cuOT
Accelerating Oblivious Transfer on GPUs for Privacy-preserving Computation

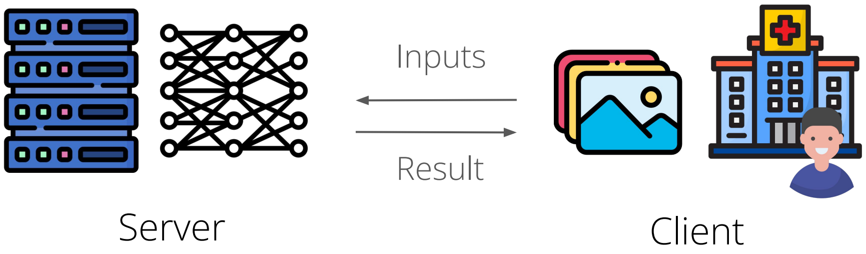
Andrew Gan, Setsuna Yuki, Timothy Rogers, Zahra Ghodsi
Purdue University

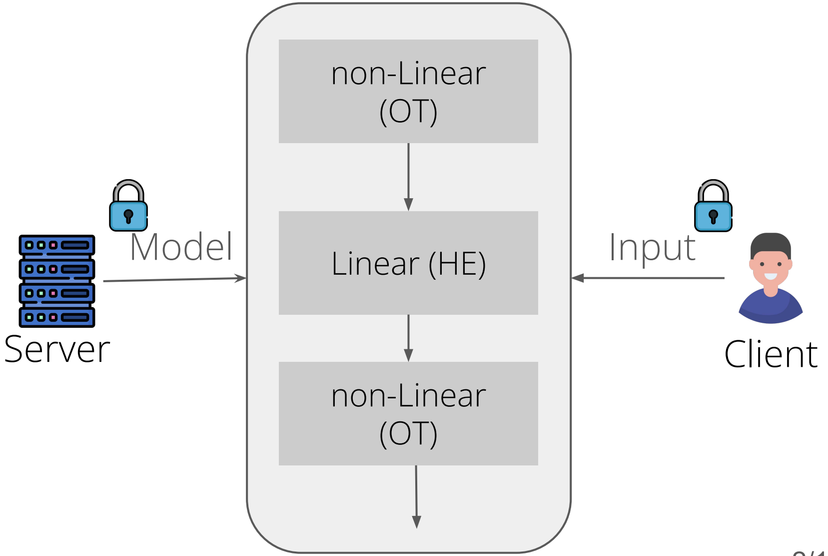
IEEE HOST 2025

PURDUE UNIVERSITY

Privacy-preserving Computation

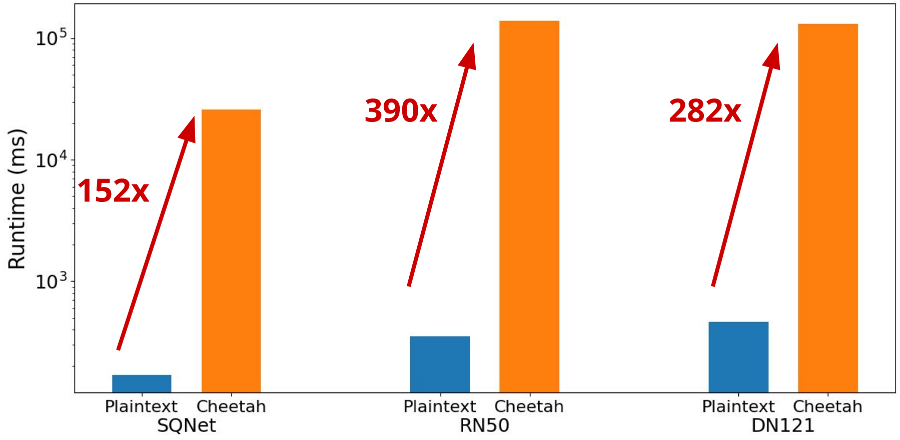
- An increasing number of applications work with sensitive user data, requiring privacy guarantees.
- Cryptographic protocols such as homomorphic encryption (HE) and oblivious transfer (OT) enable privacy-preserving applications.
- State-of-the-art frameworks use hybrid protocols:
  - HE → linear operations
  - OT → non-linear operations



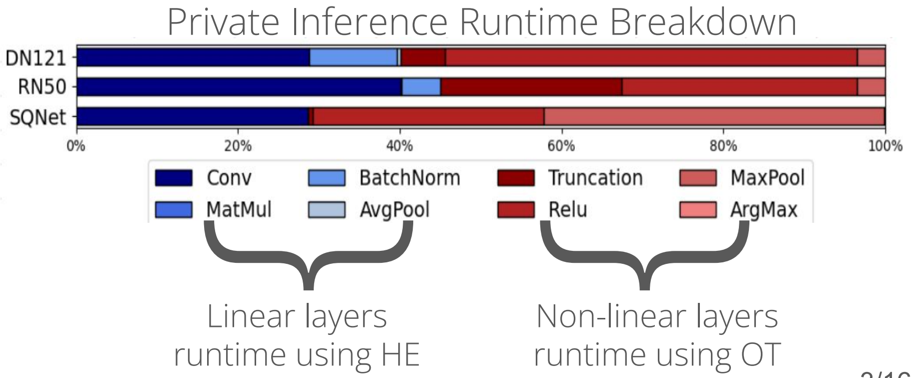


Is Privacy-preserving Computation Practical?

