Seminar on Erasure Coding in Distributed Storage Systems

236804

Our warehouse stores upwards of 300 PB of Hive data, with an incoming daily rate of about 600 TB. In the last year, the warehouse has seen a 3x growth in the amount of data stored. Given this growth trajectory, storage efficiency is and will continue to be a focus for our warehouse infrastructure. 

(Facebook, April 2014)

The traditional approach to maintain data integrity in large scale storage systems is to replicate each data item several times, so that if an item is corrupted or lost, its content can be retrieved by accessing one of its replicas. However, as storage volume in such systems continues to grow in meteoric speed, replication becomes too costly and inefficient. Erasure codes enable reconstruction of corrupted or unavailable data by algebraic manipulations on the data items that remain in the system. Their storage overhead is an order of magnitude smaller than that of replication, which makes them ideal candidates for protecting data at large scale.

Erasure coding is an important research area in information theory. Current objectives in constructing new codes and techniques are to minimize their storage overhead and reconstruction time without interfering with the system’s normal operation. The goal of this seminar is to understand the benefits, limitations, and tradeoffs of erasure codes in the context of real world, large scale, distributed storage systems. We will survey promising new coding techniques, as well as experience in employing them in large scale data centers and distributed storage systems.

The course will combine lectures by the instructors with independent reading in a seminar format. The students will read important papers in the field, will reason about the results in a critical way, and will present them in class along with their own ideas for extending the results.

The course combines techniques from the fields of coding theory, linear and modern algebra, and operating and storage systems. There are no formal pre-requisites, but the following courses provide good background:

236309 Introduction to Coding Theory
234322 Information Storage Systems

To register, email your name and ID to the instructor. Please state whether you are a graduate or undergraduate student, and other information you think is relevant, such as related courses you took.
The course will survey recent literature with the state-of-the-art works on erasure coding, for example:

- **Erasure coding theory**


- **Storage systems research**


- **Real world storage system deployment**