BOSTON UNIVERSITY

CS 561: Data Systems Architectures

Prof. Manos Athanassoulis

mathan@bu.edu

https://bu-disc.github.io/CS561/

no smartphones

no laptop

Why?

there is enough evidence that laptops and phones slow you down

BOSTON UNIVERSITY

Today

big data

data-driven world

data systems

which are the driving trends?

why do we need new designs?

CS 561 goals & logistics

I want you to speak up!
[and you can always interrupt me]

BOSTON UNIVERSITY

CS 561 philosophy

cutting-edge research

question everything (to understand it better!)

There are no *stupid questions*!

interactive & collaborative

projects, presentations, labs, OH

BOSTON UNIVERSITY

Understanding a design/system/algorithm …

system

- component 1
- component 2
- component 3

why?
why not?

algorithm

- step 1
- step 2
- step 3

understanding <u>all steps</u> and <u>all decisions</u>
helps us see the ***big picture***
and do **good research**!

(otherwise, we make ad hoc choices!)

BOSTON UNIVERSITY

Ask Questions!

… and answer my questions!

our **main goal** is to have **interesting discussions** that will help to gradually understand what the material discusses

(it's ok if not everything is clear, as long as you have questions!)

BOSTON UNIVERSITY

What do we do in this class?

reading papers

presentations

technical questions

projects

BOSTON UNIVERSITY

Reading Papers

every class **1-2 papers to discuss** in detail

in some classes the discussion will be led by a group of students

so that, each student will present one paper during the semester

(background papers also available to provide more details)

read all of them!

answer technical questions as 10-min quizzes (for a subset of the papers)

BOSTON UNIVERSITY

Paper Presentation / Student Discussion

3-4 courses (later in the semester) will be entirely led by students!

A group of ~10 students will be split in three subgroups:
- **presenters**, who will develop the slides and deliver the presentation
- **critics**, who will raise appropriate doubts or question about the paper
- **proponents**, who will explain how this paper addresses these

Note that the three subgroups can coordinate and prepare together.

Each student must participate once in a presentation/discussion in (any) one of the three roles

Prepare **slides** at least **a week before your presentation + discuss them in OH**

BOSTON UNIVERSITY

Reading a Research Paper

learn

review
what is the problem & why it is important?
why is it hard & why older approaches are not enough?
what is the key idea and why it works?

critic

what is missing and how can we improve this idea?
does the paper support its claims?
possible next steps of the work presented in the paper?

additional technical questions
to make sure the heart of the paper is clearly understood

BOSTON UNIVERSITY

remember, this will help us do **good research**!

Projects

AND

project 0

A small implementation project
to sharpen dev skills

<u>independent project</u>

Due on Jan 31, 2025

project 1

A medium project to give you a flavor of
large-scale production system

<u>groups of 3</u>

Due on Feb 14, 2025

BOSTON
UNIVERSITY

Projects

AND

project 0

A small implementation project
to sharpen dev skills

<u>independent project</u>

Due on Jan 31, 2025

project 1

A medium project to give you a flavor of
large-scale production system

<u>groups of 3</u>

Due on Feb 14, 2025

Projects

systems project

<u>groups of 2/3</u>

implementation-heavy C/C++ project

OR

research project

<u>groups of 3</u>

pick a subject (list will be available)

design & analysis

experimentation

BOSTON UNIVERSITY

Projects

OR

systems project

<u>groups of 2/3</u>

implementation-heavy C/C++ project

research project

<u>groups of 3</u>

pick a subject (list will be available)

design & analysis

experimentation

Research Project: open questions

skew-aware join optimization

context-aware spatial indexes

exploit ***near-sorted data*** with concurrency control

quantify ***Write Amplification*** in modern SSDs

come up with your **own topic!**

more on the website (soon)

BOSTON UNIVERSITY

A good project

(1) has a clear plan by project proposal by **end-February** (5%)
(2) has significant preliminary work done by **end-March** (10%)

evaluation at the **end of the semester** (25%)
(i) present the key ideas of the implementation/new approach
(ii) present a set of experiments supporting your claims

come to OH!
(more details for the projects in Class 4)

BOSTON UNIVERSITY

Class Goal

understand the internals of

<u>data systems for data science</u>

tune data systems through **adaptation** and **automation**

get acquainted with research in the area

BOSTON UNIVERSITY

Can I take this class?

background

C++ programming
data structures
algorithms
comp. architecture

pre-req

CS460/660 & CS210
contact Manos if not sure

how to be sure?

if familiar with most, then maybe!
if familiar with **none**, then no!

