COMP 6611B:
Topics on Cloud Computing and Data Analytics Systems

Wei Wang
Department of Computer Science & Engineering
HKUST
Fall 2015

Data, data, data!

Large Hadron Collider generates 40 TB data per second

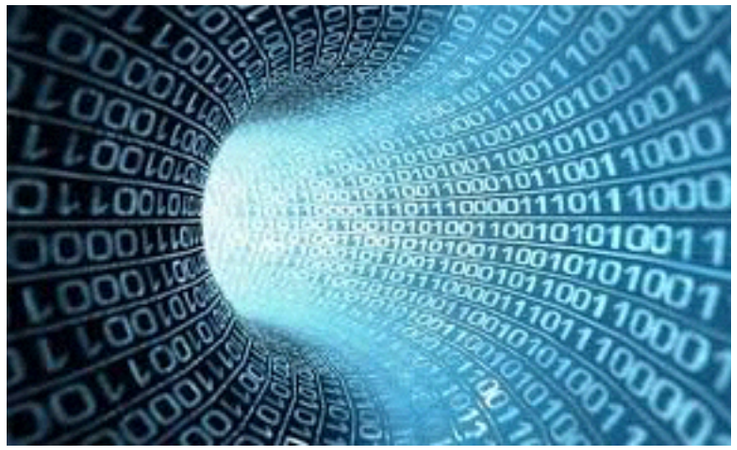
Boeing Jet Engine creates 10 TB operation information every 30 minutes

Google

Crawls 20B web pages a single day (2012)

YAHOO!

Hadoop cluster: 330K nodes, 365 PB (2014)



1.8 ZB (10^21) data created in 2011, doubling the amount of data generated in 2010

amazon.com web services

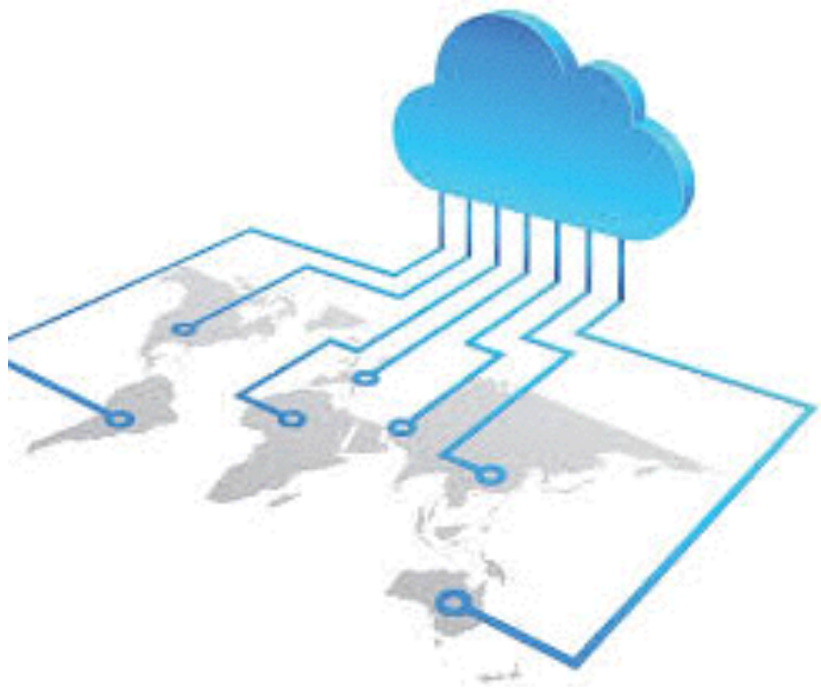
1.1M requests per second, 2T objects (2013)

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

How can we process the massive amount of data?

