UVA | SCHOOL of DATA SCIENCE

2021 IEEE 14th International Conference on Cloud Computing (CLOUD)

HySec-Flow: Privacy-Preserving Genomic Computing with SGX-based Big-Data Analytics Framework

Chathura Widanage, Weijie Liu, Jiayu Li, Hongbo Chen, XiaoFeng Wang, Haixu Tang, Judy Fox
Indiana University and University of Virginia

National Institutes of Health
Turning Discovery Into Health

Privacy-Preserving Computing

- Security and privacy issues have received increasing attention in big-data analytics performed on public or commercial clouds. Personal genomic data contain identifiable information concerning human individuals.
- **Homomorphic encryption (HE)** allows users to perform computation directly on encrypted data. HE introduces several magnitudes of computational overheads.
- A promising alternative: new hardware supporting **trusted execution environment (TEE)**, in which sensitive data are kept on secure storage and processed in an isolated environment, called the enclave.
- We use **Genomics applications (BWA)** as an example and illustrate the idea that is general for any field with data privacy components.

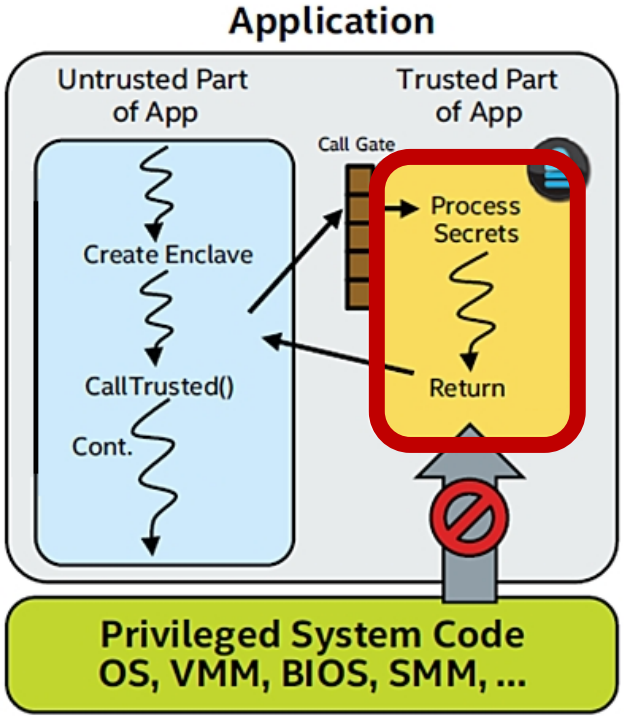
Intel®Software Guard Extensions (Intel®SGX)

Intel SGX is a set of x86 instruction extensions that offer hardware-based memory encryption and isolation for application code and data.

CPU instructions used by applications to protect critical secrets from unauthorized access:

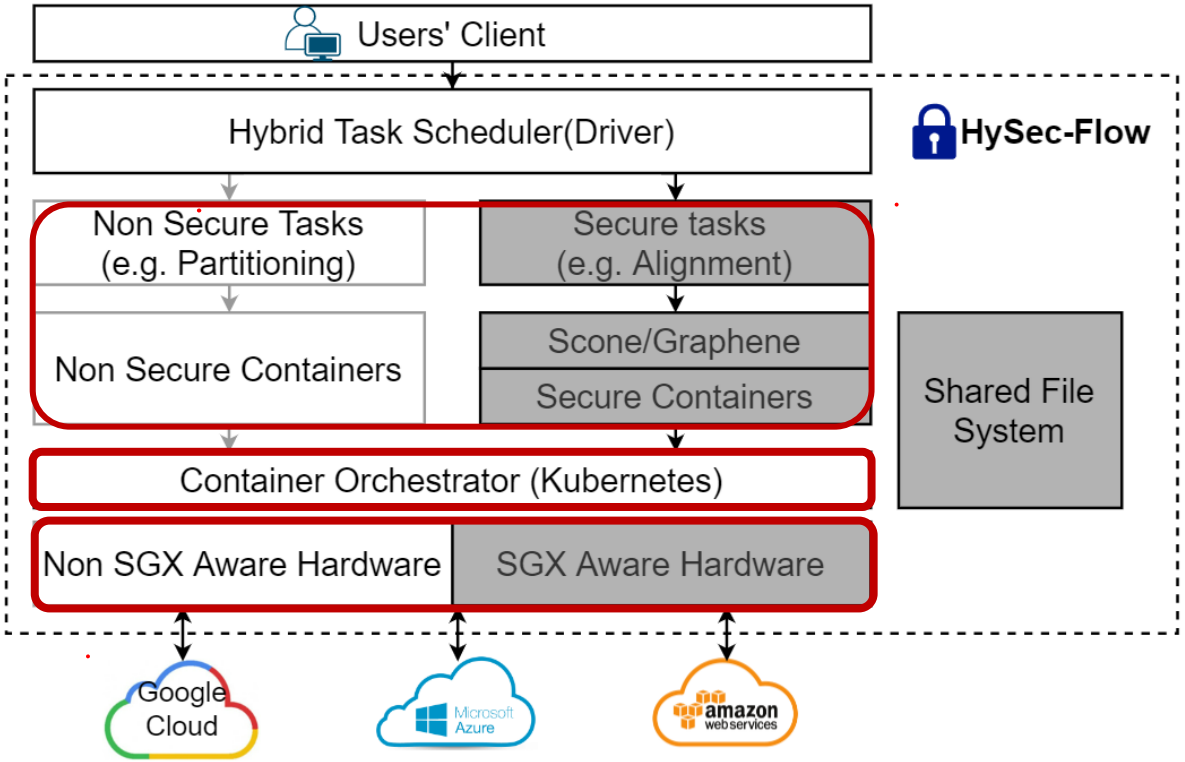
- Against software attacks originated at any privilege level
- Against many hardware based attacks

