Seminar: Topics in Cloud-Based Architectures (1851450)
October 2022 (סימסטר א תשפ"ג)
Prof. Dalit Naor
dalit.naor@mta.ac.il

Reading List and Instructions

This is the reading list for the seminar. Each student should choose one paper.

Timeline for paper and presentation assignment

- List published by **25.10.2022** ; I will go over the list in class and will say a few words about each topic.
- Presentations will start on Week 7, namely 13.12.2022
- Each student should send me as soon as possible, but no later than **13.11.2022** midnight, an email with his/her preferences for:
    - at least 2 choices papers
    - 3 choices for a date
    - Possible dates:
        - 13.12.22
        - 20.12.22
        - 27.12.22
        - 3.1.23
        - 10.1.23
        - 17.1.23
- If you need to consult with me in order to make the choice - I will be available after class, or please send me an email and we will schedule time to talk on Zoom.
- Final paper assignment and date for presentation – will be published on **15.11.2022**
- Note:
    - I will do the assignments based on your preferences
    - I will do my best to accommodate everyone as much as possible (in terms of paper choices and dates) but eventually you have to accept the assignment.
- Once the timeline is set, every student should arrange 30 mins with me (on zoom) a week before the talk to go over the slides.

How to choose a paper?

Note: if you have a problem accessing the papers please let me know.

- Read the abstracts of all the papers and select those that seem relevant to you
- For the relevant ones

    - Read the introduction & background, the conclusions, the headings, look at the figures and graphs and their captions
- Try to pick a topic that interests you. If you have prior knowledge about the topic (from work for example) it can help.

(You can read "Efficient Reading of Papers in Science and Technology")

Preparing your presentation

Read the paper and make sure you get the overall idea. Then read it again more deeply, take notes, challenge what is said in the paper until you understand it fully. Only then start answering those questions and this will guide you thru the presentation.

The presentation should have the following components:

- Details about the paper (authors & affiliations, when and where it was published, general area/topic)
- Presented by…
- Setting the stage: Motivation, background, assumptions
- Problem addressed
- Previous approaches
- In a nutshell: what's the main contribution of the paper
- Detailed section:
    - The main solution: algorithm, systems, method
- Analysis
    - General analysis
    - Measurements: What were the experiments, the workloads
    - Systems characteristics
    - Pick one or two graphs and analyze them thoroughly
- Summary
- Your opinion and reflections on the paper
- How influential was the paper and the system, e.g.:
    - Citation count
    - Blogs, articles that refer to this paper or system
    - Active implementations

Make sure not to "cut and paste" slides from publicly available sources. I don't mind that you use some pictures, figures etc. but the presentation should be YOURS, and you should think about the logic, the organization, and the content, and what you really want to say.

Reading List

Next to each paper we mark the topic(s) addressed by the paper.

- [DS] Distribute Systems, consensus
- [CS] Cloud Storage, distributed storage, file systems
- [BDF] BigData processing frameworks
- [A] BigData Analytics support

1. Dynamo (Amazon) [DS, CS]

Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, et al.: "Dynamo: Amazon's Highly Available Key-Value Store," at *21st ACM Symposium on Operating Systems Principles* (SOSP), October 2007.

2. BigTable (Google) * [CS]

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. ACM Transactions on Computer Systems (TOCS), 26(2), 1-26. Also in OSDI 2006,
https://www.usenix.org/legacy/event/osdi06/tech/chang/chang.pdf

3. Local Reconstruction Codes in Azure OS (Microsoft) * [CS]

Huang, C., Simitci, H., Xu, Y., Ogus, A., Calder, B., Gopalan, P., ... & Yekhanin, S. (2012). Erasure coding in windows azure storage. In *2012 USENIX Annual Technical Conference (USENIX ATC 12)* (pp. 15-26). (best paper)
https://www.usenix.org/system/files/conference/atc12/atc12-final181_0.pdf

4. Consensus and Coordination in large scale systems – Zookeeper (Yahoo!) [DS]

Hunt, P., Konar, M., Junqueira, F. P., & Reed, B. (2010). {ZooKeeper}: Wait-free Coordination for Internet-scale Systems. In *2010 USENIX Annual Technical Conference (USENIX ATC 10)*.
https://www.usenix.org/legacy/event/atc10/tech/full_papers/Hunt.pdf

