Advanced Cloud Computing

Introduction & Logistics

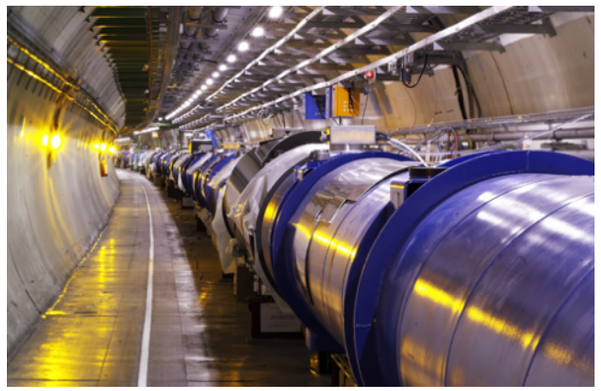
Wei Wang
CSE@HKUST
Spring 2025

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

About me

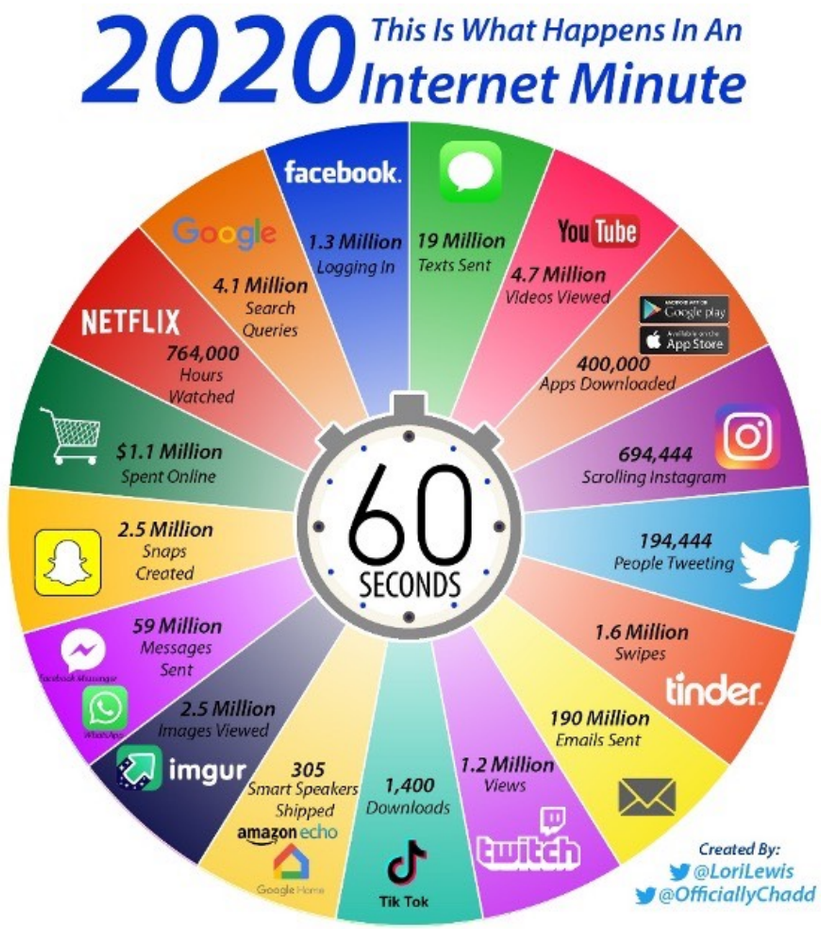
- **Wei Wang**, Associate Professor, CSE
  - PhD (2015) in Computer Engineering, University of Toronto
  - Email: weiwa@cse.ust.hk
  - Office: Rm2531
- Research interests
  - Distributed systems, with particular focus on cloud computing, big data and machine learning systems

Data, data, data!



