Web and Cloud Computing - Introduction

UBCO Master of Data Science – DATA 534

The Essence of the Course

The overall goal of this course is for you to:

How to use the web as a platform for data collection, computation, and publishing. Accessing data via scraping and APIs. Using the cloud for tasks that are beyond the capability of your local computing resources

Admin stuff…

1) Class hours

T-Th: 11:00am – 12:30pm

2) Office hours: mostafa.mohamed@ubc.ca

By appointment SCI 200E

3) Class collaboration (in-class/out-of-class Class collaboration (in-class/out-of-class)

I encourage collaboration

3) Communication: Slack, email

Grading

Assessment	Weight	Deadline (where to submit)	Type
Labs (4)	50%	Sunday at Noon (Canvas)	Individual
Project Proposal	10%	Third Week (Canvas)	Group
Group Project	40%	Last week, In-class presentation	Group

DATE	DAY	TOPIC	FORMAT
09-Jan-2024	Tues	Introduction	In-Class
11-Jan-2024	Thurs	Internet: protocol, HTML, CSS, Web scrapping	In-Class
16-Jan-2024	Tues	APIs, JSON, and how the Internet works	In-Class
18-Jan-2024	Thurs	Cloud computing platform in practice	In-Class
23-Jan-2024	Tues	Parallel Computing, Map Reduce	In-Class
25-Jan-2024	Thurs	Project's Time	Online
30-Jan-2024	Tues	Apache Spark, NoSQL	In-Class
01-Feb-2024	Thurs	Project's Time	Online
06-Feb-2024	Tues	Final Project Presentation	In-Class
08-Feb-2024	Thurs	Final Project Presentation	In-Class

Labs

#	Topic
1	Web Scraping
2	API
3	AWS
4	Map Reduce

Project

MILESTONES:

MILESTONE	INFO	%
Project Proposal	Each team will submit one page document with details about their project.	10
Project Presentation	Each team will have 10 minutes for the presentation and 2 min for Q&A.	10
Final Project	(1) Installable package in Windows, Mac, Linux, (2) Source code tested and documented, (3) Complete documentation and Vignettes (4) Individual Code Diary	30

Academic Dishonesty

Cheating is strictly prohibited and is taken very seriously by UBC.

A guideline to what constitutes cheating:

- Labs
  - Submitting code produced by others.
  - Working in groups to solve questions and/or comparing answers to questions once they have been solved (except for group assignments).
  - Discussing HOW to solve a particular question instead of WHAT the question involves.

Academic dishonesty may result in a "F" for the course and removal from the MDS program.

How to Excel in This Course

Attend ***every*** class:

- Participate in class exercises and questions.

Attend and complete all labs:

- Labs practice the fundamental employable skills as well as being for marks.

Practice on your own. Practice makes perfect.

- There are a lot of documentation to read in order to set up everything needed for the labs. Before asking the TA, make sure you have read the documentation.
- Read the additional reference material and perform practice questions.

Systems and Tools

Course material is on Canvas.

Marks are distributed on Canvas. Demos will be uploaded on Canvas

Your laptop will be used to install all software and run programs.

The Project

Do you want to form
the groups yourselves?

Project Info

Objectives

The course objectives are to train students in:

1. Scraping data from websites
2. Accessing data using APIs where available
3. Understanding the different cloud computing architectures
4. Deploying tasks in the various cloud architectures

The Web as an *Unintended* Data Source

What is the Web ?

What does the web have to do with data science ?

Why do we want to learn Cloud Computing ?

- Emerging technology
- New programming model
- Is the future of computing
- Running application on very large data-sets

What computing solutions are currently available ?

Personal Computing

- Personal computing system
- Local software installation, maintenance
- Local system maintenance
- Customizable to user needs
- Very low utilization
- High up-front cost

Reconfigurable Computing

- **Field Programmable Gate Arrays (FPGAs)**
  - Reprogrammable Hardware
  - Can exploit embarrassingly parallel code
  - Slow programming time (ms)
  - Power hungry

Mobile Computing

- You can use computing technology on the move
- Since 1990s
- Intermittent connectivity
- Limited Bandwidth
- Mobile device maturity

Utility Computing

Water, gas, and electricity are provided to every home and business as commodity services