Applications are modified split into trusted and untrusted parts



UVA | SCHOOL of DATA SCIENCE

HySec-Flow Framework Overview



The conventional (Untrusted) workflow

Distribute

the set of target sequences is partitioned into several subsets

Index

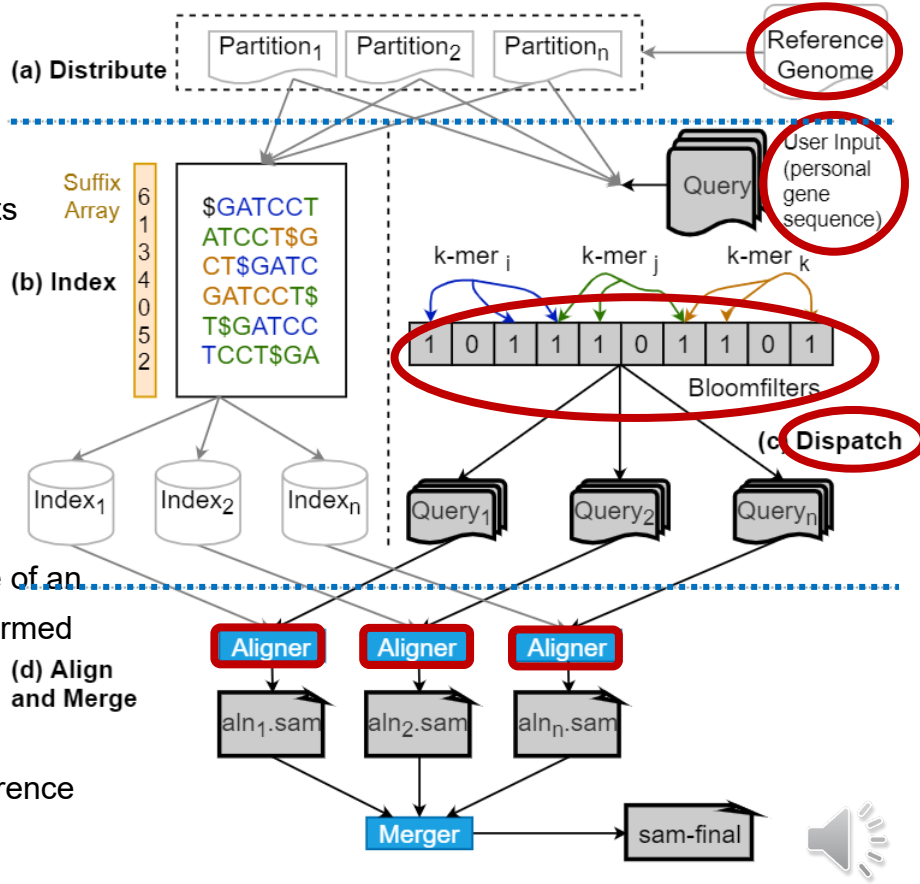
The partitions generated are indexed using a popular read alignment tool like BWA. This operation can be performed parallelly on each partition utilizing the available computing resources of the cluster.