- Privacy-preserving computation incurs a large overhead:
  - Private inference (based on Cheetah [1]) is ~390x slower than plaintext

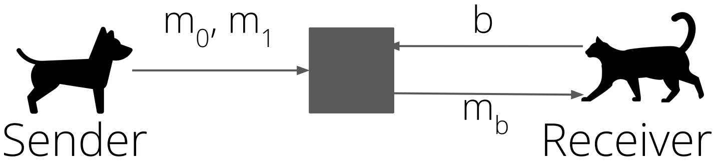


- **Solution**: Use hardware acceleration to bring privacy-preserving computation closer to practicality



[1] Z. Huang, et. al, “Cheetah: Lean and fast secure Two-Party deep neural network inference,” in USENIX Security 22.

Oblivious Transfer (OT) Protocol



- OT relies on expensive public crypto primitives

- More efficient constructions are based on OT extensions
  - use few base OTs (with public crypto)
  - extend them to many OTs (with cheap symmetric crypto)

Less computation

More communication

IKNP — Yuval Ishai, et al., Crypto 2003

Softspoken — Lawrence Roy, 2022

Ferret — Kang Yang, et al. CCS 2020

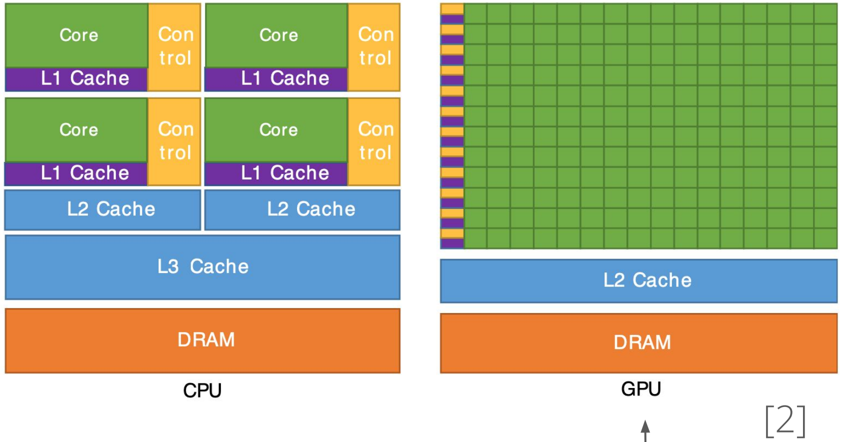
Silent — Elette Boyle, et al. Crypto 2019

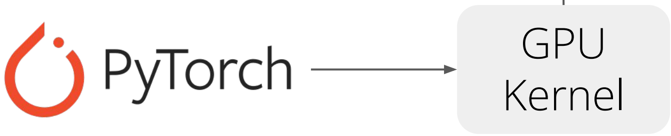
More computation

Less communication

Why Accelerate OT on the GPU?

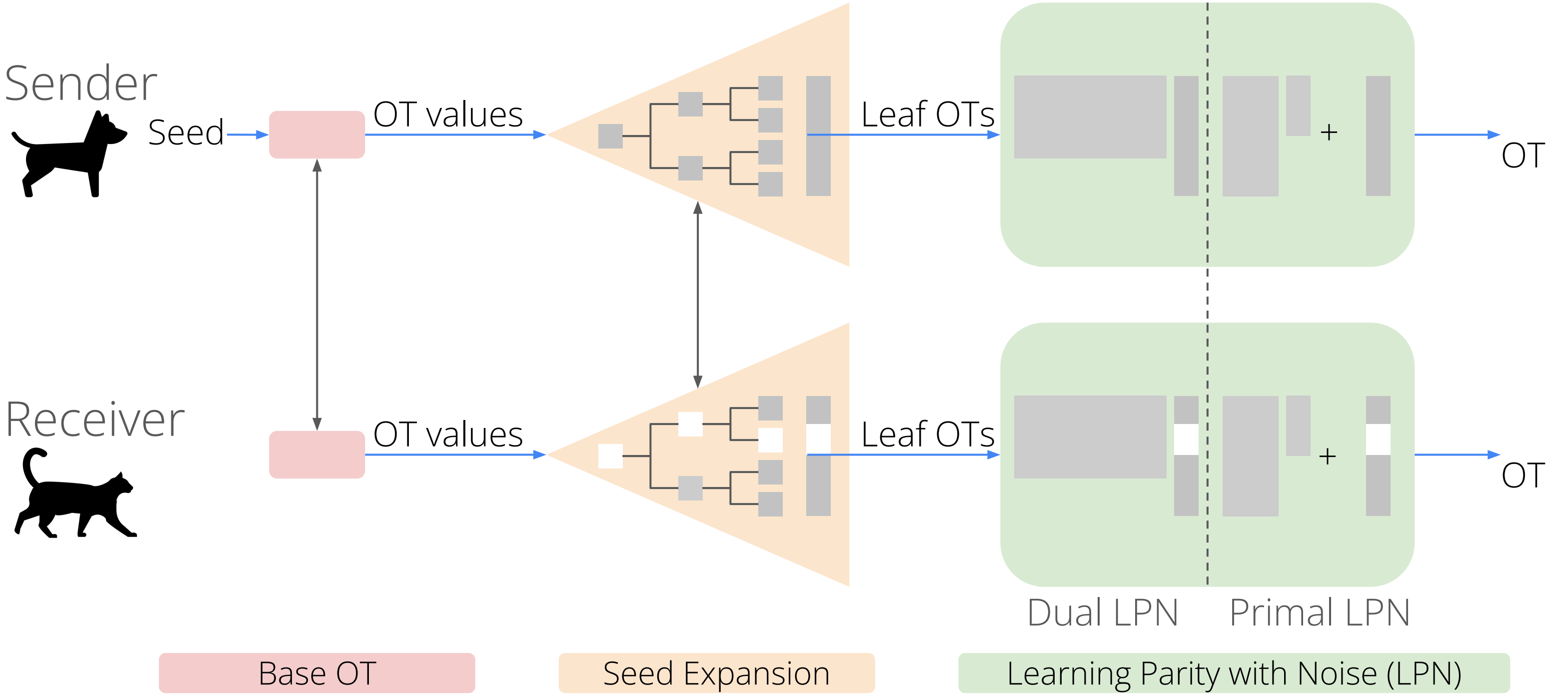
- GPUs are more common than FPGAs or other customized accelerators
- Parallelizable and reprogrammable
  - Computation pattern of OT better suited on GPU
- Popular applications like ML are executed on the GPU



