CS265 SYLLABUS - SPRING 2016

PROFESSOR
Stratos Idreos
http://stratos.seas.harvard.edu - stratos@seas.harvard.edu - Office: MD139

TEACHING FELLOWS
Manos Athanassoulis [manos@seas.harvard.edu]
Lukas Maas [maas@seas.harvard.edu]
TF’s Office: MD136

CLASS WEBSITE
http://daslab.seas.harvard.edu/classes/cs265

WHAT IS THIS CLASS ABOUT?
Big data is everywhere. A fundamental goal across numerous modern businesses and sciences is to be able to exploit as many machines as possible, to consume as much information as possible and as fast as possible. The big challenge is "how to turn data into useful knowledge". This is far from a simple task and a moving target as both the underlying hardware and our ability to collect data evolve. In this class, we will discuss how to design data systems and algorithms that can "scale up" and "scale out". Scale up refers to the ability to use a single machine to all its potential, i.e., to exploit properly the memory hierarchy and the multiple CPU and GPU cores. Scale out refers to the ability to use more than one machines (typically 100s or 1000s) effectively. We will use examples from several areas, including relational systems and distributed databases, graph processing systems (i.e., for social networks), key-value stores, noSQL and newSQL systems as well as mobile computing and interactive analytics. In a fast moving industry and research environment such skills are in high demand.

WHAT IS A DATA SYSTEM?
Data systems are literally everywhere. We are using them directly or indirectly every day all day long for numerous basic or not so basic tasks, e.g., when we are buying coffee to when we are booking airplane tickets. They provide the backbone of all modern businesses to manage their data and of course they provide the backbone of online businesses and environments such as social networks and search engines. They are also used increasingly in science as data analytics becomes more and more the fundamental barrier in generating knowledge.
WHY TAKE THIS CLASS?
As it stands, **every two days we create as much data as much we created from the dawn of humanity up to 2003** [Eric Schmidt, Google]. Sciences, businesses and everyday life are severely affected. Data systems are in the middle of all this. Data systems is how we store and access data, i.e., they are the backbone for any data-driven application. It is a $100B industry, growing 10% every year [Economist, “Data, data everywhere”].

At the same time data systems research and the whole industry are going through a major and continuous transition; given that new data-driven scenarios and applications continuously pop up, **there is a continuous need to redefine what is a good data system design** in such dynamic environments. Data system design is at the core of every major, small or emerging data driven application/industry and science lab.

This class is also a great fit for students who want to get a sense of what research is all about and are thinking about grad school in the future.

WHAT IS THE EXPECTED LEARNING OUTCOME?
The goal is to:
1) Learn state-of-the-art research and industry trends in big data systems.
2) Understand the tradeoffs in designing and implementing modern big data systems.
3) Be able to make design decisions in big data driven scenarios.
4) Develop basic research skills: reading, writing and understanding research papers.

Efficient data analytics and system design is all about how we store and access the data. In this class you are going to see how the same concepts appear again and again.

CLASS PHILOSOPHY
CS265 has unlimited office hours, unlimited late days for deliverables, relies on the latest research papers instead of a standard text book, lectures are based on interaction and discussion instead of just “lecturing and most of all it is fun!

The instructor and the TFs are here to help you all days and at all times through out the semester. You may request as many meetings as you like and as much help as you want.

The class is also geared towards engaging creative thinking and problem solving to give students a feeling of how computer science research takes place. Many of our students in the past have successfully engaged in research projects with DASlab and published research papers.
INTERACTION AND BRAINSTORMING IN EVERY CLASS
While the instructor will do a few lectures through the semester, the class is going to be primarily discussion based. Think of this as an extended brainstorming session, a round table discussion about a specific problem in each class. The goal is to create the maximum possible interaction.

Our discussion will aim at bringing up design trends and tradeoffs, as well as algorithmic issues. Another significant part of our discussions will focus on examining open problems and to highlight opportunities for innovation.

WHO CAN TAKE THIS CLASS?
If you are an undergrad and you have taken CS165 then you can make the most out of CS265. If you have not taken CS165, then good knowledge of basic data systems design and implementation techniques is required. Talk to the instructor if you have not taken CS165.

If you are PhD student and have taken a mix of systems (database and distributed systems) classes in the past, then you will be OK and we will provide enough background so you can follow. CS265 does satisfy the systems requirement towards a PhD.

