SGX and Memory Oblivious Shuffling in Apache Hadoop

PI: Yuzhe Tang
Student: Adam Piekarski

Problem: SGX (Software Guard Extensions) is a technology on recent, post-2015 Intel processors allowing for data and code to be stored in an encrypted enclave; the data is decrypted on-the-fly as needed. Apache Hadoop MapReduce is a framework for performing distributed algorithms over large sets of computers. The *map* phase partitions data into segments which are processed by each *reducer,* or computer in the cluster.

Scientific Impact: Projects involving large-scale, distributed computing—such as Big Data or artificial intelligence—rely heavily on the MapReduce pattern to handle computation. Finding ways to integrate SGX with this kind of software can result in more secure algorithms.

Solution: We explored how to leverage existing research, such as T-SGX, which makes SGX less vulnerable to side-channel attacks with the use of transactional memory, in order to enhance the security of Apache Hadoop. The map portion of the software presented the largest attack vector, as analyzing the memory accesses used in shuffling input could be used to reconstruct the data. Knowledge of SGX was employed—in particular studies on the proper placement of code vs. data in the enclave to ensure optimal performance, as seen in the figure. The integration of SGX with Hadoop is still on-going.

Impact: Integrating SGX technology with existing software is important, especially as an increasing amount of computation is done on the cloud; SGX integration may lead to more secure computing.

Educational Outreach: The potential of SGX and details about the technology are explored in a set of labs by Dr. Tang for students to use; these labs can educate and inform about SGX.