Large Hadron Collider generates 40 TB data per second

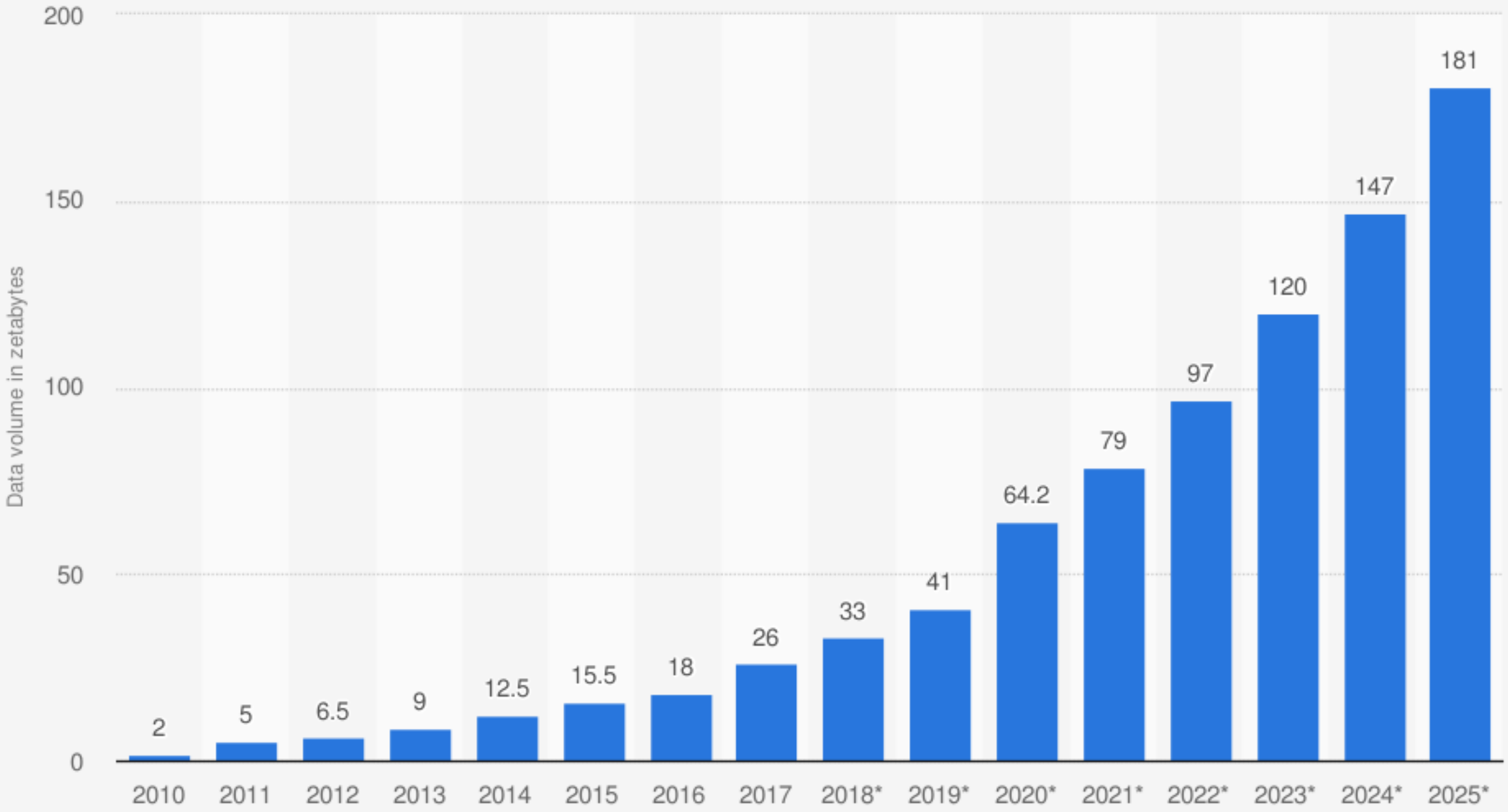
Boeing Jet Engine creates 10 TB operation information every 30 minutes