WHAT CAN I DO TO PREPARE?
Especially if you have not taken CS165 it is a good idea to spend some time preparing before the semester starts and during the early weeks of the semester. The best approach is to browse some fundamental readings in data systems architectures. We propose that you take a look at the following texts.

1) Get familiar with the very basics of traditional database architectures:

2) Get familiar with very basics of modern database architectures:

3) Get familiar with the very basics of modern large scale systems:
In addition, if you have not implemented before basic data structures such as a linked list, a hash table or a tree you will find it quite useful to do so (in C). For example you may implement basic functionalities such as insert and get.

**SEMESTER & RESEARCH PROJECTS**
Each student can choose between two kinds of semester projects:

a) a systems project

b) a research project.

Systems projects are tailored to provide background on state-of-the-art systems, data structures and algorithms. They include a design component and an implementation component in C or C++, dealing with low level systems issues such as memory management, hardware conscious processing, parallel processing, managing read/write tradeoffs and scalability. This year’s systems project is about designing and implementing one of the core components of many modern key-value stores. That is, a log structured tree that can accommodate fast reads and writes. The first part of the project is about designing the basic structure of an LSM tree for reads and writes, while the second part is about designing and implementing the same functionality in a parallel way so we can support multiple concurrent reads and writes. This is a focused project that while it is not extremely heavy in terms of how much code you have to write it will bring you against basic modern system design issues and tradeoffs. Systems projects will be done individually, i.e., each student will have to work on the project on their own.

The research project, on the other hand, is much more tailored on design and proof of concept implementations trying to solve open problems. Research projects are tailored to give a taste of research to students and lead to publications. When working on a research project, students will work closely with the instructor and members of DASlab on active research projects of the lab. Students may work on groups of three. Such projects are mainly about thinking, reading and writing and much less about coding although proof of concept implementations will be our end target.

This year we will be working on the following research projects:

1) Self-designing Data Systems
2) Designing Access Methods and Balancing Tradeoffs

Research projects will be offered to students who have taken CS165 in the past and students who already have significant systems background. This will be done in consultation with the instructor.

In mid February we will hold a special class to introduce both the systems project and the research projects in detail and this will be followed by a series of OH for
clarifications. In the meantime students may browse the websites of the two research projects to get an idea of the work involved:
1) http://daslab.seas.harvard.edu/evolution/
2) http://daslab.seas.harvard.edu/rum-conjecture/
and the class website for examples of last year’s projects.

In special cases where a student wants to work on an alternative research project, i.e., a project which is inspired by existing research that the student is already doing (e.g., as part of a PhD for a grad student or a continuation of the CS165 project for an undergrad) we will work to accommodate such requests on a case by case basis. This will be done in consultation with the instructor. Assuming there is a strong plan and drive for a specific project, such requests will most likely be granted.

**What is a successful project?** For systems projects we will give out specific functionality and performance metrics you have to achieve. For research projects we will give out specific questions you need to answer.

**Evaluation:** There is no final or midterms. At the end of the semester each student will have a 1-hour session with the instructor and another 1-hour session with the TFs where students will demonstrate their projects and answer design questions about the project. [Tip: Past experience shows that frequent participation in office hours, brainstorming sessions and sections means that the instructor and the TFs are very well aware of your system and your progress which makes the final evaluation a mere formality for these cases.]

**Collaboration policy:** The systems project is an individual project: the final deliverable should be personal, you must write from scratch all the code of your system and all documentation and reports. Discussing the design and implementation problems with other students is allowed and encouraged! We will do so in the class and during office hours and brainstorming sessions.

Research projects are going to be in groups of three and similar to the systems project we encourage discussions across teams but in the end each team should deliver a project that is clearly theirs.

**Late days policy:** All projects are due at the end of the semester and this is when they will be graded. There will be a midway check-in point in late March when we will ask you to submit a design document about the design of your project (for both systems and research projects). This will take up 10% of the project grade.
HOW MUCH WORK IS IT?
You may have heard stories about CS165 and wondering if CS265 is going to be equally hard or you may have taken CS165 and wondering if this is going to be a similar amount of work. CS165 and CS265 are different style of classes. While CS165 is much more focused on implementation leading to a full system prototype, CS265 is more focused on ideas and design. In other words, you may have written 5-10K lines of code (some even more!) for CS165 but in CS265 you are more likely going to write small amounts of code and mostly play with alternative ways to design a specific functionality, structure or algorithm to highlight the effect of different choices and to find out new ways to solve a focused problem.

