Large-Scale Data Processing and Optimisation (LSDPO)

Session 1: Introduction

Eiko Yoneki
Systems Research Group
University of Cambridge Computer Laboratory

My Trajectory
My Research Interests

- Spanning over Distributed Systems, Networking and Database
- Current Focus: Large-Scale Data Processing and Optimisation of Computer Systems
- MPhil project Suggestions
  [http://www.cl.cam.ac.uk/~ey204/teaching/Projects/2017_2018/](http://www.cl.cam.ac.uk/~ey204/teaching/Projects/2017_2018/)

My Group: Data-Centric Systems

Optimisation of Complex Data Processing in Computer Systems
- Auto-tuning to deal with complex parameter space using machine-learning
  - Structured Bayesian Optimisation, Reinforcement Learning
  - Build a solid auto-tuning platform in a complex and large parameter space
- e.g. Tuning Cluster task scheduling, ML framework, JVM garbage collector, NN model, Compiler, DB indexing...

Digital Epidemiology
- Real world mobility data collection in Africa/South America
- e.g. TB - sensing CO₂ and proximity of people → building complex networks
- Analyse network structure to understand infectious disease spread
  - Multiple modes of spread in time

Large-scale Graph Processing
- Fast, flexible, and programmable graph processing
- Cost effective but efficient storage
  - Move to SSDs from RAM
- Reduce latency
  - Runtime prefetching
  - Dynamic CPU/GPU scheduling
R244 Course Objectives

- Understand key concepts of scalable data processing
- Understand how to build distributed systems in data driven approach
- Understand a large and complex parameter space in computer system's optimisation and applicability of Machine Learning approach
- Research skills
  - Establish basic research domain knowledge in large data processing
  - Obtain your view of research area for thinking forward

Topic Areas

Session 1: Introduction
Session 2: Data flow programming: Map/Reduce to TensorFlow
Session 3: Large-scale graph data processing
Session 4: Hands-on Tutorial: Map/Reduce and Deep Neural Network
Session 5: Stream Data Processing + Guest lecture
Session 6: Machine Learning for Optimisation of Computer Systems
Session 7: Task scheduling, Performance, and Resource Optimisation
Session 8: Project Study Presentation
Course Structure

- Reading Club (not Lecture Class!)
  - ~4 or 5 Paper review presentations and discussion per session (~=20 minutes presentation + discussion)
  - Each of you will present about 1-2 reviews during the course
    - Revised (if necessary) presentation slides needs to be emailed on the following day

- Review_Log: minimum 1 per session
  - Email me by noon on Monday
  - Prepare questions

- Active participation to review discussion!

Review_Log

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name and (crsid):</td>
</tr>
<tr>
<td>Paper Title and Authors</td>
</tr>
</tbody>
</table>

1. Paper Summary (<100 words)
   Describe a brief summary (extract essentials).

2. Punch-line of the Paper (<250 words):
   What is the significant contribution?
   What is the difference from the existing work?

3. Any major criticism to the authors (<250 words)?
   Any criticism and suggestions to the authors?
**Course Work: Reports 1&2**

- **Review report** on full length of paper (1800 words)
  - Describe the contribution of paper in depth with criticism
  - Crystallise the significant novelty in contrast to the other related work
  - Suggestion for future work

- **Survey report** on sub-topic in data centric networking (<2000 words)
  - Pick up to 5 papers as core papers in your survey scope
  - Read them and expand your reading through related work
  - Comprehend your view and finish as your survey paper

**Study of Open Source Project**

- Open Source project normally comes with new proposal of system/networking architecture
- Understand the prototype of proposed architecture, algorithms, and systems through running an actual prototype
- Any additional work
  - Writing applications
  - Extending prototype to another platform
  - Benchmarking using online large dataset
- Present/explain how prototype runs
- Some projects are rather large and may require extensive environment and time; make sure you are able to complete this assignment
**Course Work: Reports 3**

- **Report on project study** and exploration of a prototype (<2500 words)
  - Project selection by **November 1, 2017**
  - Title and brief description (100 words) by email
  - Project presentation on **November 28, 2017**
  - Final report on the project study by **January 16, 2018** (by December 20 is preferable)

---

**Candidates of Open Source Project**


- List is not exhausted and discuss with me if you find more interesting one for you
- Expectation of workload on open source project study is about intensive 3 full days work except writing up report
- One approach: pick one in the session topic, which you are interested in along your survey report
**Important Dates**

- November 1 (Wednesday)
  - Project selection

- November 10 (Friday)
  - Review report

- November 24 (Friday)
  - Survey report

- January 16, 2018 (Tuesday) – December 20 (Wednesday) is preferable
  - Open source project study report

**Assessment**

- The final grade for the course will be provided as a letter grade or percentage and the assessment will consist of two parts:
  - 20%: for a reading club (presentation, participation, tutorial session exercise and *review_log*)
  - 80%: for the three reports
    - 20%: Intensive review report
    - 25%: Survey report
    - 35%: Project study
Welcome to R244

- Now tell about yourself
  - Your name and where you studied before ACS (or Part III)
  - What is your research interests (topics)
  - Why are you interested in R244

How to Read a Paper?
How to Read a Paper?