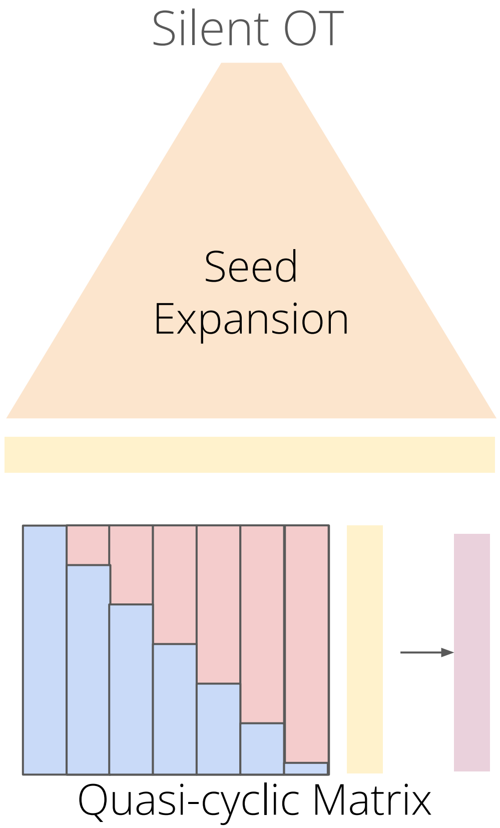


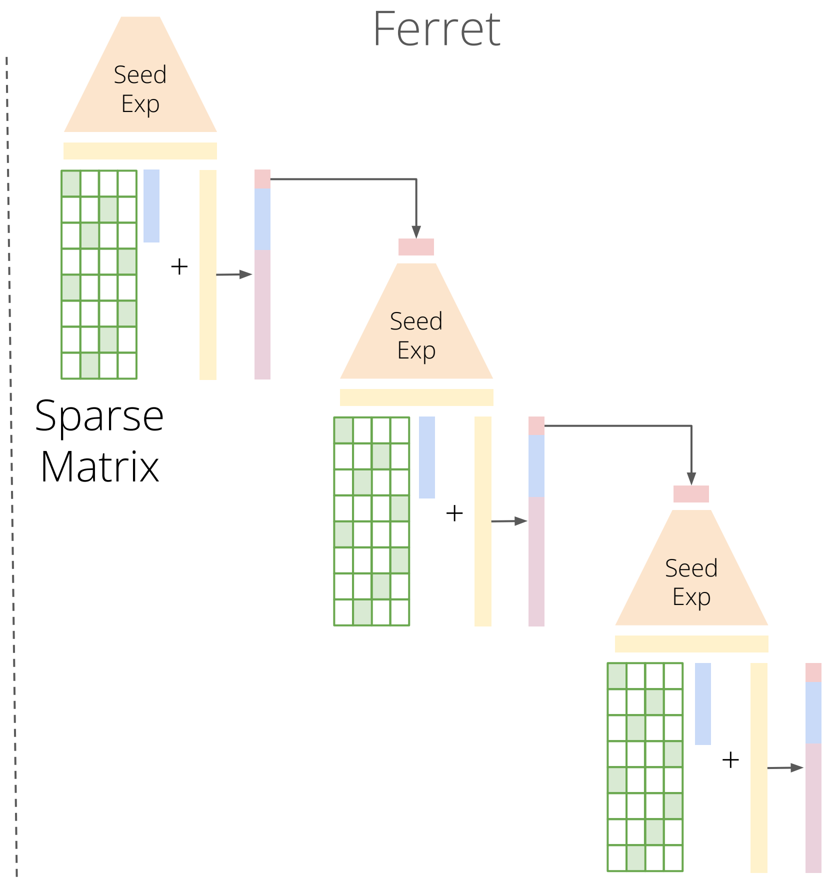
[2] https://cvw.cac.cornell.edu/gpu-architecture/gpu-characteristics/design