CLASSES
The class meets twice a week: Wednesdays and Fridays 4:00-5:30pm, room MD 221. Class starts at 4:10pm. You might be thinking that this is a weird day and timeslot. The idea is to minimize conflicts with other classes.

PRESENTATIONS AND REVIEWS
Students will be asked to present at least one research paper through the semester. The presentation will be used to trigger the discussion of the day. In addition, each student will write 5 short paper reviews through the semester.

OFFICE HOURS
We have office hours every day of the week. Starting Week 1, Prof. Stratos Idreos will hold office hours every Wednesday/Thursday/Friday 3-4pm at his office in MD139. The TFs will also hold OH twice a week: Mondays/Tuesdays 5-6pm at their office in MD136.

ATTENDANCE
Based on the philosophy of the course, attendance in both classes and OH is optional. The best way to learn, though, is through discussion and interaction with the instructor and the TFs. Our classes and lectures are not about “lecturing” – they are about interaction. We hope to see you there!

BRAINSTORMING SESSIONS
It is a tradition in CS165 and CS265 to schedule several brainstorming sessions throughout the semester. Typically we bring food and drinks and have a relaxed time discussing projects, open research topics, careers in industry and academia, grad school and anything else you may have in mind.
GRADING

- Participation in class discussions: 10%
- Paper reviews: 5%
- Presentation: 5%
- Research project 80%
- 2 Short Data Systems Projects: 80% (40% each)

Notes: Each student may do either a research project or the two short systems projects. A design document in late March will account for 10% of the project grade.

GUEST LECTURES
Every semester we are arranging 1-2 guest lectures from leaders in data system design from industry and academia. Past guest lecturers in our 2014/2015 classes include: Guy Lohman from IBM Research, Erietta Liarou from EPFL Lausanne, Alkis Simitsis and Georgia Koutrika from HP Labs and Nikita Shamgunov from MemSQL.

TEXTBOOK
The class is about state-of-the-art data systems design. There is no textbook for that. Thus, we use recent research papers and surveys which will be posted on the course website and you will have access to them through the Harvard network.

ONLINE DISCUSSIONS
We will use Piazza for online discussions.
Piazza home page: https://piazza.com/harvard/spring2016/cs265/home
Sign up: https://piazza.com/harvard/spring2016/cs265

FEEDBACK
We welcome feedback and ideas about the course at any point during the semester. Just come and chat with us during office hours!

PLAGIARISM
You are responsible for understanding Harvard and Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting "the wrong draft", or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism), where you'll find links to the Harvard Guide to Using Sources and two, free, online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.
ACCESSIBILITY
Harvard and the Extension School are committed to providing an accessible academic community. The Disability Services Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility for more information.

EXTENSION SCHOOL
Class: Lectures will be broadcasted live Wednesdays/Fridays 4-5:30pm. Lectures will also be available for on-demand broadcast within 24 hours after each class. Students will be able to watch the live or recorded broadcast through their browser using the Matterhorn player. The link to the broadcasts for CS265 will be posted before classes begin on the class website or you can find it in CS265 Canvas website virtual classroom link as of Wednesday January 27 at: https://canvas.harvard.edu/courses/8518

Capturing Discussions: Given that the majority of this class is based on interaction, extension school in cooperation with the class stuff is working to set-up a system with several microphones across the classroom so we can accurately capture all brainstorming discussions and comments during class time.

Participation in Discussions: Extension school students will also be able to participate to discussions during class time via web-conference tools: We will post the details about which conference tool we will use before classes start. The course stuff will be online with the same tools during class time and extension school students will be able to chat with the stuff directly online to pose questions: the instructor will either respond via chat or bring up the question to the whole class and respond via initiating a new discussion.

Office Hours: Extension school students will also be able to participate in office hours using the same tools as for class time. If the existing slots do not work (e.g., due to time differences), we will include additional slots for office hours that will work for those that cannot make the existing slots.

Grading: For extension school students there is no 10% “class participation” grade. This 10% goes to the project grade which accounts for 90% of the total grade. Extension school students will still be asked to provide paper reviews as well as presentations. For presentations, each extension school students will be asked to present one paper in a special session that will be arranged midway through the semester.
Piazza: To participate in piazza you need a Harvard email address. If you do not have one you can create one here: [http://www.extension.harvard.edu/resources-policies/resources/computer-e-mail-services](http://www.extension.harvard.edu/resources-policies/resources/computer-e-mail-services).

Extension school shows the class starting date to be January 28. This is when the first recorded video will be available. However, extension school students will still be able to stream the January 27 class as normal.