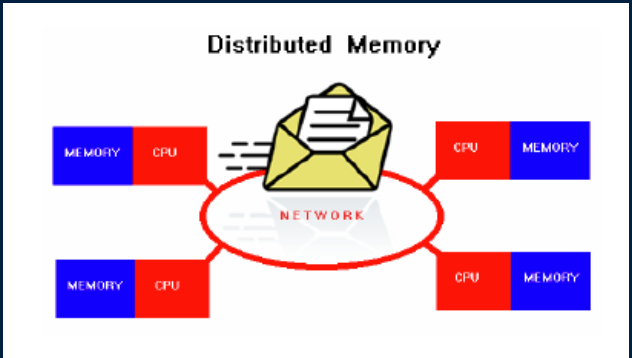
- You get connected to the utility companies' "public" infrastructure
- You get these utility services on-demand
- And you pay-as-you use

Utility Computing is doing same for computing resources (processing power, bandwidth, data storage, and enterprise software services)

Distributed Computing

Distributed Computing

- Using **distributed system**s to solve large problems.
- **Distributed System:** multiple autonomous computers connected through a communication network
- The system has a **distributed memory** where each processor has its private memory.
- Information exchanged using communication models, ex: **MPI**



Distributed Computing Contd…

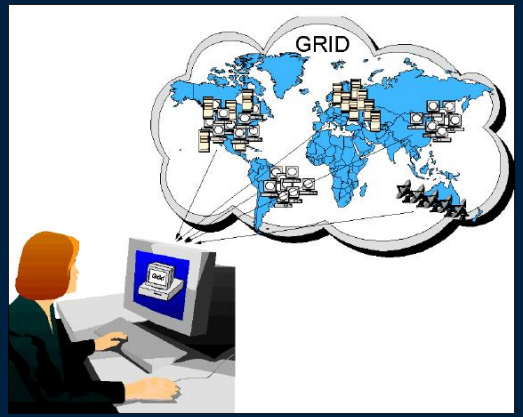
Cluster Computing:

- Characteristics:
  - tightly coupled computers
  - single system image
  - Centralized Job management & scheduling system
- Better performance and availability and more cost- effectiveness over single computer with same capabilities
- Since 1987

Distributed Computing Contd…

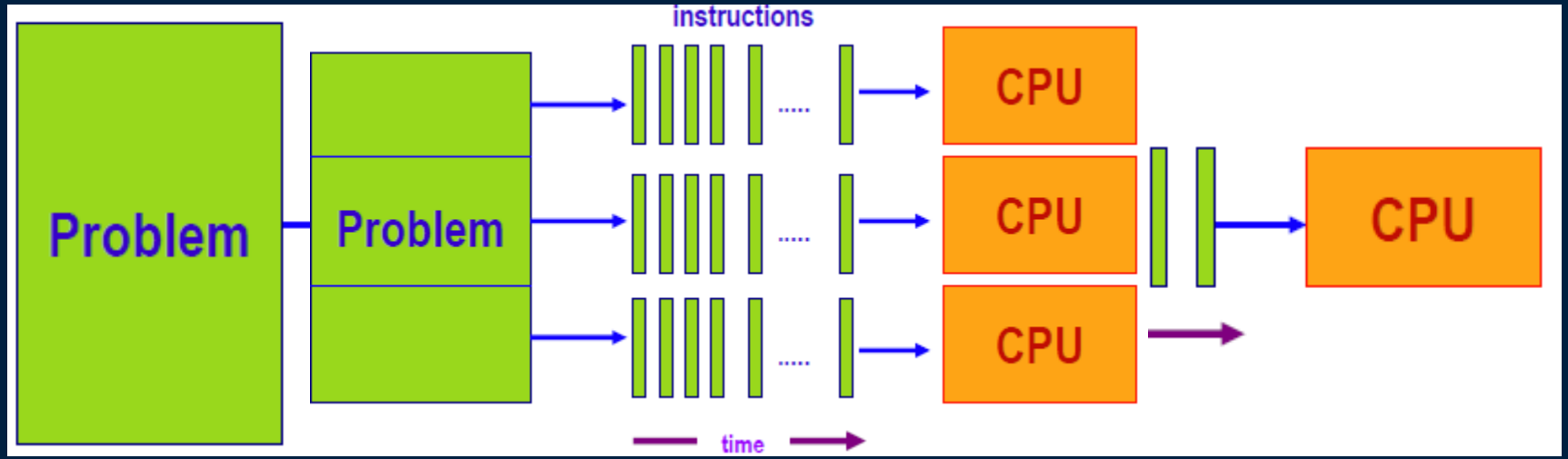
Grid Computing:

- According to Gartner, "a grid is a collection of resources owned by multiple organizations that is coordinated to allow them to solve a common problem."
  - Characteristics: loosely coupled
  - no Single System Image
  - distributed Job Management & scheduling
- Originated early 1990s



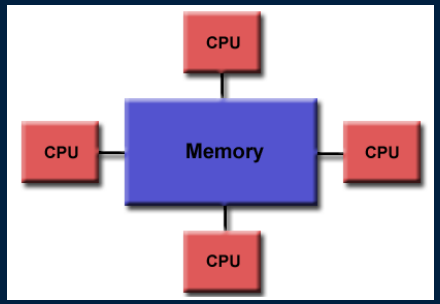
Parallel Computing

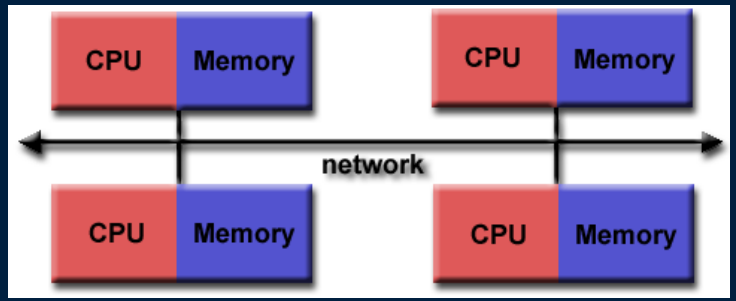
Calculations of large problems are divided into smaller parts and carried out simultaneously/concurrently on different processors.

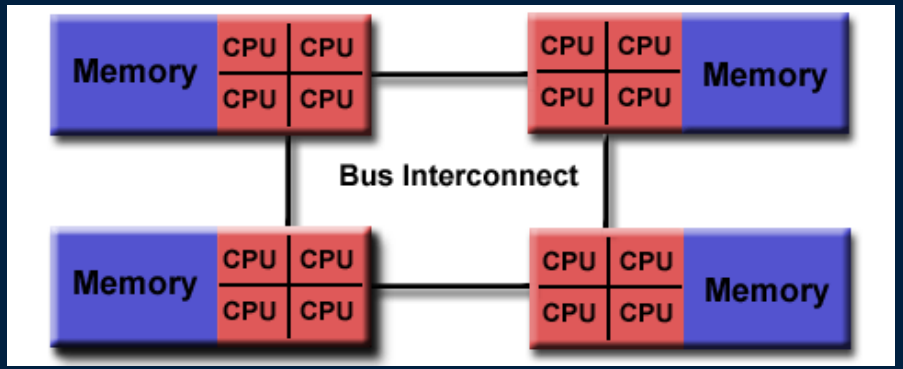


Parallel Computing

All have access to a **shared memory** that is used to exchange information between processors

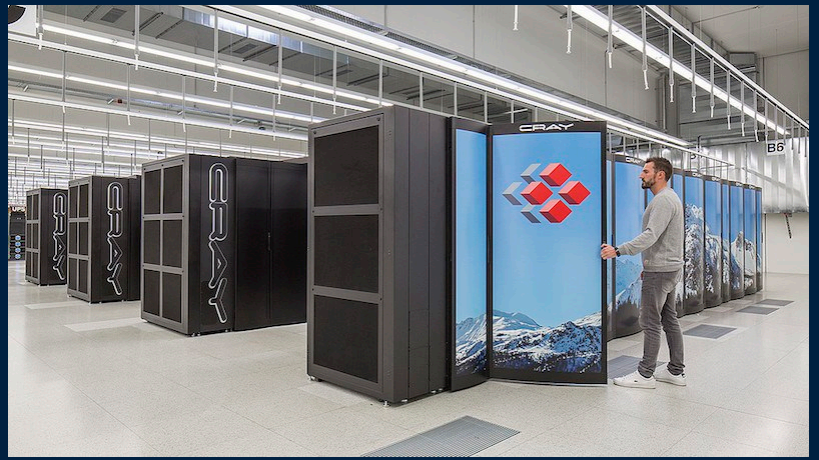






Super Computing

- Thousands of processors
- Used for compute-intensive problems
- Days instead of Years!!!
- introduced in the 1960s



Ubiquitous and Pervasive Computing

Ubiquitous= "seeming to be in all places"

Pervasive= "present or noticeable in every part of a thing or place"