System Overview of Silent OT and Ferret

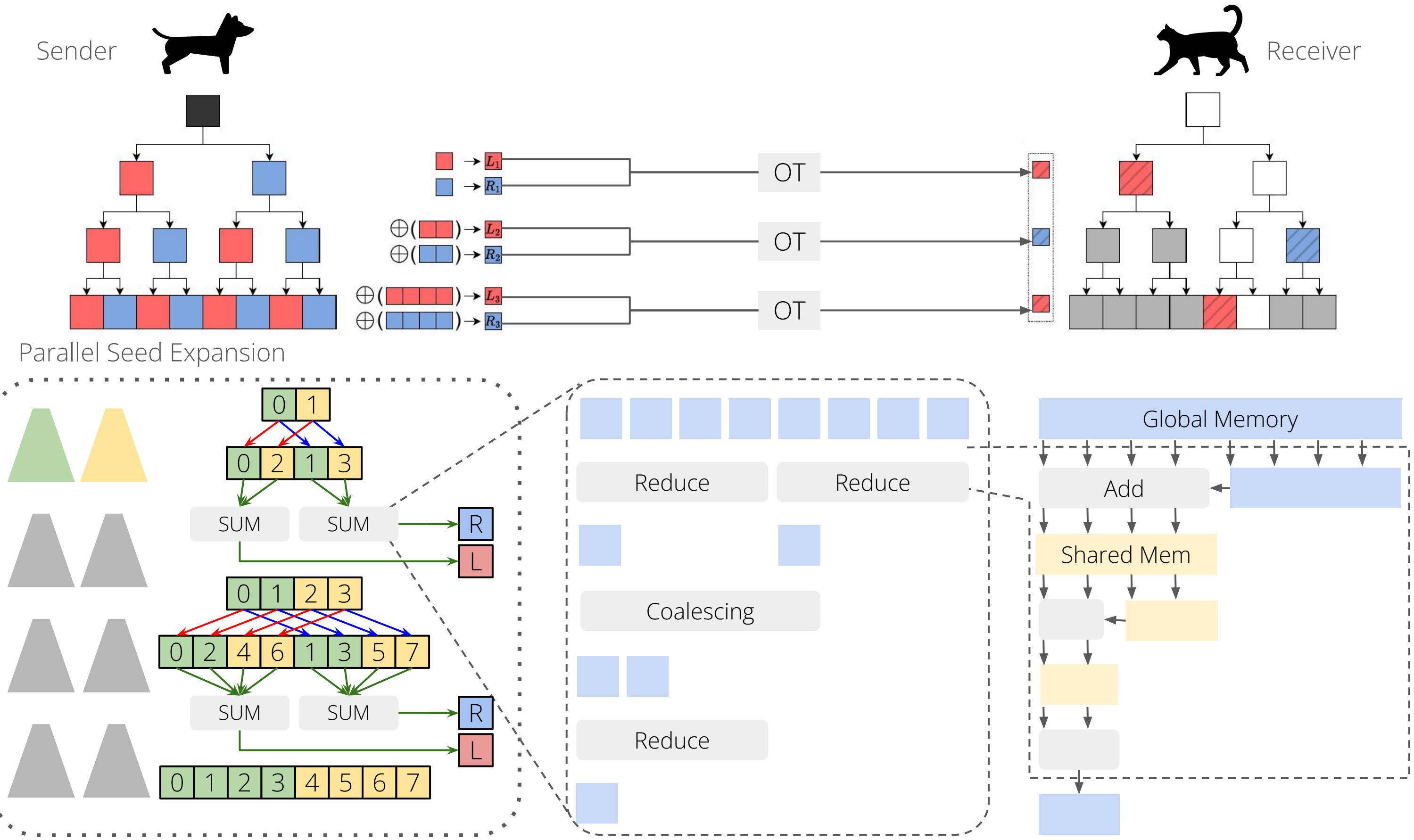


Silent OT and Ferret Computation Steps

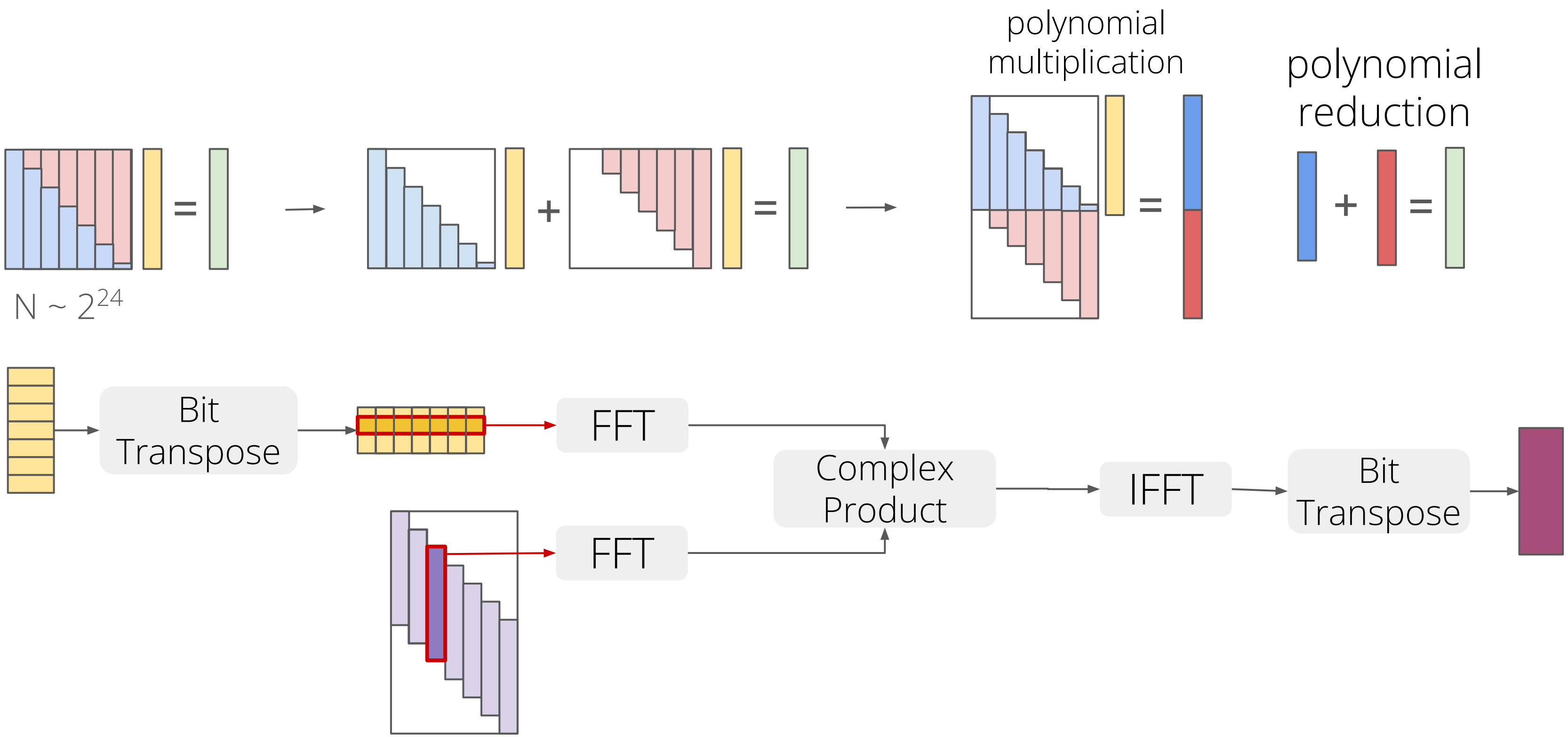




