1838083 BIGDATA: IA: Enabling Large-Scale, Privacy-Preserving Genomic Computing with a Hardware-Assisted Secure Big-Data Analytics Framework

**Major goals of the project**

1. Study secure big data analytics using trusted computing technologies, trusted execution environment in particular
2. Develop computing technologies and frameworks to support secure data analytics and its applications to different domains

**Major activities**

1. We have been investigating various attestation and trust establishment techniques critical for big data analytics on the TEE platform, SGX in particular. We also developed a proof-of-concept framework that enables MPI-based data analytics on the SGX platform and analyzed it using a set of benchmark applications.
2. We have utilized our SGX-based big-data analysis techniques to support critical bioinformatics computing tasks, including collaborative computing GLMM and privacy-preserving genotype imputation.
3. We successfully organized the iDASH genome privacy competition, involving SGX-based bioinformatics data analysis track.

**Specific objectives**

For PY2 we have set these major objectives:

1. Work on the SGX-based confidential computing platform for big-data analysis.
2. Use the confidential computing techniques to support bioinformatic computing, with effective control of side-channel leaks.

**Significant results**

1. As planned in the original proposal we are converting some popular genomics analysis applications to SGX-enabled approach to enable the privacy-preserving of the data pipeline, and studying their performance implications with the extra layers to deal with the security and privacy of the data. We chose the reads mapping application as a representative case study. To minimize the effort of porting existing applications to SGX-enabled ones, we studied the SGX runtime environments such as Scone and Graphene to utilize them in our architecture. After experimenting with multiple popular reads mapping applications, including BWA-MEM, BWA-MEM2, RMAP, and Bowties2 we decided to use BWA-MEM as the application to demonstrate our approach. We used a similar approach as proposed by DIDA, to first splitting the reference genome, and then partitioning the reads data based on the split reference genome with Bloom filter. The partitioned data are then dispatched to the multiple worker nodes. The worker processes are running in a HPC cluster environment, while using docker container and Scone to enable each individual worker process to run in the SGX environment. Between each step of this workflow, we have the intermediate data files always secured, using an encryption/decryption utility function. This whole process helps achieve the end to end secured processing of the human genome reads data. We have conducted extensive experiments to tune the performance of the framework, e.g., to find the optimal number of genome splits and the partitions of the reads data, based on the characteristics of the SGX system. We have summarized this work into a paper and will be submitting it to a suitable workshop. While the current work has demonstrated the approach to convert a traditional bioinformatics application to SGX-enabled solution, we are further improving this to make it more usable. E.g., currently we use a traditional HPC cluster environment and scripts to handle the processing of each step, as well as the worker process assignment to the compute nodes in round-robin fashion. We are now looking at utilizing a Kubernetes cluster environment and a more robust framework to deal with the distributing and management of the worker processes.
2. As part of the effort to enable elastic confidential computing, we are developing technologies that enable efficient, hierarchical attestation and data management on the cloud. The idea is to allow the user to delegate her data, under encryption to the cloud, and the management of the data for various computing jobs to a persistent enclave service. Given a job description, the service (which runs inside an enclave) automatically invokes the cloud orchestration system (such as Kubernetes) to deploy the computation topology, and attestate the states of each enclaves deployed using a novel combination of remote and local attestations. This approach enables runtime establishment of trusts with computing nodes added during the elastic computation. Also new techniques are under development to ensure the repeated attestation of the services at its runtime, which cannot be supported by a verifiable software layer on the enclave (since the hardware of SGX today does not support the operation). Another issue we are trying to address is the confidentiality of the code used in the computation, as the owner of the code may not want to expose it to the data owner yet its compliance with data sharing policies still needs to be assured.
3. As a testing case for using confidential computing in bioinformatics, we developed a privacy-preserving Expectation-Maximization (EM) algorithm to build Generalized Linear Mixed Model (GLMM) collaboratively when input data are distributed to multiple participating parties and cannot be transferred to a central server. We assume that the data are *horizontally* partitioned among participating parties: i.e., each party holds a subset of records (including observational values of fixed effect variables and their corresponding outcome), and for all records, the outcome is regulated by the same set of known fixed effects and random effects. Collaborative GWAS/PheWAS on large cohorts of patients across multiple institutions is often impeded by the privacy concerns of sharing personal genomic and other health data. Our *collaborative* EM algorithm is mathematically equivalent to the original EM algorithm commonly used in GLMM construction. While a majority computation is performed by each participating institution in its private server, the joint computation can be implemented using SGX to ensure the privacy protection. A paper on this subject has been published at the 28th Conference on Intelligent Systems and Molecular Biology (ISMB’20), a leading bioinformatics conference.
4. As another application case for applying confidential computing to protect genomic data analysis, we studied how to enable confidential genotype imputation using SGX-based big-data analysis. Genotype imputation is an essential tool in genetics research, whereby missing genotypes are inferred based on a panel of reference genomes to enhance the power of downstream analyses. Recently, public imputation servers have been developed to allow researchers to leverage increasingly large-scale and diverse genetic data repositories for imputation. However, privacy concerns associated with uploading one’s genetic data to a third-party server greatly limit the utility of these services. To protect such data, we developed SMac, an efficient, side-channel-resilient imputation algorithm designed for Intel SGX, which employs the hidden Markov model (HMM)-based imputation strategy also utilized by a state-of-the-art imputation software Minimac. SMac achieves imputation accuracies virtually identical to those of Minimac and provides protection against known attacks on SGX while maintaining scalability to large datasets. We additionally show the necessity of our strategies for mitigating side channel risks by identifying vulnerabilities in existing imputation software and controlling their information exposure. Overall, our work provides a guideline for practical and secure implementation of large-scale genetic analysis, representing a step toward privacy-preserving analysis services that can facilitate data sharing and accelerate genetics research. A paper on this subject has been accepted by 25th INTERNATIONAL CONFERENCE ON RESEARCH IN COMPUTATIONAL MOLECULAR BIOLOGY (RECOMB2021), another leading bioinformatics venue.
5. In collaboration with UCSD and UTHealth in Houston, we successfully organized the 7th iDASH Genomics Privacy and Security competition. This year the competition contains three tracks, including a track on privacy-preserving clustering of single-cell transcriptomics data in SGX, in which participating teams are challenged to implement a given clustering algorithm on Intel’s SGX platform.

