DSECL ZG517 - Systems for Data Analytics

Session #1 – Course Introduction

Murali P

muralip@wilp.bits-pilani.ac.in

[Saturday – 04:30 PM]

BITS Pilani

Pilani Campus

Agenda

- About the course and operation
- Motivation
- Systems Attributes for Data Analytics - Single System
  - Kinds of storage
  - Levels of storage
  - Processing

About

Learn about fundamentals of data engineering; Basics of systems and techniques for data processing - comprising of relevant database, cloud computing and distributed computing concepts.

Objectives of the course

- Introduce students to a systems perspective of data analytics: to leverage systems effectively, understand, measure, and improve performance while performing data analytics tasks
- Enable students to develop a working knowledge of how to use parallel and distributed systems for data analytics
- Enable students to apply best practices in storing and retrieving data for analytics
- Enable students to leverage commodity infrastructure (such as scale-out clusters, distributed data- stores, and the cloud) for data analytics.

Administrivia

What to expect in this course?

- Analysing scenarios based on Hardware concepts
- Designing solutions for a particular scenario
- Analysing the optimality of the hardware solution proposed
- Analysing the cost effectiveness of the solution

What We'll Cover in this Course

- **Introduction to Data Engineering**
  - Systems Attributes for Data Analytics - Single
  - System Systems Attributes for Data Analytics - Parallel and Distributed Systems
- **Systems Architecture for Data Analytics**
  - Introduction to Systems Architecture
  - Performance Attributes of Systems
- **Data Storage and Organization for Analytics**
- **Distributed Data Processing for Analytics**
  - (Re-)Designing Algorithms for Distributed Systems
  - Distributed Data Analytics

Books

Text books and Reference book(s)

T1	Kai Hwang, Geoffrey Fox, and Dongarra. Distributed Computing and Cloud Computing. Morgan Kauffman
T2	

Expect to see more references in CANVAS

Books[2]

Reference book(s)

R1	Nikolas Roman Herbst, Samuel Kounev, Ralf Reussner. Elasticity in cloud computing: What it is, and what it is not. 10th International Conference on Autonomic Computing (ICAC '13). USENIX Association.
R2	Mohammed Alhamad, Tharam Dillon, Elizabeth Chang.Conceptual SLA Framework for Cloud Computing.4th IEEE International Conference on Digital Ecosystems and Technologies. April 2010, Dubai, UAE.
R3	Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google File System. SOSP'03, October 19–22, 2003, Bolton Landing, New York, USA.
R4	Apache CouchDB. Technical Overview. http://docs.couchdb.org/en/stable/intro/overview.html
R5	Apache CouchDB. Eventual Consistency. http://docs.couchdb.org/en/stable/intro/consistency.html
R6	Seth Gilbert and Nancy A. Lynch. Perspectives on the CAP Theorem. IEEE Computer. vol. 45. Issue 2. Feb. 2012
R7	Werner vogels. Eventually Consistent. january 2009. vol. 52. no. 1 Communications of the acm.
R8	Eric Brewer.CAP Twelve Years Later: How the "Rules" Have Changed. IEEE Computer. vol. 45. Issue 2. Feb. 2012
R9	M. Burrows, The Chubby Lock Service for Loosely-Coupled Distributed Systems, in: OSDI'06: Seventh Symposium on Operating System Design and Implementation, USENIX, Seattle, WA, 2006, pp. 335–350.
R10	MATEI ZAHARIA et. al. Apache Spark: A Unified Engine for Big Data Processing .COMMUNICATIONS OF THE ACM | NOVEMBER 2016 | VOL. 59 | NO. 11.
R11	YASER MANSOURI, ADEL NADJARAN TOOSI, and RAJKUMAR BUYYA. Data Storage Management in Cloud Environments:Taxonomy, Survey, and Future Directions . ACM Computing Surveys, Vol. 50, No. 6, Article 91. December 2017

Evaluation Plan

Name	Type	Weight
Quiz-I	Online	5%
Assignment-I	Take Home	25% both A1+A2
Assignment-II	Take Home	
Mid-Semester Test	Closed Book	30%
Comprehensive Exam	Open Book	40%

Lab Plan

innovate achieve lead

Lab No.	Lab Objective
1	Programming exercises on map-reduce
2	Setting up a simple 3-tier application on the Cloud
3	Synchronization exercise on CouchDB
4	Pen-and-paper exercise on Locality, Memory Contention, and Communication Requirement
5	Pen-and-paper exercise on calculations of speedup, MTTF, and MTTR.

- **Labs not graded**
- **Lab recordings will be available**
- **Webinars will be conducted for lab sessions**

What's already done

- 2 year, 4 semester programme

Year	First Semester		U	Second Semester		U
I	DSE* ZC415	Data Mining	3	DSE* ZC413	Introduction to Statistical Methods	3
	DSE* ZC416	Mathematical Foundations for Data Science	4	DSE* ZG523	Introduction to Data Science	3
	DSE* ZG519	Data Structures and Algorithms Design	5		Elective –I	
	DSE* ZG516	Computer Organization & Software Systems	5		Elective-II	
		Total	17		Total	15 (min)
II	Elective-III			DSE*ZG628T	Dissertation	16
	Elective-IV					
	Elective-V					
	Elective-VI		16 (min)			16

Administrivia

What to NOT expect in this course?