cuOT: GPU Acceleration of Seed Expansion

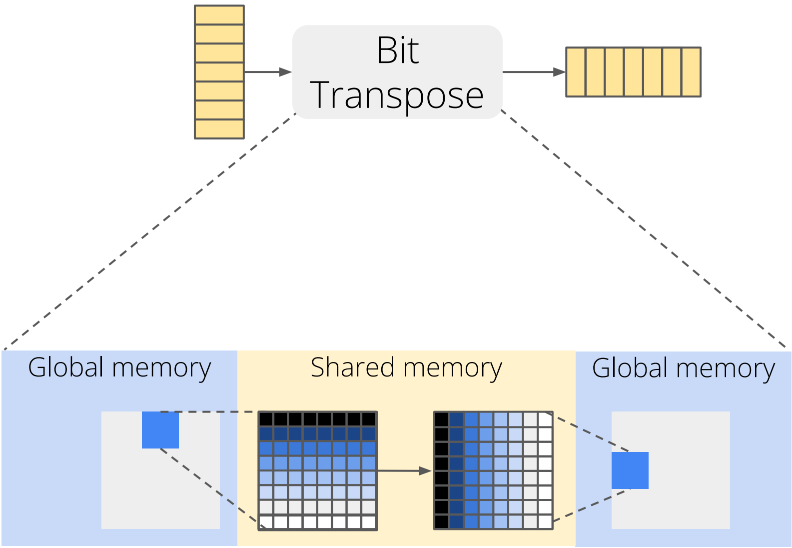


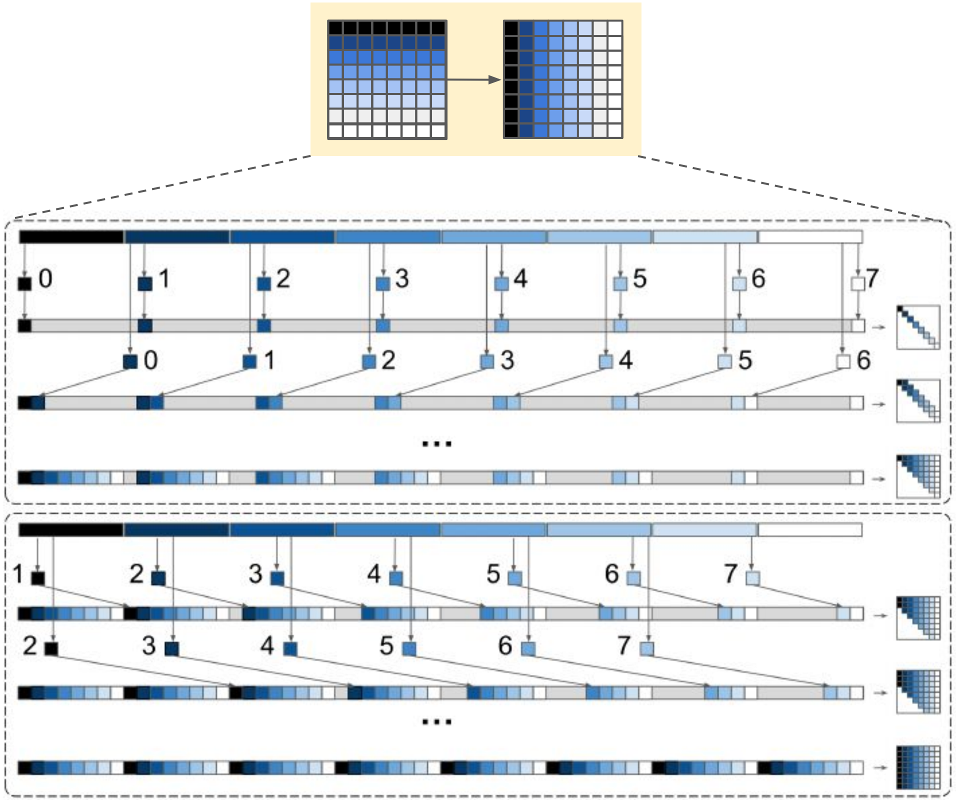
cuOT: GPU Acceleration of Dual LPN