**Key outcomes**

1. We developed a workflow to run a genomics reads mapping application in the SGX-enabled environment. This demonstrated the approach to convert a popular bioinformatics application to an end-to-end privacy-preserving workflow. We have conducted extensive experiments to study the performance implications with various parameter settings. We plan to release the code once the system development is complete.
2. We published a paper on SGX-based distributed GLMM analysis on ISMB’20 and have another paper on SGX-based genotype imputation accepted by RECOMB’21.
3. We organized the high-profile iDASH competition, attracting over 100 teams around the world the participate.

**Training and Development**

The PIs advise a postdoc fellow, a programmer and a PhD student on the project.

**Outreach Activities**

We have successfully organized the iDASH competition, which has been participated by 115 teams from 21 countries in 4 continents. The participants include leading institutions (Yale, Sandia National Lab, Seoul National University, Tsinghua University, etc.) and industry (Samsung, Alibaba, etc.). The whole competition has spanned 6 months.

**New report period**

1. We will continue to develop the confidential elastic computing platform and submit papers on the research and release code once the development is complete.
2. We will further apply the platform to protect bioinformatics data analysis and sharing. For this purpose, we will also study the techniques to mitigate the unique threats these applications are facing.

**Impact on the development of the principal disciplines of the project**

**Impact on other disciplines**

**Development of human resources**

**Technology transfer**

**APPENDIX:**

(workshop paper manuscript here)