BOSTON UNIVERSITY

Next classes

Class 1-2
logistics, big data, data systems, trends and outlook

Class 3
more basics on data systems, systems classification, graph, cloud

Class 4
intro to class project

Class 5 and beyond
present and **discuss** research papers from Papon/Zichen + students + guest lectures

BOSTON UNIVERSITY

who doesn't have a lot of data?

big data?

So what do we do with this data?

BOSTON UNIVERSITY

BOSTON UNIVERSITY

is data analysis new?

what is *really* new?

BOSTON UNIVERSITY

Every day, we create 2.5 exabytes* of data — 90% of the data in the world today has been created in the last two years alone.

[Understanding Big Data, IBM]

*exabyte = 10^9 GB

BOSTON UNIVERSITY

data management skills needed

100s of entries	**pen & paper**
10^3-10^6 of entries	**UNIX tools and excel**
10^9 of entries	**custom solutions, programming**
10^{12+} of entries	**data systems**

BOSTON UNIVERSITY

size (volume)

rate (velocity)

sources (variety)

big data

(it's not only about size)

all of the above plus …

BOSTON UNIVERSITY

our ability to collect ***machine-generated*** data

scientific experiments

sensors

social

monitoring

micro-payments

Internet-of-Things

cloud

BOSTON UNIVERSITY

BOSTON UNIVERSITY

BOSTON UNIVERSITY

what is a **data system**?

BOSTON UNIVERSITY

a **data system** is a large software system
(a collection of algorithms and data structures)
that **stores data**, and provides the **interface** to
update and **access** them **efficiently**

the end goal is to make **data analysis** easy

BOSTON
UNIVERSITY

“relational databases are the foundation of western civilization”

Bruce Lindsay, IBM Research

ACM SIGMOD Edgar F. Codd Innovations award 2012

+ growing need for tailored systems

BOSTON UNIVERSITY

Why?

more data

new hardware

new applications

new performance goals

BOSTON UNIVERSITY

The big success of 6 decades of research

a declarative interface!

"ask and thou shall receive"

ask ***what*** you want

data system

system decides ***how***
to store & access

is this good?

BOSTON UNIVERSITY

"three things are important in the database world: ***performance****,* ***performance****, and* ***performance****"*

Bruce Lindsay, IBM Research

ACM SIGMOD Edgar F. Codd Innovations award 2012

CS561: data systems **kernel** under the looking glass

this is where we will spend our time!

system architecture (row/column/hybrid)

indexing

relational/graph/key-value

scale-up/scale-out

goal: learn to design and implement a DB kernel

BOSTON UNIVERSITY

how to design a data system kernel?

what are its basic components?

algorithms/data structures/caching policies

what decisions should we make?

how to combine? how to optimize for hardware?

designing a DB kernel is ***complex***

BOSTON UNIVERSITY

data system design **complexity**

application

performance

budget

thousands of **options**
millions of **decisions**
billions of **combinations**

BOSTON UNIVERSITY

let’s think together: a simple DB kernel

a key-value system, each entry is a {key,value} pair

main operations: *put, get, scan, range scan, count*

workload has both reads (get, scan, range scan) *and writes (put)*

data

how to store and how to access data?

how to efficiently delete?

BOSTON UNIVERSITY

designing a simple key-value system

what is the key/value?

are they stored together?

can read/write ratio change over time?

what to use? b-tree, hash-table, scans, skip-lists, zonemaps?

how to handle concurrent queries? million concurrent queries?

what happens if data does not fit in memory?

how to compress data?

what about privacy and security?

how to offer robustness guarantees?