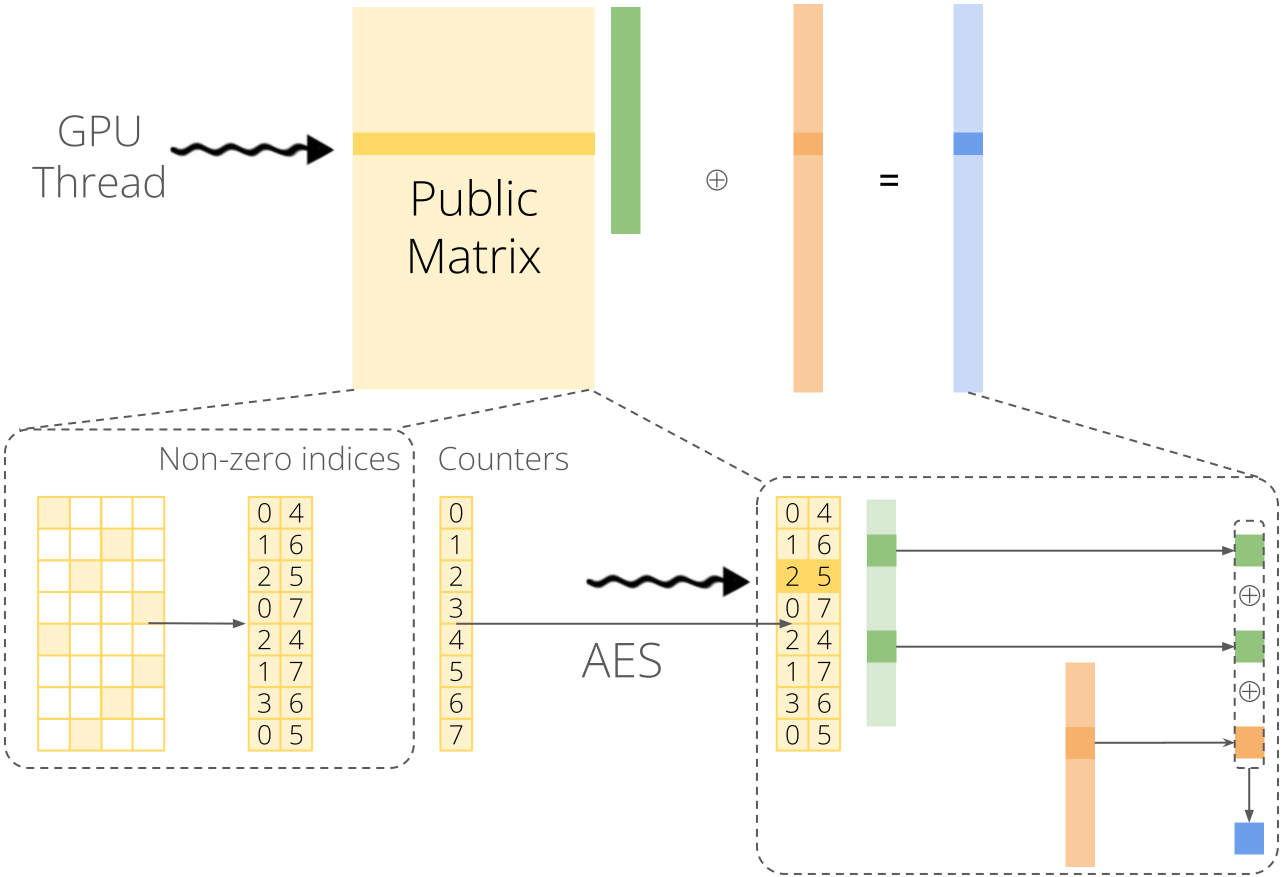
PURDUE UNIVERSITY

cuOT: Custom GPU Kernels for Dual LPN



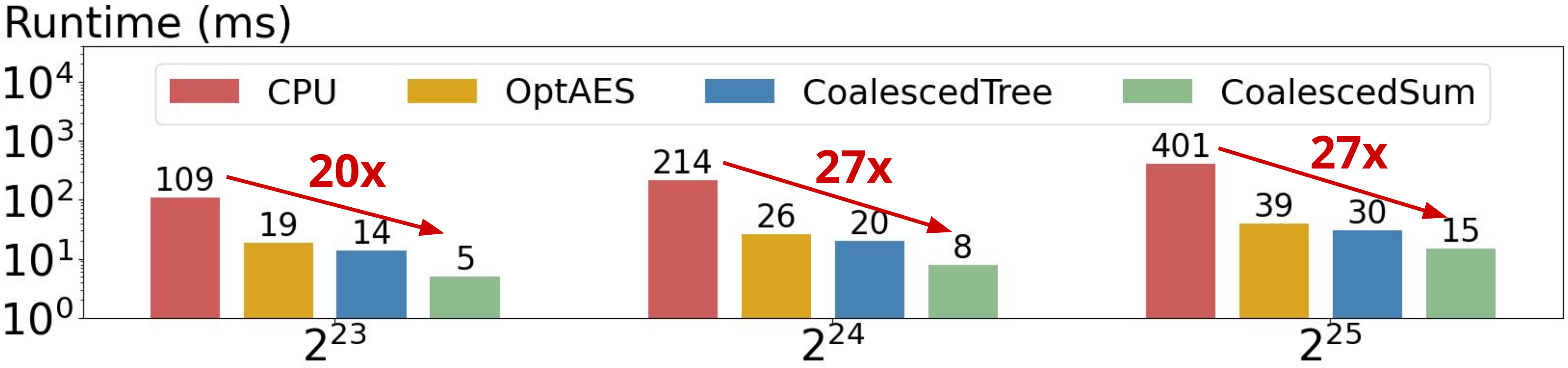


