Fall 2019 EECS E6897: Topics in Information Processing

Distributed Storage Systems for Big Data

The next generation of cloud applications, such as self driving cars, Internet of Things, and personalized social media, rely on running sophisticated hyperscale AI and analytics algorithms in real-time. To make their predictions, the algorithms need access to vast amounts of data. This data is stored on distributed storage systems that span tens of thousands of servers across data centers in multiple regions. These massive distributed storage systems underpin the performance, reliability and precision of these algorithms.

This course investigates how next generation distributed storage systems are designed and operated. The course spans the gamut from how to store bits on an individual storage device to how to replicate data across tens of thousands of servers.

In this course we will answer the following questions:

- How are distributed storage systems designed and architected?
- How do they remain available and durable despite frequent server and hardware failures?
- How do they maintain performance across multiple remote regions?
- How can they take advantage of new hardware trends?

Time and place: Tuesday 4:10 – 6:30 PM, location TBD

Instructor: Asaf Cidon

Prerequesites: Appropriate for grad students or advanced undergrads with previous classwork in computer systems. Suggested prerequisite is an operating systems class (COMS W4118 or equivalent). Students from non-systems/networking areas are welcome.

Format: Reading, writing about, and discussing research papers. Short presentations. Research projects in small groups.

Schedule

Course introduction

September 3: Course Introduction

1. Overview, logistics, goals
2. Storage technology trends
    a. HDD vs SSD: What does the future of storage hold? Part 1, Part 2.
    b. Analyzing Intel-Micron 3D XPoint: The Next Generation Non-Volatile Memory.
3. How to read papers
4. End-to-end argument

September 10: Indices and Data Structures

- LFS: The Design and Implementation of a Log-Structured File System.
- Log-structured Merge Tree: Optimizing Space Amplification in RocksDB.
- Filters: SILT: a memory-efficient, high-performance key-value store.

September 17: Distributed File Systems

1. Network file system: Scale and Performance in a Distributed File System.
2. Data center file system:
    a. Designs, Lessons and Advice from Building Large Distributed Systems (Slides 1-27).
    b. The Google File System.

September 24: Consistency and Consensus

1. **Project proposals due**
2. Weak consistency: Dynamo: Amazon's Highly Available Key-Value Store.
3. Consensus: In Search of an Understandable Consensus Algorithm (RAFT)
4. Transactions: Spanner: Google's Globally-Distributed Database.

October 1: No class: We will have two classes on the week of October 15.

October 8: Single Node Reliability

1. Disk reliability:
    a. Flash Reliability in Production: The Expected and the Unexpected.
    b. The Bleak Future of NAND Flash Memory.
2. File System reliability: IRON File Systems.

October 15: Replication

1. Replication for server failures: Copysets: Reducing the Frequency of Data Loss in Cloud Storage.
2. Replication for disk failures: Who's Afraid of Uncorrectable Bit Errors? Online Recovery of Flash Errors with Distributed Redundancy
3. Replication + strong consistency: HyperDex: A Distributed, Searchable Key-Value Store

Extra class this week:

- Mid-semester project presentations.

October 22: Coding

1. RAID: A Case for Redundant Arrays of Inexpensive Disks (RAID).
2. Distributed erasure codes: Erasure Coding in Windows Azure Storage.
3. Using codes for performance: EC-Cache: Load-balanced, low-latency cluster caching with online erasure coding.

October 29: Caching

1. Memory allocation: Cliffhanger: Scaling Performance Cliffs in Web Memory Caches.
2. Cache modeling: Cache Modeling and Optimization using Miniature Simulations.
3. Eviction policy: LHD: Improving Cache Hit Rate by Maximizing Hit Density.

November 5: University Holiday

November 12: Memory

1. Distributed memory management: Memory Resource Management in VMware ESX Server.
2. Persistent key-value: Fast Crash Recovery in RAMCloud.
3. RDMA and disaggregated memory: Efficient Memory Disaggregation with INFINISWAP.

November 19: Deduplication

- Deduplication for lower storage: Venti: a new approach to archival storage.
- Deduplication optimized for storage + network: A Low-bandwidth Network File System.
- Encryption + deduplication: DupLESS: Server-aided Encryption for Deduplicated Storage.

November 26: New storage technologies

- Replacing DRAM in KV stores: Reducing DRAM Footprint with NVM in Facebook.
- Replacing Write Ahead Logging: Write-behind Logging.
- Consistency challenges: NOVA: A Log-structured File System for Hybrid Volatile/Non-volatile Main Memories.

December 2: ML

- Storage partitioning: Bandana: Using Non-volatile Memory for Storing Deep Learning Models
- Database tuning: Automatic Database Management System Tuning Through Large-scale Machine Learning.
- Learned data structures: SageDB: A Learned Database System.

December 17: Project Presentations