5. Cloud file systems: GFS (Google) and HDFS (Yahoo!) * [CS]

[SOSP 2003]
Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles* (pp. 29-43).
https://dl.acm.org/doi/pdf/10.1145/945445.945450?download=true

[MSST 2010]
Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)* (pp. 1-10). Ieee.
https://storageconference.us/2010/Papers/MSST/Shvachko.pdf

6. Batch processing: MapReduce (Google) [BDF, CS]

Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. OSDI 2004: 137-150

https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full_papers/dean/dean.pdf

Also, see the CACM paper 4 years later
Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107–113.
https://dl.acm.org/doi/pdf/10.1145/1327452.1327492

7. Batch processing: Spark (Berkeley, AMP Lab) [BDF]

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., ... & Stoica, I. (2012). Resilient distributed datasets: A {Fault-Tolerant} abstraction for {In-Memory} cluster computing. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)* (pp. 15-28). (best paper)

https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf

8. Resource Management (Berkeley, AMP Lab, Databricks) [DS, BDP]

Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A. D., Katz, R., ... & Stoica, I. (2011). Mesos: A Platform for {Fine-Grained} Resource Sharing in the Data Center. In *8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 11)*.

https://www.usenix.org/legacy/events/nsdi11/tech/full_papers/Hindman.pdf

9. Data Reduction for backup systems (DataDomain, Dell) * [CS]

Benjamin Zhu, Kai Li, and Hugo Patterson. 2008. Avoiding the disk bottleneck in the data domain deduplication file system. In Proceedings of the 6th USENIX Conference on File and Storage Technologies (FAST'08). USENIX Association, USA, Article 18, 1–14.

https://www.usenix.org/legacy/event/fast08/tech/full_papers/zhu/zhu.pdf

10. Security/Data Privacy (Microsoft) [A]

McSherry, F. D. (2009, June). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (pp. 19-30).

https://dl.acm.org/doi/pdf/10.1145/1559845.1559850

11. Graph Processing (Google, GraphLab) [BDF, A]

Pregel, SIGMOD 2010
Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., & Czajkowski, G. (2010, June). Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 135-146).

https://dl.acm.org/doi/pdf/10.1145/1807167.1807184

PowerGraph, OSDI 2010
Gonzalez, J. E., Low, Y., Gu, H., Bickson, D., & Guestrin, C. (2012). {PowerGraph}: Distributed {Graph-Parallel} Computation on Natural Graphs. In *10th USENIX symposium on operating systems design and implementation (OSDI 12)* (pp. 17-30).

https://www.usenix.org/system/files/conference/osdi12/osdi12-final-167.pdf

12. P2P Data location (Chord) [DS]

Stoica, I., Morris, R.T., Karger, D.R., Kaashoek, M.F., & Balakrishnan, H. (2001). Chord: A scalable peer-to-peer lookup service for internet applications. SIGCOMM '01.

https://pdos.csail.mit.edu/papers/chord:sigcomm01/chord_sigcomm.pdf

13. Storage Reliability (NetApp) * [CS]

Lakshmi N. Bairavasundaram, University of Wisconsin—Madison; Garth Goodson, Network Appliance, Inc.; Bianca Schroeder, University of Toronto; Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau, University of Wisconsin—Madison; (2008). An Analysis of Data Corruption in the Storage Stack. *Proceedings of the 6th USENIX Conference on File and Storage Technologies, February 2008.* ACM Trans. Storage 4, 3, Article 8 (November 2008), 28 pages.

https://www.usenix.org/legacy/events/fast08/tech/full_papers/bairavasundaram/bairavasundaram.pdf

14. Data Formats (Google) [A]

Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., & Vassilakis, T. (2010). Dremel: interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, *3*(1-2), 330-339.

https://storage.googleapis.com/pub-tools-public-publication-data/pdf/36632.pdf

(Parquet) Dremel made simple with Parquet, 2013

https://blog.twitter.com/engineering/en_us/a/2013/dremel-made-simple-with-parquet