cuOT: GPU Acceleration of Primal LPN



cuOT Seed Expansion Speedup

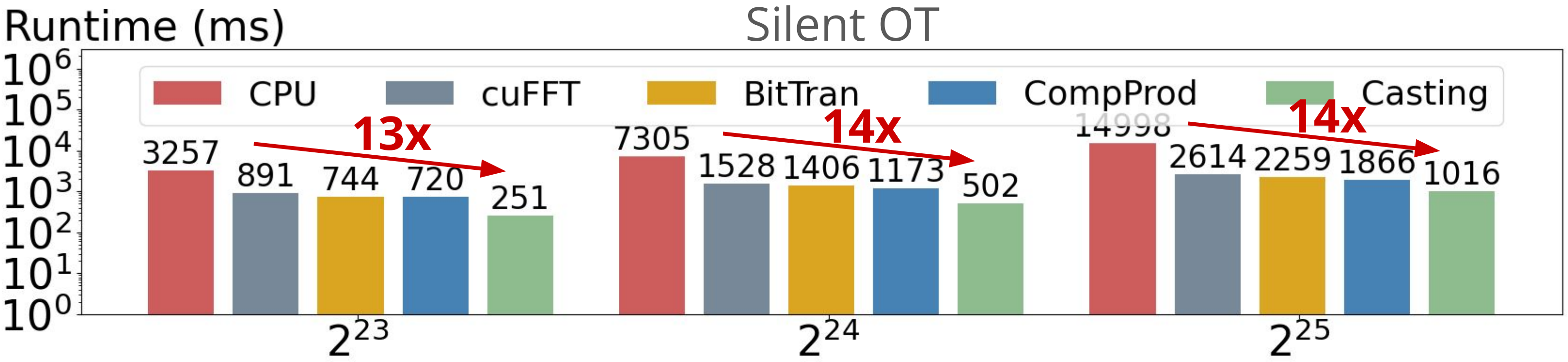
- cuOT achieves up to 27x speedup for seed expansion