Dispatch

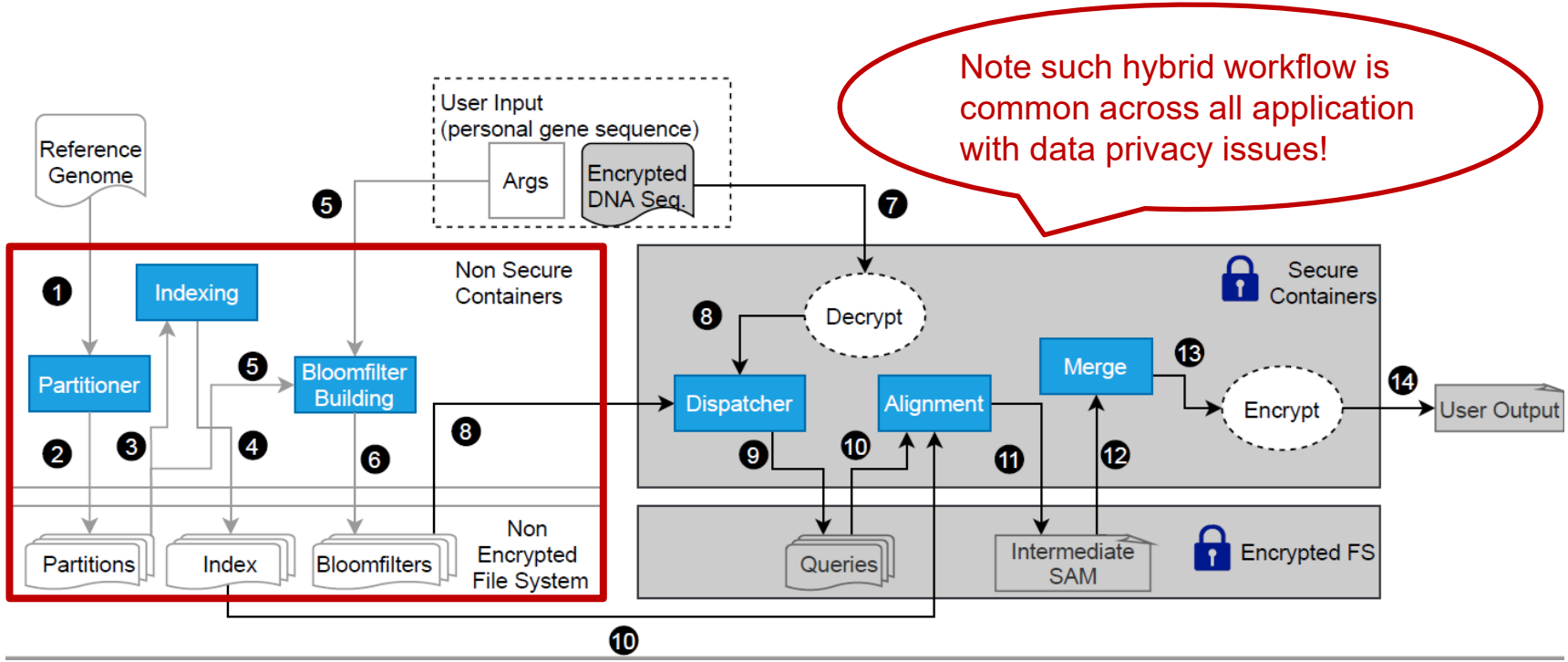
The dispatch stage is performed to reduce the search space of an input DNA sequence within each partition. This can be performed by utilizing many application-dependent techniques.

Align & Merge

aligns short personal sequencing reads against Human reference genome with the Burrows-Wheeler Aligner (BWA).

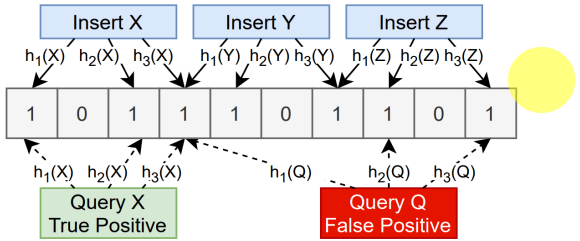


Workflow of Privacy-Preserving Computing Framework



Partition, Indexing and Bloom Filter Building [1 - 6] (non secure)

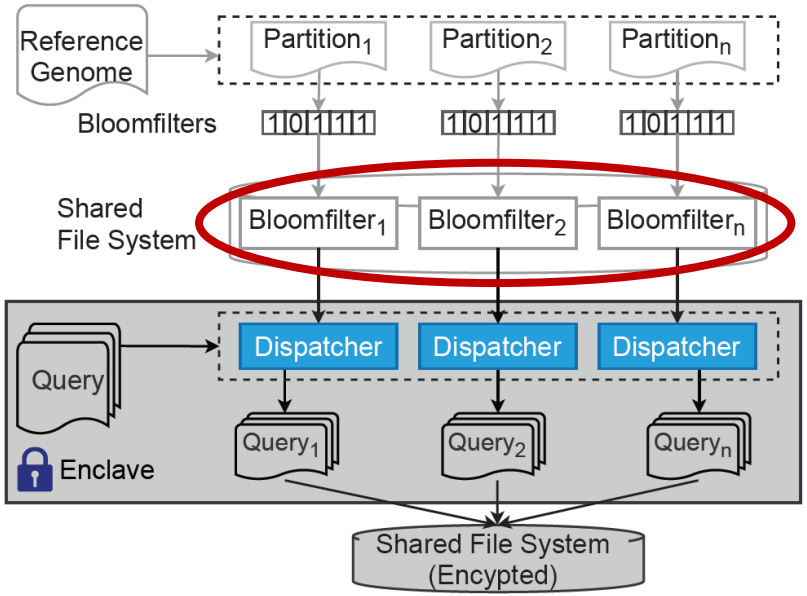
- Split the reference genome sequence into multiple p number of partitions such that each partition can be individually indexed and searched on different nodes of the cluster.
- The partitions generated are indexed using a popular read alignment tool like BWA. This operation can be performed parallelly on each partition utilizing the available computing resources of the cluster.
- Compute a bloom filter for each partition by inserting sub-sequences of length 'b' of the reference genome partition with overlaps of length 'l'.
- These tasks works only on non-sensitive data which are
- All the tasks are executed one time for the same referen



