Aalto University

UNIVERSITY OF WATERLOO

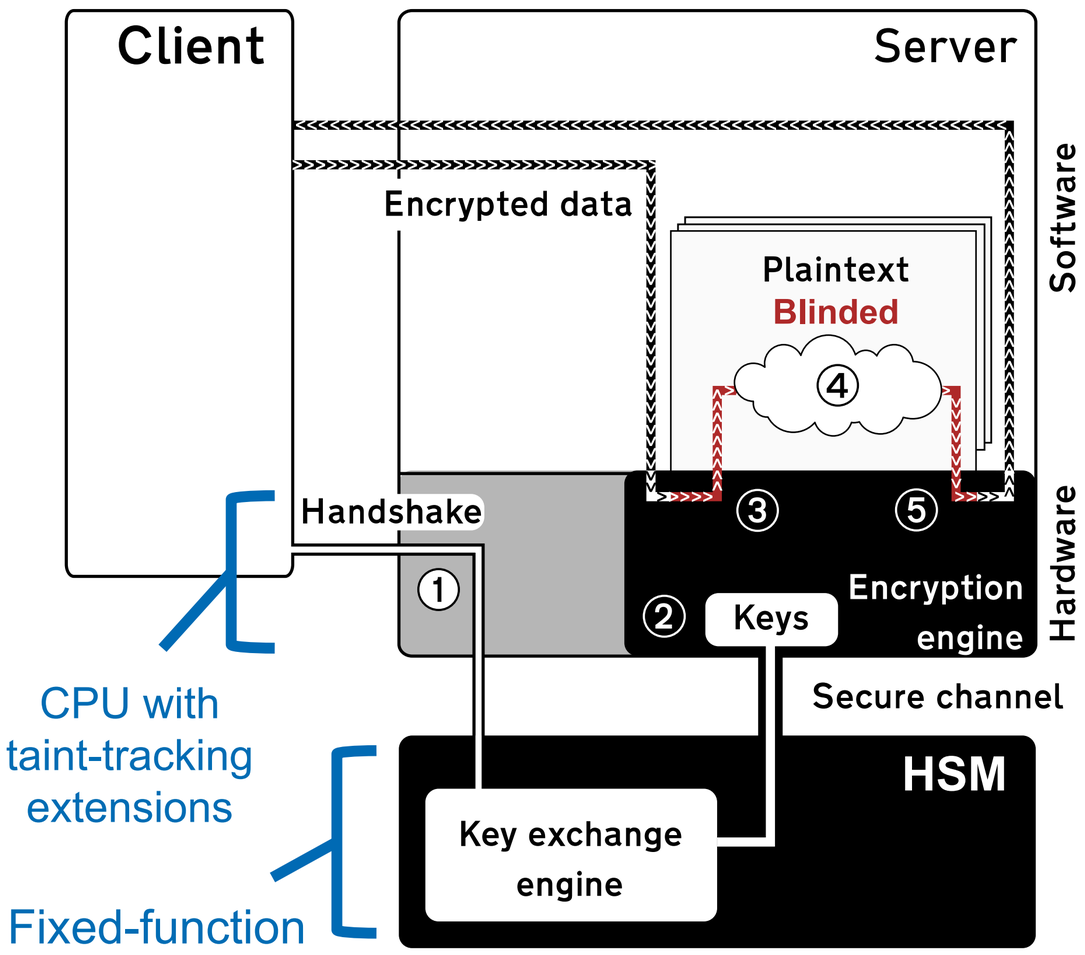
Hossam ElAtali, John Z. Jekel, Lachlan J. Gunn, N. Asokan

