A Case for End-to-End Deduplication

Submitted by Ravi Prakash Giri (rgiri8)

The paper starts by talking about the Deduplication and its presence in almost every backup environment. The author has categorized the uses of deduplication in three different categories: Storage, Data Distribution and in Network Communication. He further described the overlap among theses categories and the hybrid deduplication approach that he has proposed. He presented three different end-to-end deduplication use cases followed by their performance analysis and discussion on confidentiality and integrity.

The paper has successfully categorized the deduplication into three different categories. It provides a good literature survey on the topic. The main strength of the paper is the hybrid approach proposed by the author. In his approach, the deduplication information available to the end systems in a network communication can be extended to the network itself. So, the approach that is currently specific to storage and data distribution would be generalized to the network layer, allowing the advantages of deduplication to apply to all network protocols. Although this approach reduces both storage and bandwidth requirements, but it also increases the needs of a large index to perform efficiently, with significant memory requirements and overheads due to cache misses.

Deduplication and it's hybrid approach is definitely a panacea for reducing the storage problem especially in data centers. Although it offers various advantages in context of saving memory, it suffers from various problems from security point of view. Deduplication opens a door for the attackers to perform various cache-based side-channel attacks. Since two processes sharing a library will have a same mapped addresses. This property leads to various cache based attacks by monitoring the private and shared cache memory.