Conventional Bloom filter with $k = 3$ that illustrates the true positive, and false positive.

Dispatch [7 - 9]

- Dispatch is the process of partitioning the user's query(DNA sequence) into p partitions.
- The dispatch step can be parallelly run for each reference genome partition.
- Dispatch works on sensitive data and needs to run inside intel SGX.
- $Query_n$ contains subsequences that would possible present in $Partition_n$
- Outputs are written into a shared file system in an encrypted format.



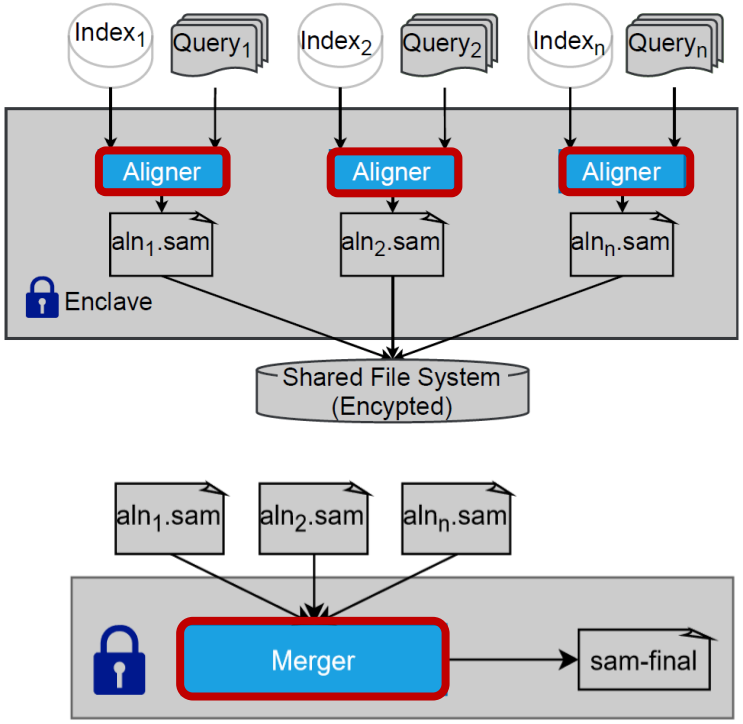
Alignment & Merge [10 - 14]

- Alignment is performed using BWA for each partition.

Algorithm 2: Internal operations within the framework

input : $G = \{g_1, g_2, \ldots, g_p\}$: Reference genome partitions;
I: Input DNA Sequence

```
1  Function DISPATCH(b, I, args):
2      q = []
       // reading sequences of the input
3      for seq in I do
4          for bmer in seq do
5              if b.test(bmer) then
6                  q.append(i)
7      return q
8  Function ALIGNMENT(g, q):
9      return bwa(g, q)
10 Function MERGE(M):
11     S = merge(M) // call DIDA merge
12     return S
```



UVA | SCHOOL of DATA SCIENCE

Security Analysis

- SGX Enclave can protect the code/data integrity even when the executable is loaded into a library OS.
- Disk I/O has been safeguarded by Scone/Graphene's protected filesystem, which utilizes AES-GCM to encrypt user data and immediate data during the computation.
- Under our threat model, the only security risk is key delivery, which is protected by the secure channels we built after trust establishment. Therefore, file tampering attacks can be defeated
- Side channels have been considered to be a threat to trusted execution environments, including SGX.

Security Design

Protected File System

We use SCONE/GrapheneSGX's protected file system to guarantee the all outputs are encrypted.

Attestation and Secret Provisioning

Our attestation and key provisioning mechanism can provide and manage keys for I/O encryption.