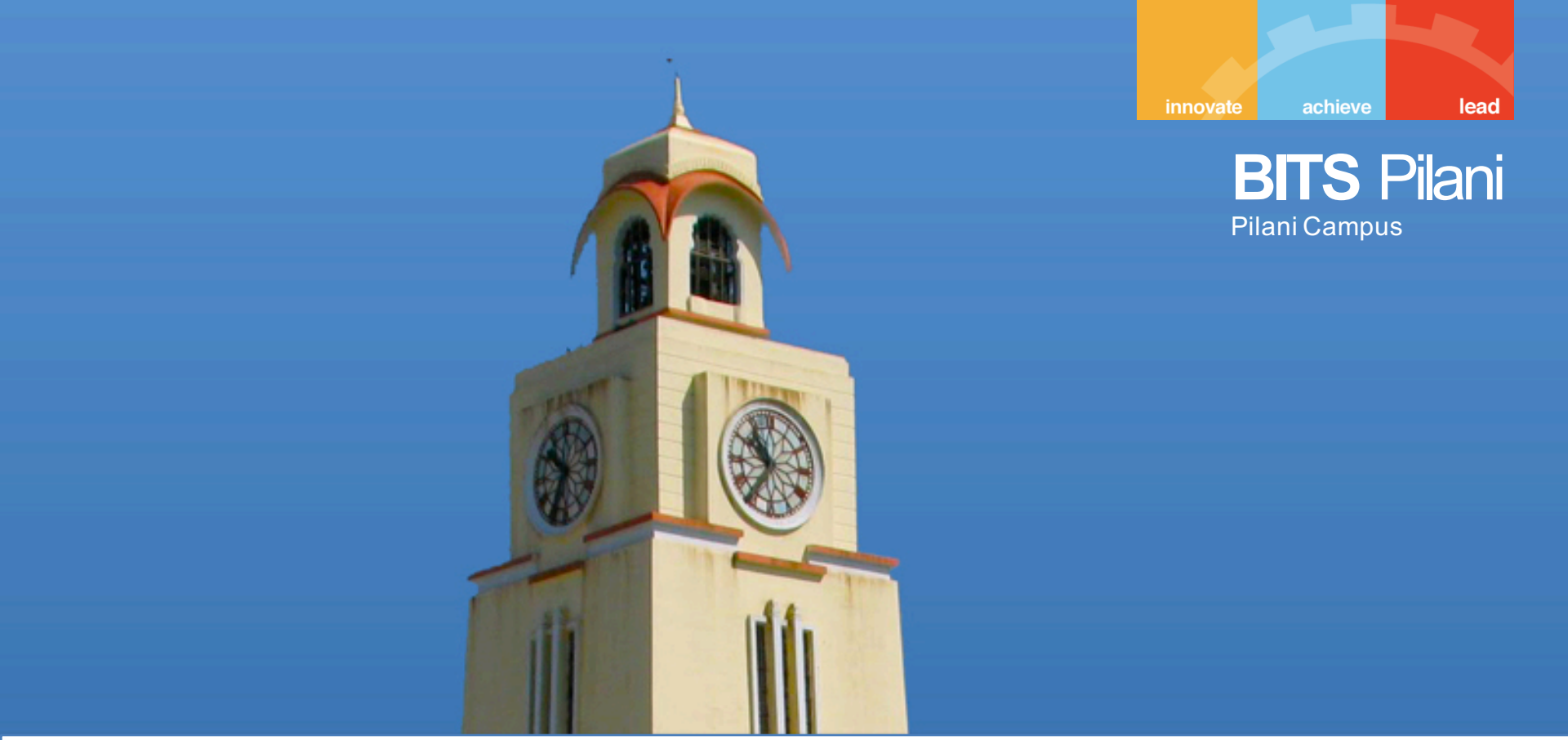
- Preliminary knowledge on Hardware concepts and fundamentals
- Preliminary knowledge on Data Engineering and Data Science

We expect you to have done well
Computer Organization and Software Systems
You should know processor, memory, storage, performance metrics and measurement.

Systems for Data Analytics

BITS Pilani
Pilani Campus

Motivation



innovate achieve lead

BITS Pilani

Pilani Campus

Systems Attributes for Data Analytics - Single System

Courtesy: Prof Sundar B slides

Storage for Data

- **Structured** Data
  - E.g. Relational Data
- **Semi-structured** Data
  - E.g. HTML pages, XML data, JSON, CSV files, Email, NoSQL DB, etc.
- **Un-structured Data**
  - E.g. X-ray images, audio/video/photo files, word processing docs, books, journals, health records, metadata, etc.

Anecdotal Evidence

- *Most of the data today is semi-structured / unstructured.*

Kinds of Data and Forms of Storage

- Historically,
  - **Relational Databases** were used for storing *structured data*
  - **File Systems** were used for *unstructured data*

Question

What are the typical characteristics of (relational) databases vs. file systems?

