Fault Tolerance and Attack Resilience on Big Data Storage

Yunghsiang S. Han
Department of Electrical Engineering
National Taiwan University of Science and Technology

Abstract Recently, organizations need to manage, process, and store huge amounts of data. Since these data are large and complex, they are very difficult to process, manage, and store by traditional database tools and data processing applications. Large data centers with storage nodes (disks) have been built to store "big data." One critical requirement of a data center is to assure data integrity. Due to the use of commodity software and hardware, crash-stop and Byzantine failures (or attacks) are likely to be more prevalent in today's large-scale data centers or distributed storage systems. Regenerating codes have been shown to be a more efficient way to disperse information across multiple storage nodes and recover crash-stop failures in the literature. Recently, regenerating codes and their variants have been proposed to be employed in data centers to prevent big data storage from crashing. It has been shown that regenerating codes can be designed to minimize the per-node storage (called MSR) or minimize the communication overhead for regeneration (called MBR). In this talk, we first review regenerating codes and their security issues. Then we propose a new encoding scheme for $[n, d]$ error-correcting MSR codes that generalizes our earlier work on error-correcting regenerating codes. Error-correcting regenerating codes are not only capable of resisting crash-stop failures but also Byzantine attacks. We show that by choosing a suitable diagonal matrix, any generator matrix of the $[n, \alpha]$ Reed-Solomon (RS) code can be integrated into the encoding matrix. Hence, MSR codes with the least update complexity can be found. An efficient decoding scheme is also proposed that utilizes the $[n, \alpha]$ RS code to perform data reconstruction. The proposed decoding scheme has better error correction capability and incurs the least number of node accesses when errors are present.

Biography Yunghsiang S. Han received B.Sc. and M.Sc. degrees in electrical engineering from the National Tsing Hua University, Taiwan, in 1984 and 1986, respectively, and a Ph.D. degree from the School of Computer and Information Science, Syracuse University, NY, in 1993. He was with Hua Fan College of Humanities and Technology, National Chi Nan Univer-

sity, and National Taipei University, Taiwan. From August 2010, he is with the Department of Electrical Engineering at National Taiwan University of Science and Technology.

Dr. Han's research interests are in error-control coding, wireless networks, and security. Dr. Han has conducting state-of-the-art research in the area of decoding error-correcting codes for more than sixteen years. He first developed a sequential-type algorithm based on Algorithm A* from artificial intelligence. At the time, this algorithm drew a lot of attention since it was the most efficient maximum-likelihood decoding algorithm for binary linear block codes. Dr. Han has also successfully applied coding theory in the area of wireless sensor networks. He has published several highly cited works on wireless sensor networks such as random key pre-distribution schemes. He also serves as the editors of several international journals.

Dr. Han was the winner of the Syracuse University Doctoral Prize in 1994 and a Fellow of IEEE. One of his papers won the prestigious 2013 ACM CCS Test-of-Time Award in cybersecurity.