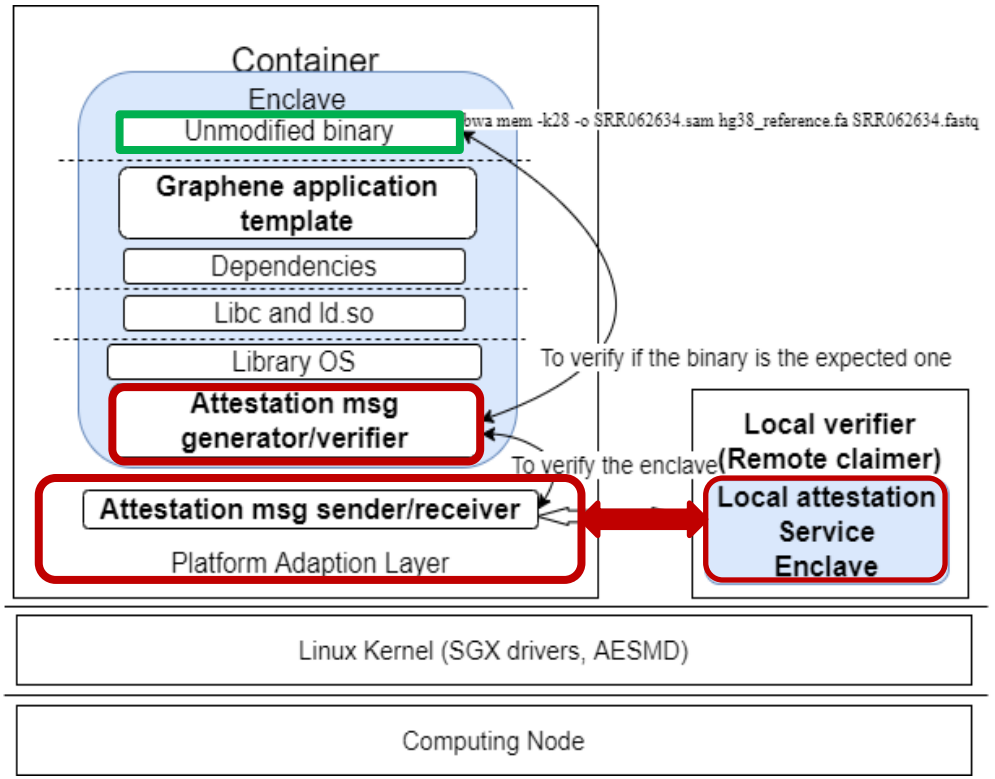
More specifically, we devise an attestation plane, including

- An attestation management service,
- A local attestation service on each computing node,
- A modified Graphene container with attestation interfaces.

Security/Graphene-SGX

Attestation Interfaces

- In Graphene's LibOS:
  - Attestation msg generator
  - Attestation msg verifier
- In Graphene's PAL:
  - Attestation msg sender
  - Attestation msg receiver
- In Graphene's LAS:
  - Local Attestation Server



UVA | SCHOOL of DATA SCIENCE

Experimental Results

1000 Genomes Project

- The 1000 Genomes Project (1000 Genomes | A Deep Catalog of Human Genetic Variation (internationalgenome.org)) is an international research effort to establish largest public catalogue of human variation and genotype data.
- A catalogue of common human genetic variation, using openly consented samples from people who declared themselves to be healthy.
- The reference data resources generated by the project remain heavily used by the biomedical science community.

Experimental setup & data set

Our experiments are conducted on a **10-nodes SGX-enabled cluster**, with each node has an **Intel(R) Xeon(R) CPU E3-1280 v5 @ 3.70GHz CPU** and **64G RAM**. The SGX enclaves are initialized with 8GB heap space with both Scone and Graphene.

Data Set	Source	# Reads	Base pair/read
SRR062634.filt.fastq	1000 Genomes	309K	100
SRR062634_1	1000 Genomes	24M	100
SRR062634_2	1000 Genomes	24M	100

SGX overhead

Overhead from enclave initialization time

- We measure by varying the HeapMaxSize 16M, 64M, 256M, 1024M, 4096M.
- 0.04 seconds per MB of 4096M heap size.

Overhead from OCall/Ecall (per million calls)

- Ocall: 5.27 seconds
- Ecall: 4.65 seconds
- Call in Untrusted environment: 1.3 milliseconds

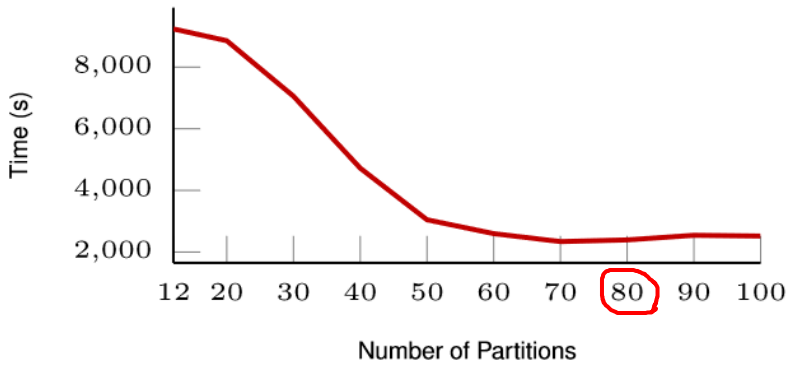
Overhead from EPC page swapping

- An enclave utilizes Processor Reserved Memory (PRM), which is **128MB**.
- The usable memory size for an SGX application is only around **90MB**.
- EPC page swap occurs when a larger data set may not fit into this space.