Storage for Data- Attributes

- Granularity
- Way of accessing
  - Random Access
  - Sequential
- Structure of DB and facilities as compared to file systems
  - Querying

There is a link between
form of data and form of storage

Kinds of Data and Forms of Storage [2]

- Later,
  - XML databases were introduced to store semi-structured data
  - Object storages were introduced to store unstructured data (with ***object type*** information)

Exercise

Find examples of these two types of storage and their typical characteristics!

Kinds of Data and Forms of Storage [3]

- Today,

Kind of Data	Form of Storage	Example (Products)
Structured (Relational)	Relational Databases	Oracle, MSSQL, and MySQL; SimpleDB (Amazon)
Semi-structured / Unstructured	File Systems or Object Storages or NOSQL databases (including XML databases)	MongoDB; S3, Elastic FS (Amazon)

Discussion/Assignment:

How we access data in relational vs semi-structured vs unstructured data?

Compare with real examples

Data Location: Memory vs. Storage

- Computational Data is stored in
  - ***Primary Memory*** (a.k.a. Memory)

vs.

- Persistent Data is stored in
  - i.e. ***Secondary Memory*** (a.k.a Storage)

Data Location: Memory vs. Storage

- Computational Data is stored in
  - ***Primary Memory*** (a.k.a. Memory)

Use and Throw

vs.

- Persistent Data is stored in
  - i.e. ***Secondary Memory*** (a.k.a Storage)

Multiple runs

Questions / Exercises

1. What does "persistent" refer to? Is it same as non-volatile?
2. Identify examples of these two kinds of data
3. Identify technologies suitable for the two kinds of data

Data Location: Memory vs. Storage vs. Network

- Data accessed from a ***local store***
  - i.e. storage attached to a computer

vs.

- Data accessed from a ***remote store*** / remote processor
  - i.e. storage hosted on the network (or *storage attached to a computer hosted on the network*)

Question

- What is the difference in the form of access?

There is a link between
form of data and form of storage

Cost of Access: *Memory vs. Storage vs. Network*

- Exercise:
  - What are the typical access times?
    1. RAM
    2. Hard Disk
    3. Ethernet LAN
  - Access Time Parameters: ***Latency*** and ***Bandwidth***
    - When and how do these parameters matter?
  - Identify mechanisms used to alleviate the access time delays in each case.

Memory Bandwidth Requirement [1]

- How does all this impact processing capability?
  - Design intellect, money => processor is the key
- Let's take a typical processor
  - Consider a 2.5 GHz processor with 4 cores:
    - each with a CPI of 1.25 and
  - a RISC ISA where
    - 1 out of 4 instructions require a memory access and
    - word size is 4 bytes.
    - pipelining
  - Calculate the memory bandwidth required!

Memory Bandwidth Requirement [2]

- Processor Clock: 2.5 GHz
- One cycle: 0.4 ns
- In one core:
- CPI = 1.25 (conservative estimate)
- Typical instruction executes hence in 1.25*0.4 = 0.5ns
- RISC ISA
  - 1 out of 4 instructions requires memory data access
- 4 cores
  - 4 instruction access + 1 data access = 5 accesses.

Memory Bandwidth Requirement [3]

- Total bandwidth to constantly feed processor:
  - 40GBps

How to give a typical 4-core processor about 40GB of data every second?

Memory Hierarchy – Motivation

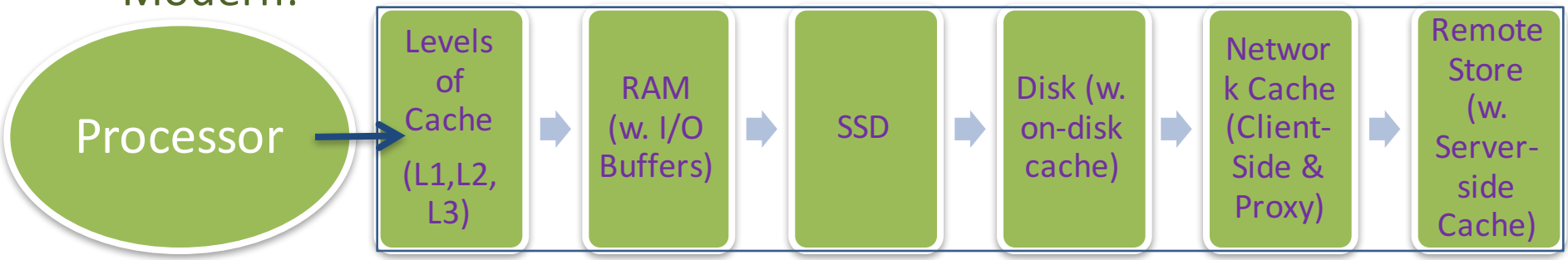
- A **Memory Hierarchy** amortizes cost in computer architecture:
  - fast (*and therefore costly*) but small-sized memory

  to
  - large-sized but slow (*and therefore cheap*) memory

Memory Hierarchy

- Original:

- Modern:



Discussion/Assignment:

How do we
Reconcile Memory Bandwidth Requirement
with the Memory Hierarchy?

Thank you !