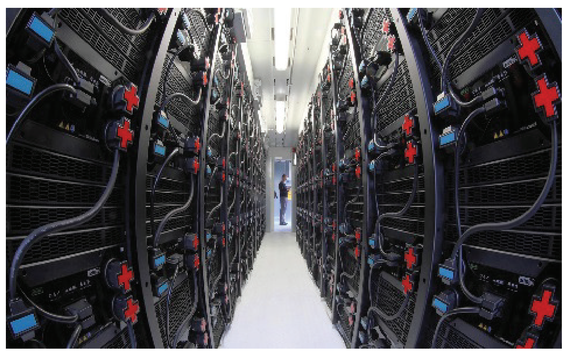
On Computing the Capacity of Finite-State Channels

Dr. G. Han, Department of Mathematics

Though it is merely the dawn of the coming big data era, we have already been swamped in the massive amount of data produced on astronomical scales. Before embarking on the mission of extracting useful information from the data, the first and foremost problem is how one can store and retrieve the data in an efficient and reliable manner. Recently distributive storage systems for big data have been emerging as a cost-effective solution, utilizing the existing storage systems in a distributive fashion.

Distributive storage systems have been seen in a great variety of real-life applications ranging from portable consumer electronics to personal or even enterprise computing. The insatiable demand of greater affordability from consumers has been driving the industry and academia, who have been relentlessly making use of aggressive technology in the bit-cost reduction process. On the other hand though, as their costs continually reduce, distributed storage systems have been more vulnerable to various device or system level noises, and maintaining the overall system reliability and performance has become a major concern.

To combat this increasingly imminent issue, various fault-tolerance techniques have been employed, however, at the expense of reduced memory storage efficiency. As the level of sophistication of the systems drastically escalates, it is of central importance to know their theoretical limit in achieving the maximal storage efficiency.



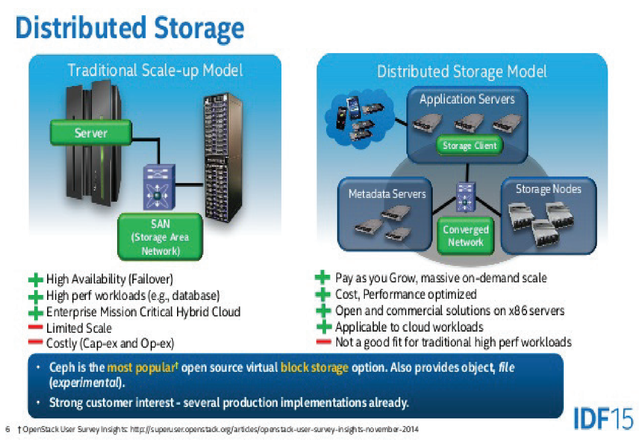
A cluster of a distributed storage system

Over the past several years, the research group led by Dr. Han has been devoted to conducting a thorough study on the fundamental limits of input-constrained finite-state channels, one class of communication channel models for distributed storage system for massive data. The tools that Dr. Han employed in his research are mainly coding and information theory.

More specifically, in our study, we have modeled a distributed storage system as a communication model where the behaviors of the channel may be affected by the channel inputs and outputs and certain constraints must be met for data storage efficiency and reliability. However, compared to a traditional storage system with a single channel state, there are a number of obstacles posing great challenges for us: non-stationarity of the model, the presence of memory and interference, node failures in the system and costs incurred by repair bandwidth.

The level of sophistication of the system merits a serious study in the long run. At this stage, Dr. Han's research team has been focusing on a prototypical model with simplified assumptions. In particular, we have established certain mathematical properties of finite-state channels under rather mild assumptions, based on which, we have recently proposed a recursive algorithm to compute the fundamental limit of certain simplified system, which proves to be convergent.

In the near future, we would like to further integrate the results we have obtained in the past few years, and overcome a number of issues to examine the distribute storage systems from an information-theoretic perspective. For instance, we will consider the effects of the channel inputs and outputs under the assumptions that the noise is both multiplicatively and additively Gaussian, which is of great importance to a class of practical flash memory channels. For this, my student, Yonglong LI and I have obtained some preliminary results. Our investigation in this direction reveals that for a full-fledge distributed storage model many established techniques seem to fall short of achieving a thorough analysis and novel ideas and new mathematical tools need to be developed.



It is of fundamental importance of understand the limit of distributed storage systems for massive data. We have made certain progress in analyzing such systems with simplified assumptions. Our ultimate goal is to give an accurate characterization of the omnipresent distributed storage system. We are confident that the framework we adopting, and the tools we have employed for this purpose are effective and will eventually bring us to fruition.

Further Readings:

1. G. Han. A randomized algorithm for the capacity of finite-state channels. IEEE Transactions on Information Theory, vol. 61, no. 7, pp. 3651-3669, 2015.
2. Y. Li, A. Kavcic and G. Han. Capacity of Multilevel NAND Flash Memory Channels. To be submitted to IEEE Transactions on Information Theory.
3. Y. Li and G. Han. Asymptotics of Input-Constrained Erasure Channel Capacity. To be submitted to IEEE Transactions on Information Theory.