Optimal partitions for splitting the reference genome

The runtime is measured by sequentially run the alignment for dispatched reads on one single node using SGX via Scone. When the number of partitions is greater than 60, it got flattened.

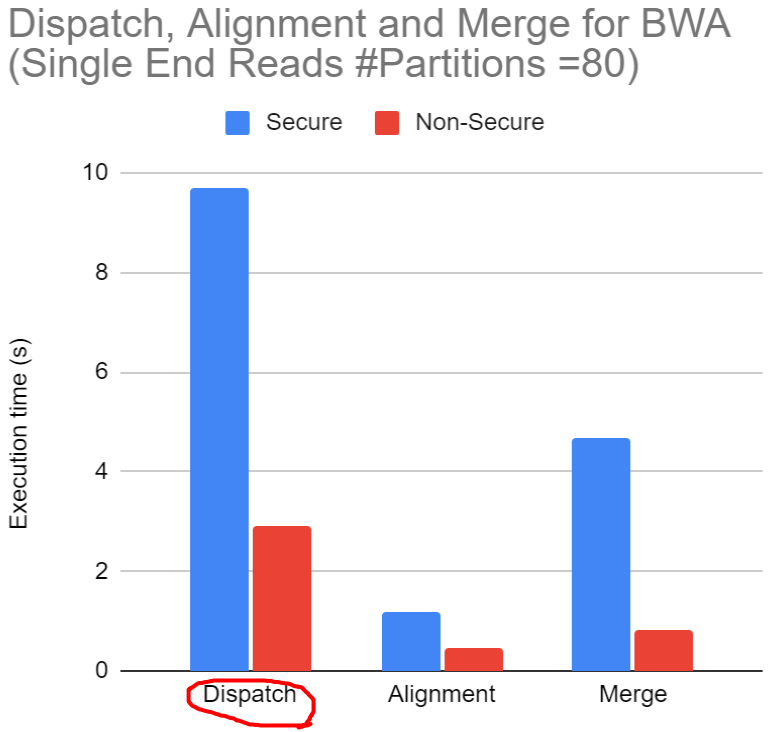
Human reference genome data is about **3.2 GB**, this translates to the reference partition size around or smaller than 50 MB. With the usable memory space around **90 MB** for SGX, this optimal configuration suggests that the entire indexing table can fit into the SGX EPC to minimize the unnecessary EPC swapping, thus improving the overall performance.



