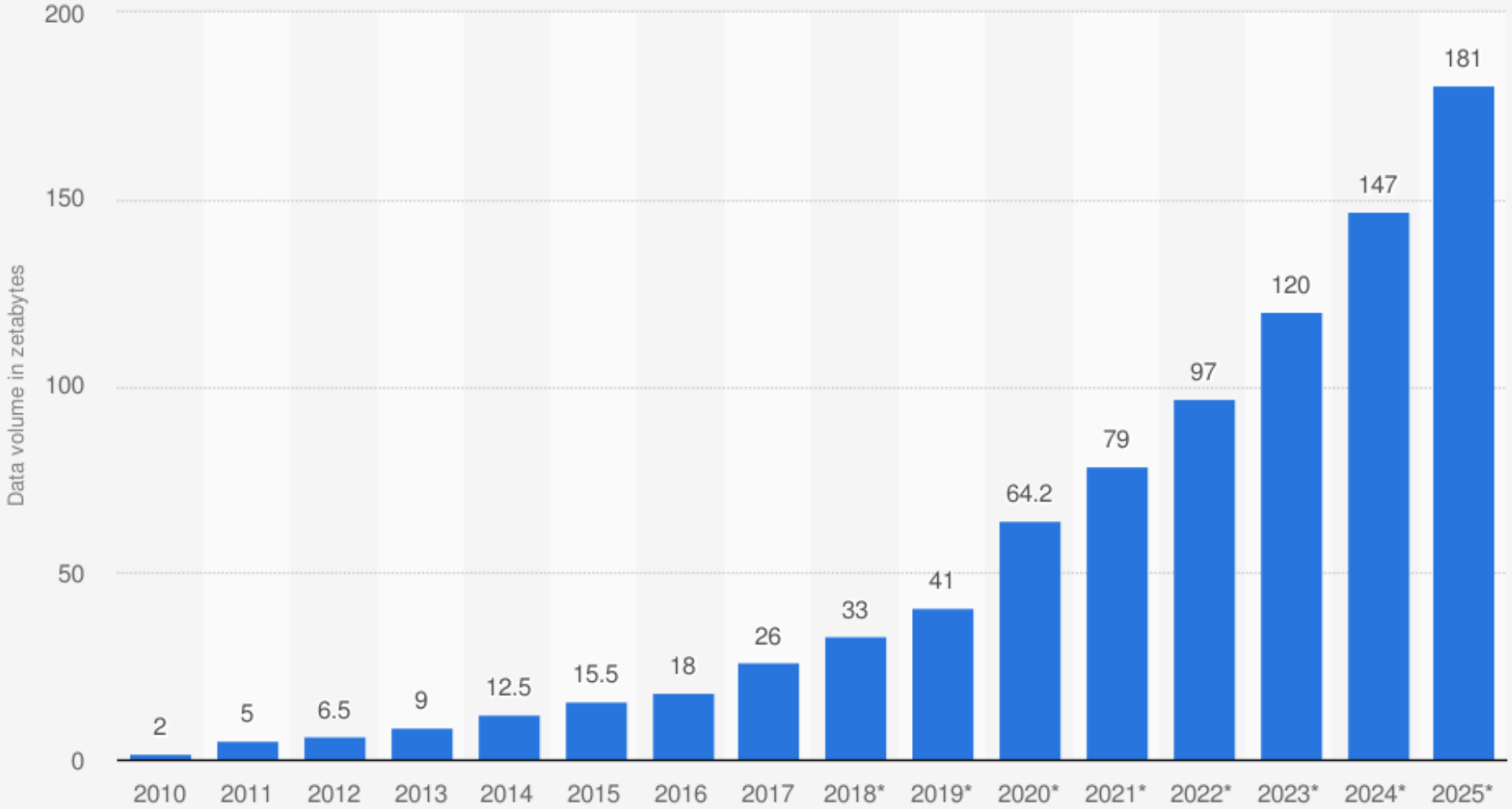
Large Hadron Collider generates
40 TB data per second

Boeing Jet Engine creates 10 TB
operation information every 30 minutes

Google

Search index contains 100s billions
($>10^{11}$) webpages and is well over 100
petabytes ($>10^{17}$) in size

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (in zettabytes)



Sources
IDC; Seagate; Statista estimates

Additional Information:
Worldwide; 2010 to 2020

“640K ought to be enough for anybody.”
— Bill Gates (1981)

How can we crunch the massive amount of data?


Cloud Datacenter
Google
google.com/datacenters

Datacenters

- >100K servers
- Costs in billions of dollars
- Geographically distributed

It is estimated that >94% global workloads has been processed in datacenters in 2021

Estimated # servers

> 1M

Microsoft ~ 1M

YAHOO!

facebook

amazon

Several 100,000s each

Source: http://www.datacenterknowledge.com/archives/2013/07/15/ballmer-microsoft-has-1-million-servers/

“I think there is a world market for maybe five computers.”