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

Datacenters

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

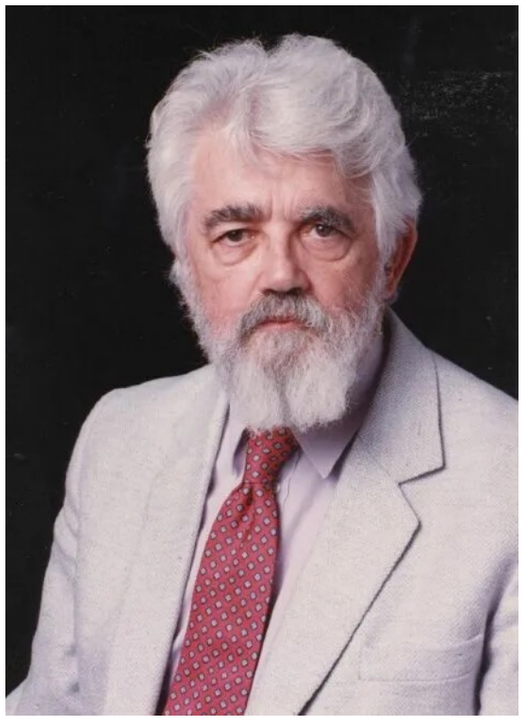
It is estimated that >94% global workloads was processed in datacenters in 2021

“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

- Allows everyone access to
  - unlimited computing resources
  - unlimited storage resources
  - All delivered over internet
- No upfront cost
  - pay only for what you use

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?

OS for the cloud: Cluster management systems & distributed computing frameworks (for big data and machine learning)

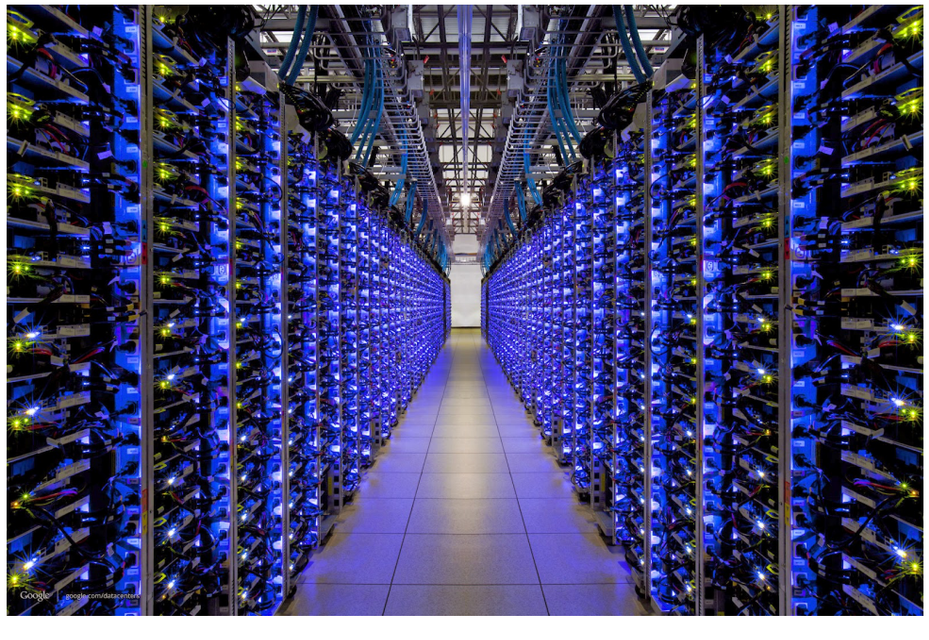
The datacenter **is** a computer

Windows

OS X Yosemite

hadoop Spark TensorFlow

kubernetes



About the course

- Website: in Canvas, CSIT5970
- Announcements and course materials are posted online on a regular basis
- TA:
  - Tianyuan Wu (twubt@connect.ust.hk)

Prerequisites

- Object-oriented programming
- Data structures and algorithms
- Comfortable with Python/Java programming
- Comfortable with Unix/Linux
- A laptop that can host a Linux VM with at least 2 cores and 8 GB RAM (or a Mac/Linux laptop)
  - has **`VirtualBox`** installed
  - has **`Git`** installed

Accept an email invitation from AWS Academy Learner Lab to get $100 USD credit

A note on lab environment

- AWS Academy Learner Lab
  - Offers a long-running lab environment for students to learn cloud and AWS production services
    - Each student will have a **$100** AWS Platform Credit to spare
    - Have access to a **restricted set of AWS services**
      - we will mainly use EC2, S3, EMR, and Lambda
- Extra expense at your own cost. **We (me, or the dept) cannot financially help you in any means!**

Textbook/References

- No official textbook
  - Cloud computing is a rapidly evolving technology
- Best way to learn it is to read research papers
  - Landmark and cutting-edge research work that developed the cloud technology
  - e.g., Google Filesystem, MapReduce, Spark
  - Reading list will be posted online

Learning by reading & doing,
and learn things online!