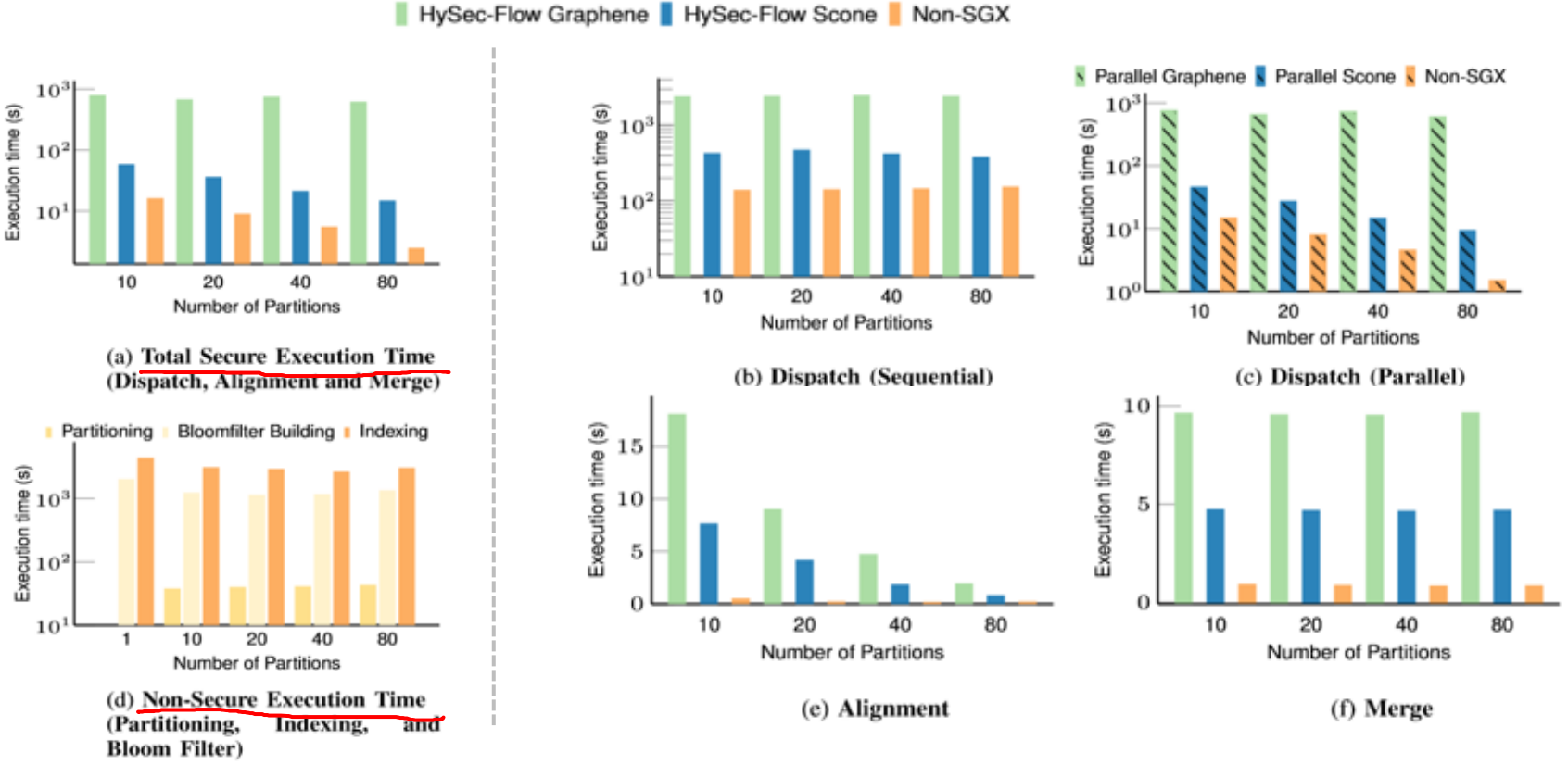
Hybrid-Secure vs. Non-Secure

In the best case, we can run the single end alignment pipeline securely in **15.5** seconds (9.68s in parallel dispatching, 1.18s in parallel alignment over 80 nodes and 4.66s in merging) by partitioning the problem into 80 subtasks.
In the worst case: **793** seconds on one SGX enabled node.

This result will be dramatically improved with new Intel hardware with larger enclaves.

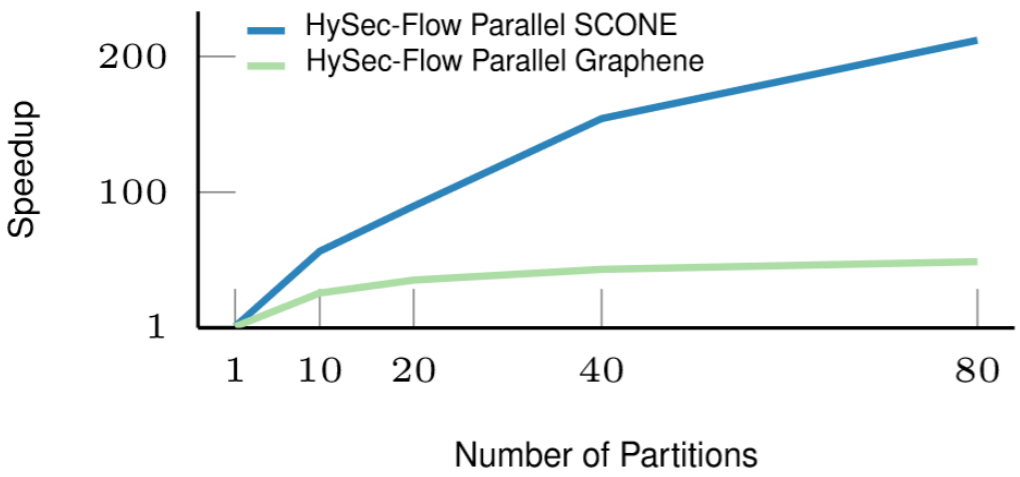


Comparison of the HySec-Flow execution time of Scone & Graphene



Speedup of HySec-Flow over Scone and Graphene

The best-case of HySec-Flow execution time (15.52 seconds) is **212x** speedup compared to Scone execution (3291s) respectively.



Intel's new hardware and large Enclave

We plan to conduct experiments on the latest Intel 3rd Gen Scalable Xeon Processors, since they are equipped with larger **EPC size up to 512 GB**(or 1TB for 2-socket system). Thus, we can hold larger genomic dataset totally in the enclave and avoid extensive paging overhead.