BOSTON UNIVERSITY

what happens when we move to the cloud?

hardware at massive scale

performance tradeoffs different

10GB app: 1% less memory in your machine so what?

10GB app: 1% less memory in 1M instances

1M*10GB*1%=100TB!

~800k$ in today's price

BOSTON UNIVERSITY

DB systems history line

BOSTON UNIVERSITY

class key goal

understand **system design tradeoffs**

design and **prototype** a system

with other **side-effects:**

sharpening your systems skills

(C/C++, profiling, debugging, linux tools)

data system designer & researcher

any business, any startup, any scientific domain

CS 561: more logistics

topics

storage layouts, HTAP systems, adaptive indexing, solid-state storage, data integration, data skipping, data systems and ML, learned index

past but still relevant topics

relational systems, row-stores, query optimization, concurrency control, SQL

no textbook – only research papers

BOSTON UNIVERSITY

grading

class participation: 5%

quizzes: 10%

project 0: 5%

project 1: 10%

paper presentation/student discussion: 15%

project proposal: 5%

mid-semester project report: 10%

project: 25%

exam: 15%

BOSTON UNIVERSITY

Survival Guide

class website: https://bu-disc.github.io/CS561/

Project 0 [5%]

- Individual
- Due on **Jan 31**

Project 1 [15%]

- 3 persons per group
- Due on **Feb 14**

Paper Presentation/Student Discussion [15%]

- Groups of ~10 in subgroups: presenters/critics/proponents
- Signup soon here: http://tinyurl.com/S24-CS561-presentations

Quizzes on Technical Questions [10%]

- 12 individual in-class Quizzes
- Lowest two scores will be discarded

Class Participation [5%]

Class Project [40%]

- 2-3 persons per group
- Project proposal (**5%**), due on **Feb 23**
- Mid-way Report (**10%**), due on **Mar 22**
- Final Report + Presentation + Contribution (**25%**), due on **Apr 26**

BOSTON UNIVERSITY

Piazza

all discussions & announcements

https://piazza.com/bu/spring2025/cs561

also available on class website

BOSTON UNIVERSITY

How can I prepare?

1) Read background research material

- **Architecture of a Database System**.
  By J. Hellerstein, M. Stonebraker and J. Hamilton.
  Foundations and Trends in Databases, 2007
- **The Design and Implementation of Modern Column-store Database Systems**.
  By D. Abadi, P. Boncz, S. Harizopoulos, S. Idreos, S. Madden.
  Foundations and Trends in Databases, 2013
- **Data Structures for Data-Intensive Applications: Tradeoffs and Design Guidelines.**
  By M. Athanassoulis, S. Idreos, D. Shasha.
  Foundations and Trends in Databases, 2024

2) Start going over the papers

BOSTON UNIVERSITY

class summary

2 classes + 2 OH + 1 Lab (5 days) per week

Quizzes with Technical Questions
Paper Presentation/Student Discussion

project 0 + project 1 + ***systems or research project***

||

proposal + mid-semester report + final report + project presentation

BOSTON UNIVERSITY

what to do now?

A) **read the syllabus and the website**
B) **register to Piazza + Gradescope**
C) **start working on project 0**
D) register for the presentation (week 2)
E) start submitting paper reviews/answering tech. questions (week 3)
F) go over the class project (end of next week will be available)
G) start working on the proposal (week 3)

BOSTON UNIVERSITY

Resources

class website: https://bu-disc.github.io/CS561/

piazza website: https://piazza.com/bu/spring2025/cs561

presentation registration: http://tinyurl.com/S25-CS561-presentations

gradescope: https://www.gradescope.com/courses/918719 (**code in Piazza**)

office hours: https://piazza.com/bu/spring2025/cs561/staff

material: papers available from the BU network

BOSTON UNIVERSITY

Welcome to CS 561: **Data Systems Architectures!**

Prof. Manos Athanassoulis

mathan@bu.edu

next time: more detailed logistics and start with data systems design