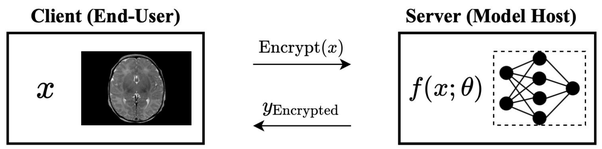
Characterizing and Optimizing End-to-End Systems for Private Inference

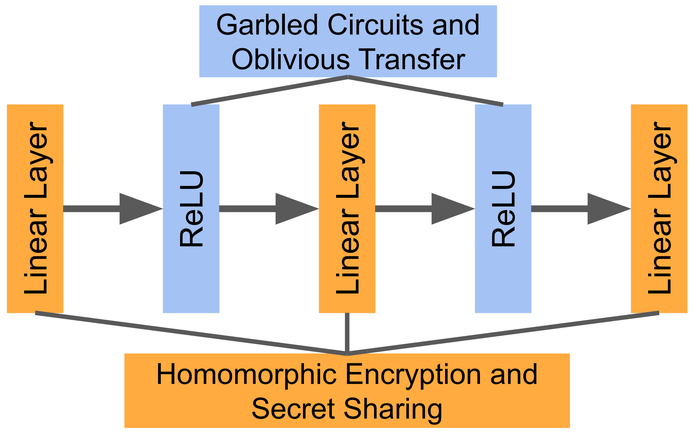
Karthik Garimella[1], Zahra Ghodsi[2], Nandan Kumar Jha[1], Siddharth Garg[1], Brandon Reagen[1]
kg2383@nyu.edu
New York University[1], Purdue University[2]

Introduction

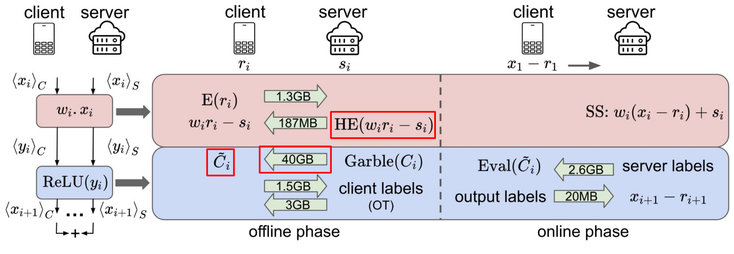
Privacy concerns in client-server machine learning services have given rise to private inference (PI), where neural inference occurs directly on encrypted inputs [1]:



Private Inference is constructed using cryptographic techniques: **homomorphic encryption (HE) and additive secret sharing** for the linear layers and **garbled circuits (GC) with oblivious transfer** for ReLU activations:



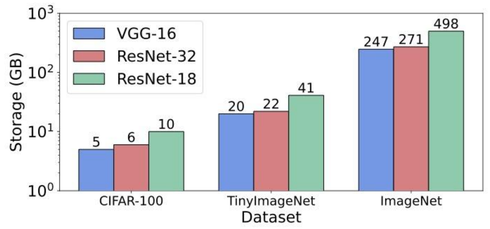
A Hybrid PI Protocol



Most PI Protocols are divided into an offline and online phase. The components highlighted in **red** show the bottlenecks that prevent PI protocols from handling realistic workloads of inference requests. Prior work assumes preprocessing costs can be ignored.

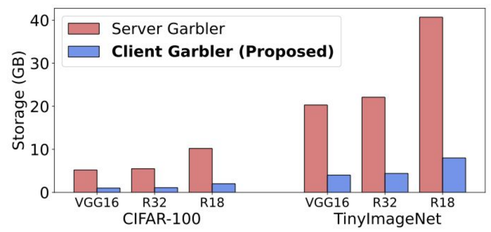
System Costs of Private Inference

Client-Side Storage

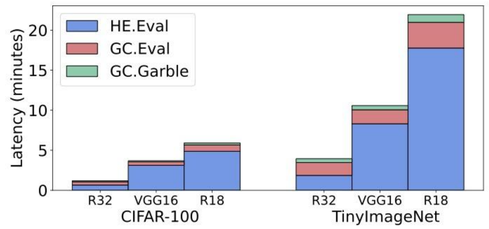


During the offline phase, the client must store the garbled circuits representing ReLU (18 KB /ReLU).

Client-Garbling: By switching the roles of the GC garbler and evaluator, the server can store the garbled circuits rather than the client. **A 5X increase in number of of available precomputes.**

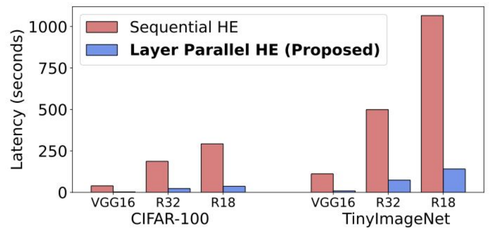


Compute

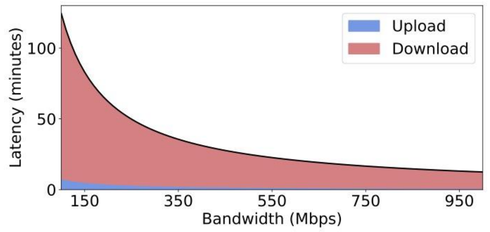


The homomorphic evaluation of linear layers in the offline phase accounts for most compute latency.

Layer Parallel Homomorphic Encryption: The homomorphic evaluation of the linear layers are independent of each other and can be evaluated in parallel. **HE evaluation speedup of 10X.**

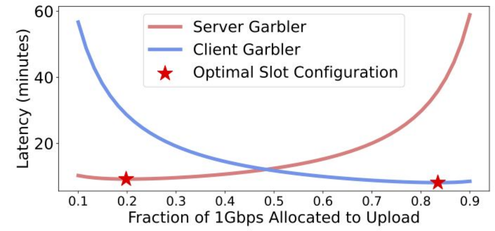


Communication



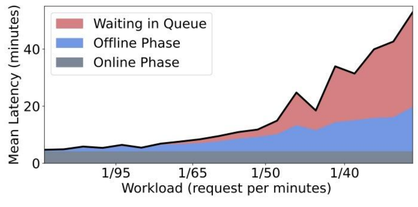
The imbalance in uploaded and downloaded data is caused by the transmission of GCs to the client.

Wireless Slot Allocation: Current 5G standards allow for allocating more upload bandwidth to account for garbled circuit transmission. **WSA reduces communication latency by up to 35%.**



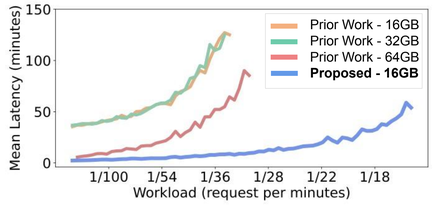
Results

Prior Work



Prior work optimizes for the online phase. However, offline phase costs affect the client's observed latency when considering multiple inferences requests.

Our Optimizations



Our proposed optimizations enable a higher maximum sustainable throughput and lower latencies. We also reduce client-side storage requirements [2].

References

[1] Pratyush Mishra, Ryan Lehmkuhl, Akshayaram Srinivasan, Wenting Zheng, and Raluca Ada Popa. Delphi: A cryptographic inference service for neural networks. In 29th USENIX Security Symposium, 2020

[2] Karthik Garimella, Zahra Ghodsi, Nandan Kumar Jha, Brandon Reagen, and Siddharth Garg. Characterizing and Optimizing End-to-End Systems for Private Inference. ASPLOS, 2023.

Paper

Cryptonite (code)