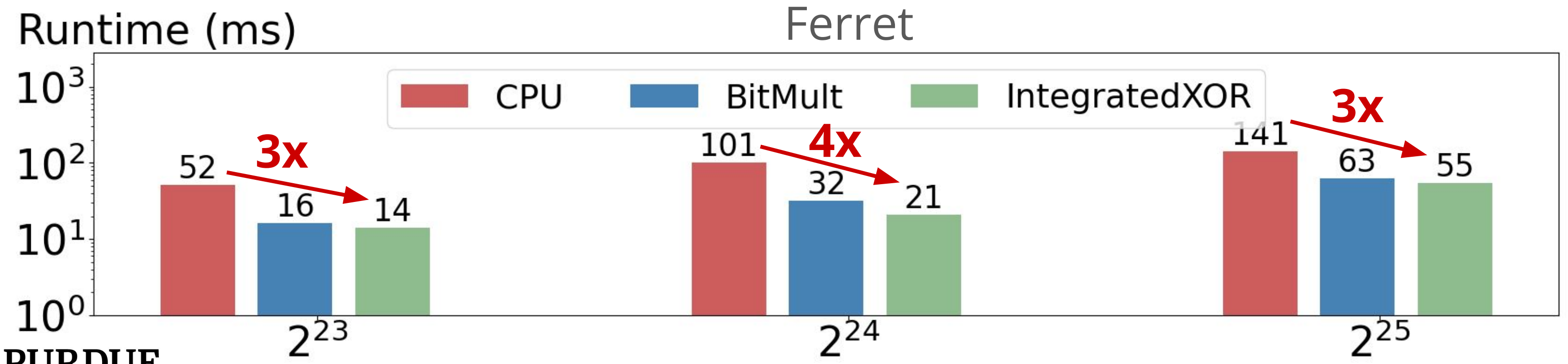


evaluation results collected on RTX A6000 GPUs

cuOT LPN Speedup

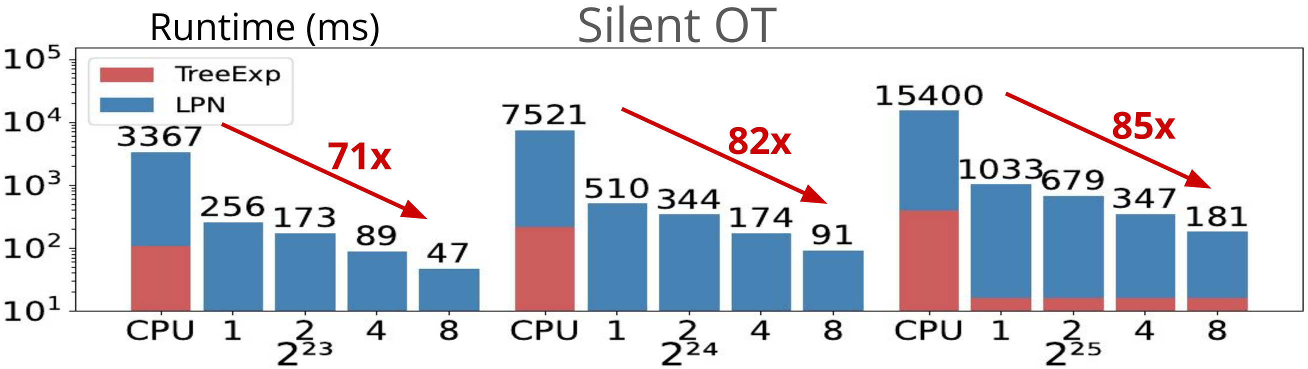
- cuOT achieves up to 14x speedup for LPN

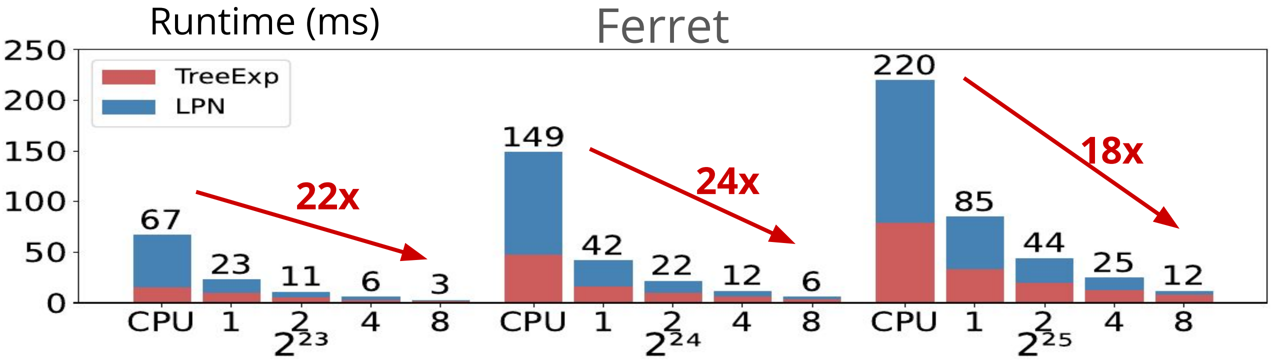




cuOT End-to-end Runtime Benefits

- cuOT achieves up to 85x speedup on 8 GPUs for Silent OT, 24x for Ferret





- cuOT results in 42% reduction in non-linear layer runtime for DenseNet121

Conclusion

- Protocols for privacy-preserving computation can benefit from acceleration on ubiquitous platforms like GPUs

- cuOT achieves an order-of-magnitude speedup in generating millions of OTs compared to CPU baseline

Backup slides

Silent OT and Ferret Communication Overhead Comparison

For generating 2^25 OTs (~33 million), Ferret incurs 200x more communication overhead

	Silent OT	Ferret
Total communication	13.3 kB	3089.3 kB
Delay (300 Mbps bandwidth)	0.4 ms	82.8 ms