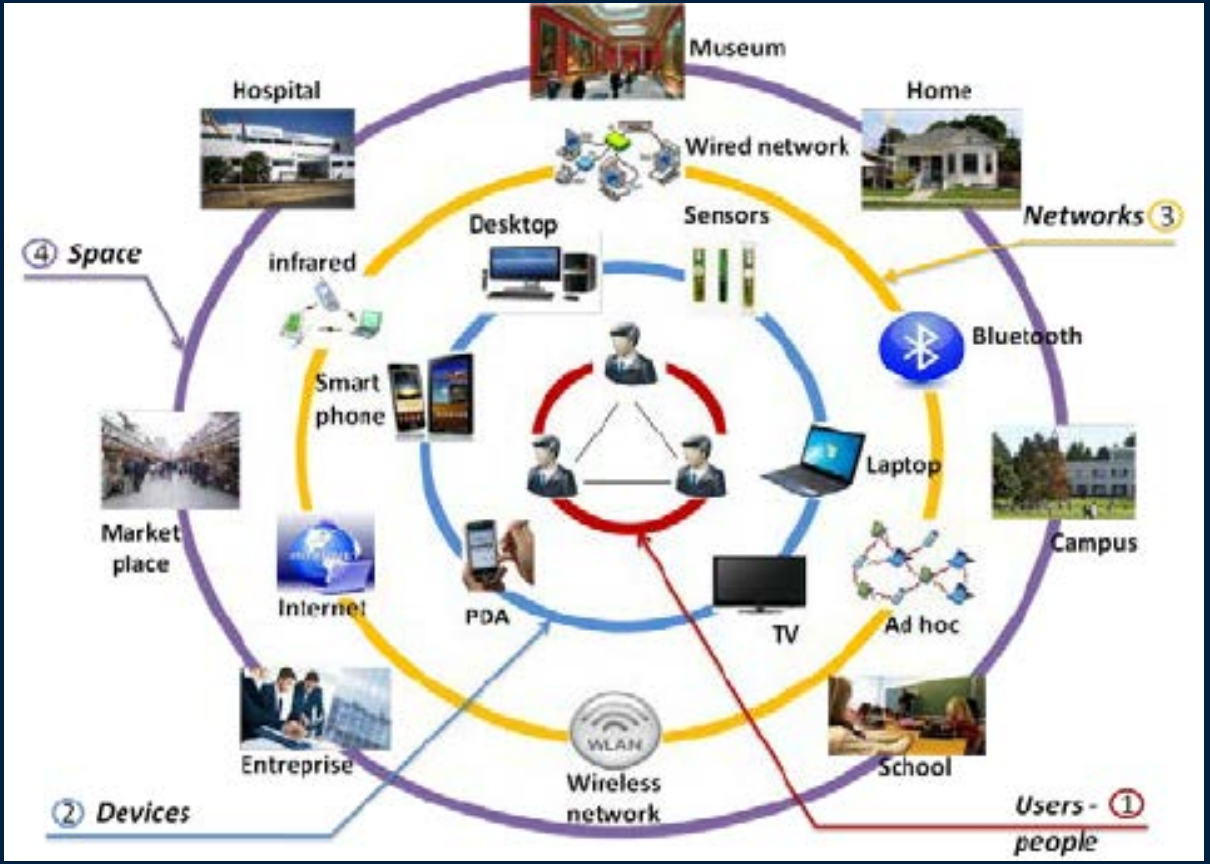
Information processing engaged in everyday's activities and objects.

Term used since 1980s

Different models but same vision:

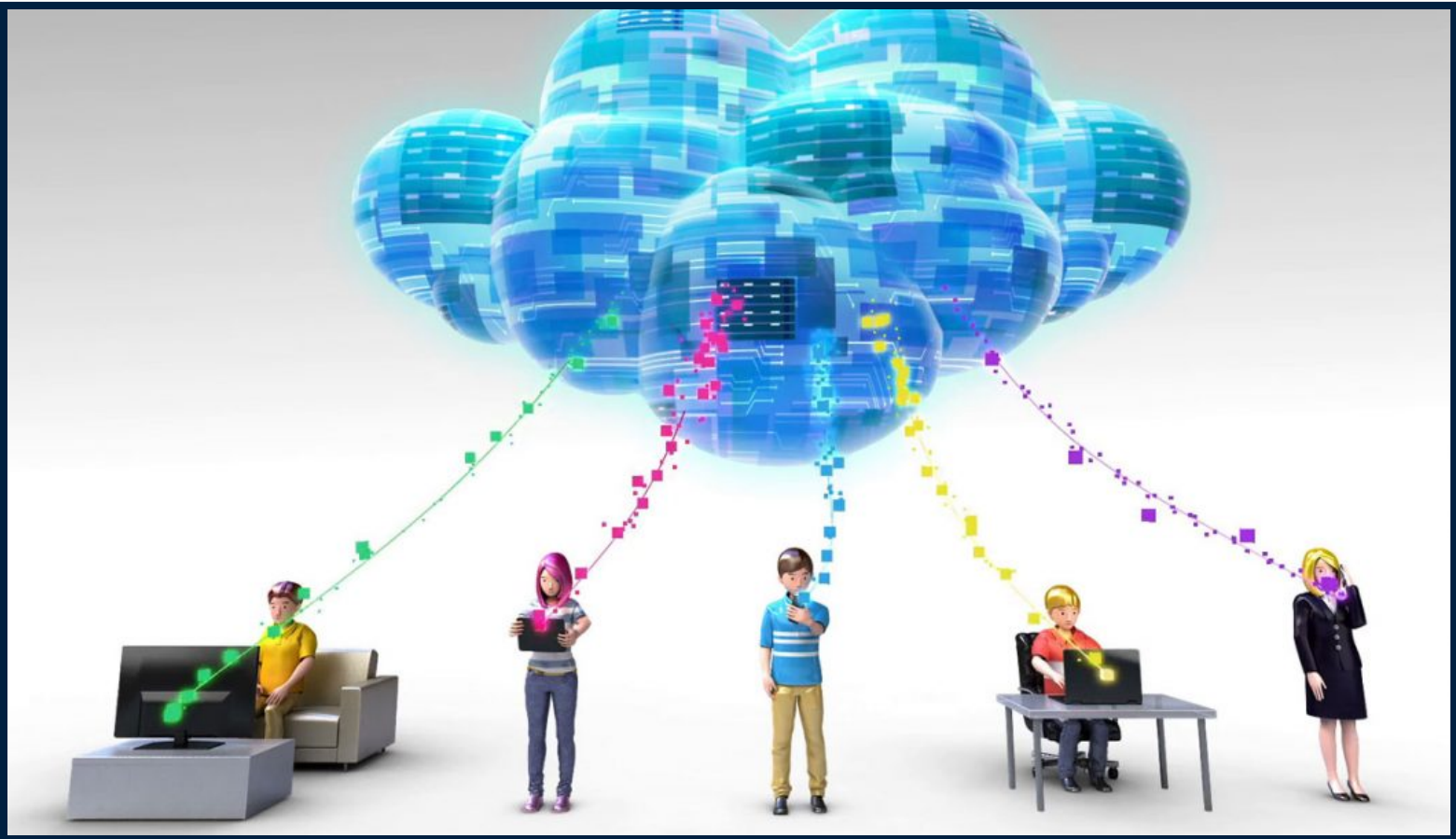
- Small, inexpensive, robust devices distributed throughout everyday's life

Ubiquitous and Pervasive Computing




Personal Computing
Reconfigurable Computing
Parallel Computing
Distributed Computing
Ubiquitous Computing
Autonomic Computing
Super Computing
Grid Computing
Cluster Computing
Utility Computing
Cloud Computing
Pervasive Computing
Mobile Computing

Cloud Computing



Cloud Computing

Think of it as Internet Computing

- Computation done over the internet

Enabled through:

- High Bandwidth and High Speed Internet
- Utility Computing
- Virtualization
- …

Cloud Computing Services

Three basic services:

- **Software as a Service (SAAS) model**
  - Apps through browser
- **Platform as a Service (PAAS) model**
  - Delivery of a computing platform for custom software development as a service
- **Infrastructure as a Service (IAAS) model**
  - Deliver of computer infrastructure as a service
- **XAAS, (the list continues to grow)**

Some useful videos…

SaaS:

https://www.youtube.com/watch?v=kGUPSvswmY0&feature=related

Virtualization:

https://www.youtube.com/watch?v=p11lJOnALS4&feature=related

Cloud Computing:

https://www.youtube.com/watch?v=M988_fsOSWo

Summary

We looked at various computing frameworks

We learnt about application domains for cloud computing

Next Lecture

- HTML, XML, CSS
- Web Scraping
- Some Demos


THE UNIVERSITY OF BRITISH COLUMBIA