Assessment

- Homework and labs (30%)
  - programming on AWS and locally hosted VMs
- Midterm exam (30%)
- Open-ended course project (40%)
- No final

Course project

- Teamwork
  - a group of 2-4 students
- Open-ended (sample topics available as well)
  - Must be related to cloud computing
  - Engineering/research
- Project report (25%) + video presentation (15%)

Course project

- Example topics
  - Data analytics on public datasets (e.g., AWS Public Datasets)
  - A cloud-based service application (e.g., reimplementing MapReduce/Spark framework using AWS Lambda)
  - Cloud resource management and scheduling using Kubernetes
  - Reproducing a published cloud system work
  - …

To pass

- Attend the lecture and tutorial
- Get your hands dirty
- Learn things online
- Do all assignments by yourself
- Do well in the exam

You cannot have a good understanding of Cloud without 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

- We will detect possible plagiarism in your code/reports.
- Suspicious cases will be directly reported to the CS general office. A panel will be formed to deal with all cases.
- Minimum penalty: zero mark for the assignment/ homework.

Tentative lecture schedule

Week	Topic
1	Logistics, Cloud concept and characteristics
2	Cloud fundamentals and Service models
3	Virtualization
4	Cloud Storage Systems
5	MapReduce
6	Hadoop
7	MapReduce Algorithm Design
8	RDD and Spark
9	Spark Programming
10	Graph Analytics
11	Distributed Machine Learning
12	Serverless Computing
13	Advanced Topics

Other matters

- Try to come on time
- Participate as much as you can in the classroom. It’s a two-way avenue.

S. Keshav, “How to Read a Paper,” ACM SIGCOMM Comput. Comm. Rev. 2007

The three-pass approach

- **The first pass** (5 - 10 min): get the general idea of the paper

- If needed, go to **the second pass** (1 hour): grasp the paper's content, but not details

- If needed, go to **the third pass** (several hours or days): *virtually re-implement* the ideas and technical details

The first pass is to get a bird's eye-view of the paper (5 - 10 min)

The first pass

- Carefully read the title, abstract and introduction
- Only read the section and sub-section headings
- Read the conclusions
- Glance over the references

Able to answer the five C's

- **Category:** What type of paper is this? Measurement, theory, system, protocol, algorithm, or a survey?

- **Context:** Which other paper is it related to?

- **Correctness:** Do the assumptions appear to be valid?

- **Contributions:** What are the main contributions? Are they significant?

- **Clarity:** Is the paper well written?

Reasons NOT to read further

- Not interesting or irrelevant to my research
- Technically unsatisfied
- The assumptions appear to be invalid
- Not well written or poorly organized
- The contributions seem to be incremental

The second pass: read with greater care but not every detail (1 hour)

The second pass

- Grasp the content while ignoring technical details such as proofs and implementation

- Pay special attention to the figures, diagrams and other illustrations — they contain important information based on which the conclusions are drawn

- Mark relevant unread references for further reading

Able to summarize the main thrust

- Is the paper solving a “right” problem?
- Are the claimed contributions significant/valid with convincing supporting evidence?
- Is the approach/evaluation technically sound and novel?
- What is the potential impact of the paper?

Do I need to go to the third pass to digest the technical details?

Yes, only if

- You are interested in the technical details and have time
- You want to do some followup work
- The results are groundbreaking but somehow out of surprise or counter-intuitive
- The proof techniques, implementation details, and/or experiments turn out to be useful

The third pass: *virtually re-implement* the paper (several hours or days)

Recap

- **The first pass** (5 - 10 min): get the general idea of the paper

- If needed, go to **the second pass** (1 hour): grasp the paper's content, but not details

- If needed, go to **the third pass** (several hours): *virtually re-implement* the ideas and technical details