Cloud Computing

- **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

Cloud Datacenter

google.com/datacenters

Datacenters

- >10K servers
- Costs in billions of dollars
- Geographically distributed

Estimated # servers

Google > 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, Head of IBM (1943)

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

Big data systems: OS for the cloud

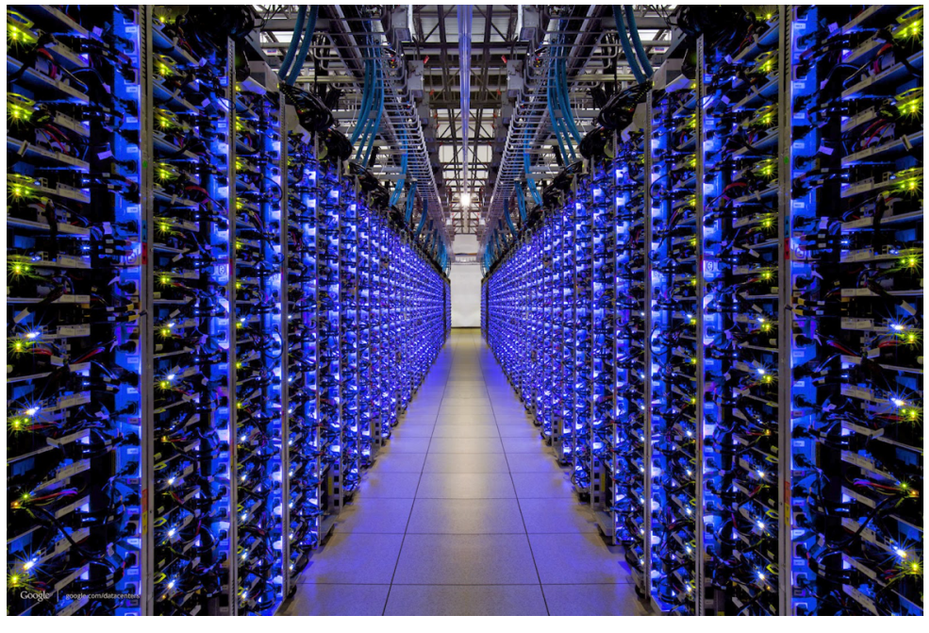
The datacenter **is** a computer

Windows

OS X Yosemite

hadoop

Spark



Focus of this course

Focus of this course

- Examine advanced research topics in cloud systems, data processing frameworks, networking, storage, etc.

- Understanding the key challenges that arise in the architecture design, system implementation, and performance optimization

Paper reading-based seminar course

Reading list

- ~30 top conference papers covering various research topics
  - Datacenter architecture
  - State-of-the-art data processing frameworks
  - Workload characteristics
  - Resource management and scheduling

http://www.cse.ust.hk/~weiwa/teaching/Fall15-COMP6611B/readinglist.html

Course requirements

Paper reading

- Each week covers a group of papers focusing on a specific research topic

- Before the class
  - Read **all** papers
  - Choose one to write a **review** and submit it to the instructor's email: weiwa@cse.ust.hk

Paper review

- Paper summary
- Strengths
- Weaknesses
- Detailed comments

Paper presentation

- Each student will present **at least one** paper
- In the Monday lecture, we will determine the presenters and papers to be presented in the Friday lecture and Monday lecture in the following week
- Maximum **25 min** for each presentation
- We will **randomly** choose students to ask/answer questions after the presentation

Course project

- Term-long, open-ended course project
- Topics depend on you, but must be approved by the instructor
  - Sample topics will be provided
- Work alone or collaborate with another student

The delivery

- One page proposal due at **the end of week 3**
- **3-page** midterm report
- **6-page** course thesis at the end of the term
- Final presentation

Final presentation

- 10 min for the single-author work, 15 min for the collaboration work
  - The time allocation depends on you
- Marked by both the instructor and the audiences

Grading

- Class participation and discussion: 10%
- Paper review: 20%
- Presentation (including papers and project thesis): 25%
- Course project: 45%
  - Proposal: 5%
  - Midterm report: 10%
  - Final thesis: 20%

Questions?

http://www.cse.ust.hk/~weiwa/teaching/Fall15-COMP6611B/home.html

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): *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 assumption appear to be valid?

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

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

Now decide if it is needed to go to the second pass with more details

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

Take away: The paper will never be read if the problem and/or the contributions cannot be understood in five minutes.

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?

You may get an idea why the paper is accepted

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)

The third pass

- Make the same assumptions as the authors, re-create the work
  - Identify and challenge every assumption in every statement
  - How would I solve the problem and do the experiment?
  - How would I present the paper if I were to write it?

You should able to

- Reconstruct the entire structure of the paper
- Identify the strong and weak points, e.g.,
  - implicit assumptions
  - miss citations
  - potential issues with experimental or analytical techniques

The weak points might suggest a new problem for further research!

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