— Thomas Watson (1943)

Utility Computing

Computing may someday be organized as a public utility just as the telephone system is a public utility … Each subscriber needs to pay only for the capacity he actually uses, but he has access to all programming languages characteristic of a very large system … The computer utility could become the basis of a new and important industry

— John McCarthy, 1961

Cloud Computing: One Step Forward

- **Computing as a utility:** deliver computing resources over the Internet, as a metered service
  - Dynamic provisioning: pay-as-you-go
  - Scalability: “infinite” capacity
  - Elasticity: scale up or down

amazon web services | EC2

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.micro	1	Variable	1	EBS Only	$0.013 per Hour
t2.small	1	Variable	2	EBS Only	$0.026 per Hour
t2.medium	2	Variable	4	EBS Only	$0.052 per Hour
t2.large	2	Variable	8	EBS Only	$0.104 per Hour
m4.large	2	6.5	8	EBS Only	$0.126 per Hour
m4.xlarge	4	13	16	EBS Only	$0.252 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.504 per Hour
m4.4xlarge	16	53.5	64	EBS Only	$1.008 per Hour
m4.10xlarge	40	124.5	160	EBS Only	$2.52 per Hour
m3.medium	1	3	3.75	1 x 4 SSD	$0.067 per Hour
m3.large	2	6.5	7.5	1 x 32 SSD	$0.133 per Hour
m3.xlarge	4	13	15	2 x 40 SSD	$0.266 per Hour
m3.2xlarge	8	26	30	2 x 80 SSD	$0.532 per Hour

Now that we have computing resources in cloud. What's next?


Big data systems: OS for the cloud

The datacenter is a computer

OS X Yosemite

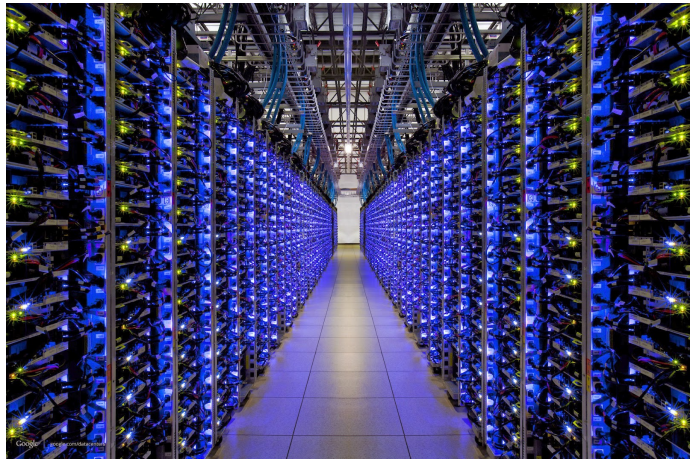
Windows

hadoop

Spark

TensorFlow

kubernetes



About the course

- Website: see Blackboard
- Announcements and course materials are posted online on a regular basis
- TA: Zhiqing Zhong
- USTF: Juan Albert Wibowo

Prerequisites

- Comfortable with Unix/Linux
- Comfortable with Python programming

Accept an email invitation to get enrolled into an AWS Academy Learner Lab

A note about EC2

- Amazon EC2 credit
  - $100 USD credit from AWS Academy Classroom upon enrollment
  - AWS free tier: 750 hours per month for the **first 12 months**
- Shut down your virtual instances whenever you are done
- Extra usage at your own cost. We cannot financially help you in any means!

Textbook/References

- No official textbook
  - Cloud computing are emerging technologies under heavy development.
- References (details on the course website):
  - Armbrust et al., “Above the clouds: A berkeley view of cloud computing,” Technical report, UC Berkeley, 2009.
  - Chambers et al., “Spark: The Definitive Guide,” O’Reilly, 2018.
  - State-of-the-art research papers (posted online)
- Learn things online

Assessment

- Assignments
  - 25%
  - Programming
- Open-ended course project
  - 30%
- Final exam
  - 45%

Course project

- Teamwork
  - a group of up to 3 students
- Open-ended (sample topics available as well)
  - Tech/business/research
  - Any topics related to cloud computing systems
- Maximum 5-page project report + demo video

Course project

- Example topics
  - Analyzing public cloud traces (e.g., AWS Public Datasets, Google Cluster Workload traces)
  - Performance benchmarking of cloud services (e.g., EC2, Lambda)
  - Building your own systems/applications atop cloud services
  - Performance optimizations of cloud services (e.g., based on open-source versions)
  - Topics in cloud competitions are welcome (e.g., Alibaba Tianchi)

  …

Tips

- Attend the lecture and tutorial
- Learn things online
- Do all assignments by yourself
- Do well in the final

Learn cloud computing by trying it yourself!

Academic honesty

- In short, don't cheat!
- Don't copy code or solutions from your classmates or third-party sources, and don't let others copy yours. Both cases are plagiarism and penalized in the same way

Protocol for Plagiarism

- TAs will detect possible plagiarism in your code/reports.
- Suspicious cases will be directly reported to the school.
- Minimum penalty: zero mark for the assignment.

Objectives

- Understand the following aspects of cloud computing
  - Characteristics and service models
  - Underlying cloud technology in computing, storage, networking, resource management, etc.
  - State-of-the-art big data systems, e.g., Spark
  - GPUs in the cloud

Tentative lecture schedule

Week	Topic
1	Logistics, cloud concept and fundamentals
2	Service models
3	Virtualization
4	Container & Kubernetes
5	Serverless computing
6	Cloud storage systems
7	MapReduce
8	Spark
10	Cluster management
11	GPU cloud
12	Guest lecture, sky computing
13	Review
14	Final exam

Subject to change.
Check back frequently!

Tentative tutorial schedule

- Monday 6-6:50 PM
- Teaching C, 208
- The first tutorial starts next week (9.9)
- Tutorials may be skipped with notification in advance

Any questions?