- Scope of LSDPO is wide
  - ...includes distributed systems, OS, networking, programming language, database...

- Type of papers
  - Building a real system
  - Proposing algorithm/logic on architecture design
  - Optimising computer systems
  - New idea

Critical Thinking

- Reading a research paper is not like reading a text book
- But the most important one is that the paper is not necessary the truth
  - there is no right and wrong, just good and bad
  - There are inherently subjective qualities...but you can’t get away with just your opinion: must argue
- Critical thinking is the skill of marrying subjective and objective judgment of a piece of work
**First Let’s Argue for...**

- What is the problem?
- What is important?

- Why isn’t it solved in previous work?
  - Why graph specific parallel processing? MapReduce is not good enough?

- What is the approach?
  - Graph specific MapReduce

- Why is this novel/innovative?
  - Iterative operation for graph parallel

**And Now against...**

- Problem is overstated (or oversold)
- Problem does not exist
- Approach is broken
  - It does not work for all the algorithms...
- Solution is insufficient
  - Only works when data is in memory...
- Evaluation is unfair/biased
  - Use HPC for experiment
So Which is RIGHT Answer?

- There isn’t one!
  - Most of arguments are mostly correct...

- Your judge on what is valuable on topic

- In this course, we’ll be reviewing a selection of ~20 papers (4-5 per week)
  - All of these papers were peer-reviewed and published
  - However you can pick your opinion on papers!

Reviewing Tips & Tricks

- Identify a core/major idea of the topic
- Read related work and/or background section and read key other papers on the topic
- Capture the author’s claim of contribution in introduction section and judge if it is delivered
- Understand the methodology that demonstrates paper’s approach
- Capture what authors evaluate and judge if that is a good way to evaluate the proposed idea
- For theory/algorithm paper, capture what it produces as a result (rather than how)
Key in Review Comments

- What do YOU think?
  - Where you finally get to explain your opinion!
  - You should aim to give a judgement on the work
  - Your judgement should be backed by your argument

- Questions for the authors

How to Review a Paper Aid...


- Simon Peyton-Jones: How to write a great paper and give a great talk about it, Microsoft Research Cambridge.

- David A. Patterson: How to Have a Bad Career in Research/Academia, 2001.

See course web page for the paper links.
Structure of Presentation

1. Background/context
   - What motivated the authors?
   - What else was going on in the research community?
   - How have things changed since?

2. What is problem to be tackled?
   - What is the problem they tried to solve?
   - What are the key ideas?
   - What did the authors actually do?
   - What were the results?

3. Your opinion of the paper
   - What you agree and what you disagree?
   - What is the strength and weakness of their approach?
   - What are the key takeaway?
   - What was the impact (possible impact)?

Preparing...

- Not too much basics: remember, others would have read the paper
  - Brief overview
  - Do not make exact repeat of the paper

- Aim: generate discussion – spit your straight opinion about the paper to stir the discussion
  - Explore the arguments they make and the conclusions they draw. What is your opinion on it?
  - When you argue, state clearly the point of argument
Presenting...

- Practice beforehand to ensure length of your presentation

- Getting nervous is normal!
  - We are in the same boat and we help each other to understand the paper
  - Presentation is a tool to provide a discussion forum

- Try not to get defensive or angry at questions
  - It is not your paper!

Listening Presentation...

- You need to get involved

- Ask questions from your review – bring your review_log copy

- Always be respectful of the speaker

S. Hand’10
**How to Write Reviews (Report 1)**

- **Paper Summary**
  - Provide a brief summary of the paper
  - At this stage you should try to be objective

- **Problem**
  - What is the problem? Why is it important? Why is previous work insufficient?

- **Solution or Approach**
  - What is their approach?
  - How does it solve the problem?
  - How is the solution unique and/or innovative?
  - What are the details?

- **Evaluation is unfair/biased**
  - How do they evaluate their solution?
  - What questions do they answer?
  - What are the strength/weakness of the system and evaluation itself?

---

**How to write Survey paper (Report 2)**

- Demonstrate a summary of recent research results in a novel way that integrates and adds understanding to work in the research area

- Must expose relevant details associated, but it is important to keep a consistent level of details and to avoid simply listing the different works

- For example:
  - Define the scope of your survey
  - Classify and organize the trend
  - Critical evaluation of approaches (pros/cons)
  - Add your analysis or explanation (e.g. table, figure)
  - Add reference and pointer to further in-depth information
Summary

- R244 course web page:
  http://www.cl.cam.ac.uk/~ey204/teaching/ACS/R244_2017_2018
  Email: eiko.yoneki@cl.cam.ac.uk

- Slides of presentation, forms, other information will be on the web