The tentative experiment platform looks like this:

- Server: Supermicro X12 Ultra System
- CPU: Intel®Xeon®Gold 5318S Processor (36M Cache, 2.10 GHz)
- RAM: 256GB DDR4 ECC

UVA | SCHOOL of DATA SCIENCE

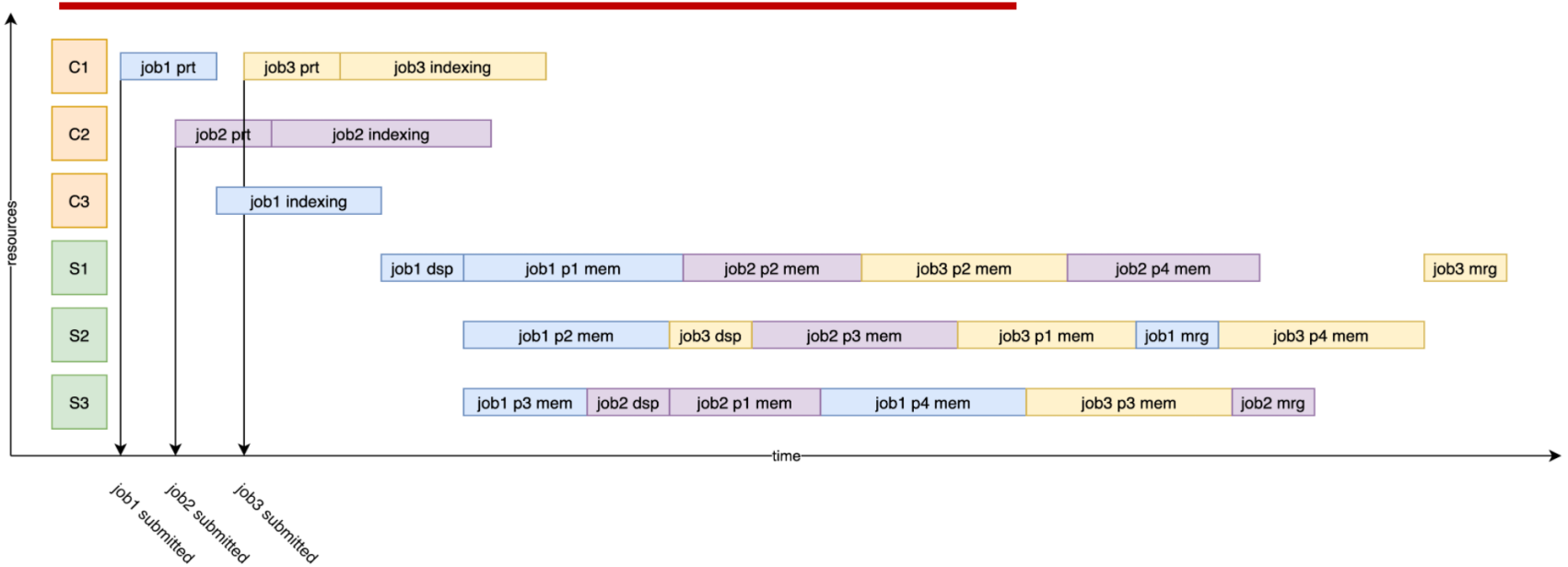
Conclusions & Future Work

HySec-Flow Conclusions

- A novel workflow architecture using **hybrid computing to address heterogenity and performance issues in privacy-preserving computing** .
- Speedup is up to 212x (for 80 partitions) executing BWA sequence alignment using human reference genome, in contrast to running directly in the Scone framework on Intel's SGX hardware.
- The speedup is mainly achieved from the hybrid execution with process level parallelism as well as significantly reduced search space from the bloom filter based dispatch step.
- HySec-Flow can be easily adapted to many other unmodified genomics applications, including cases where the algorithms are data-parallel:
  - genome variation calling
  - gene expression analysis using RNA-seq data
  - peptide identification in clinical proteomics

Future HySec-Flow Work

- By adding a new 'driver' component, HySec-flow can securely accept jobs from users and assign unmodified applications on demand from a pool of secure and non-secure containers.



Acknowledgement

This work is partially supported by

- NSF grant No.1838083 on BIGDATA: IA: Enabling Large-Scale, Privacy-Preserving Genomic Computing with a Hardware-Assisted Secure Big-Data Analytics Framework,
- NSF grant CCF-1918626 Expeditions: Collaborative Research: Global Pervasive Computational Epidemiology,
- NSF grant No. 1835631 CINES: A Scalable Cyberinfrastructure for Sustained Innovation in Network Engineering and Science,
- NIH R01HG010798: Secure and Privacy-preserving Genome-wide and Phenome-wide Association Studies via Intel Software Guard Extensions (SGX).

National Institutes of Health
Turning Discovery Into Health