Dolma: Data-Oblivious ML Accelerators using Hardware Security Extensions

1. Motivation

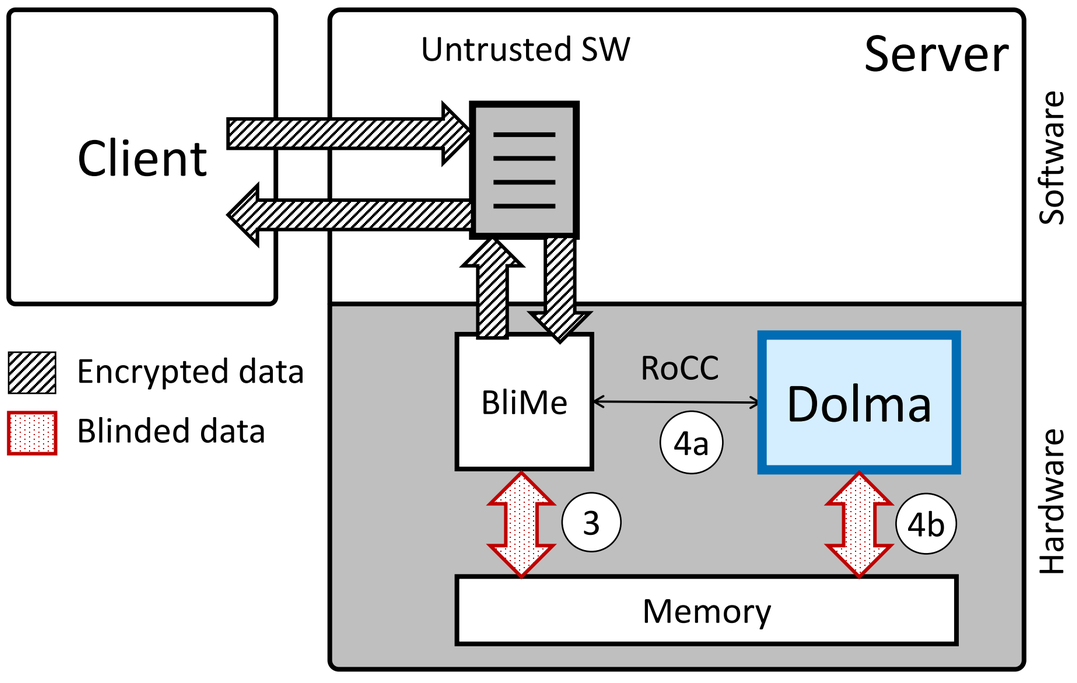
- Outsourced computing is everywhere
- Service providers don't expose models/code
- BUT clients expose sensitive data to providers
- Prior solutions:
  - Crypto solutions (e.g. FHE) still very slow
  - TEEs are prone to side-channel attacks
- State-of-the-art solution: BliMe [1]
  - Taint-tracking-based security policy in HW limits sensitive data to "safe places"
  - BUT only supports CPU workloads

2. Background: BliMe



1. **Handshake (incl. remote attestation)**
2. **Shared secret key**
3. **Atomic data import (inputs)**
   - Decrypt & blind (Blinded ← true)
4. **Safe ("blinded") computation**
   - Enforced by BliMe HW extensions
5. **Atomic data export (result)**
   - Encrypt & unblind (Blinded ← false)

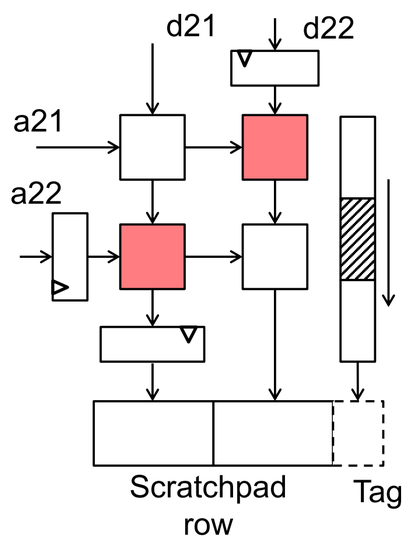
3. Our solution

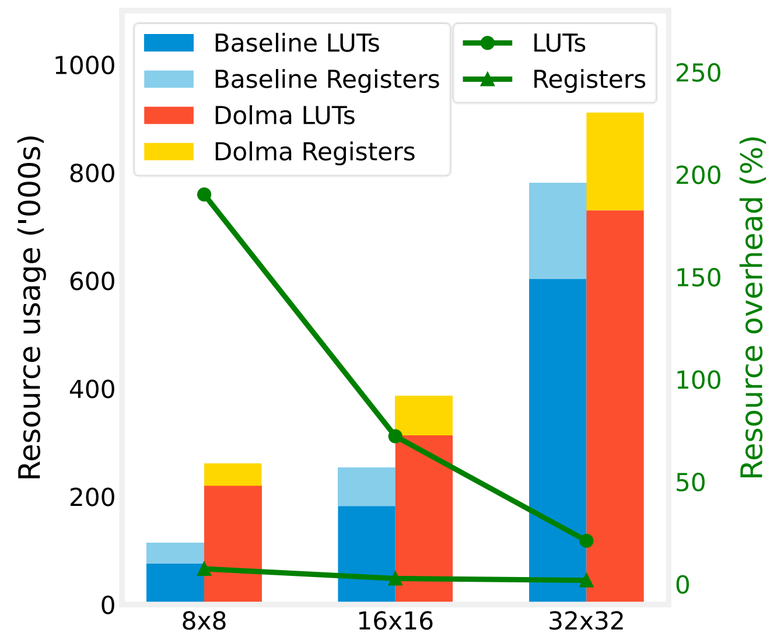


- Adapt taint-tracking-based security policy to matrix-multiplication ML accelerator
- **Prohibit** leaking sensitive data into any observable output, e.g. execution time, memory access patterns

4. Implementation & Results

- Implemented in RTL on Gemmini
- Row-wise taint-tracking
- **Optimization:** Exploit fixed-functionality of systolic array to reduce taint-tracking logic





Average Overheads (vs insecure baseline)	
Performance*	5.6%
Power	14.6%

* Perf. overhead for ResNet-50 classification

[1] ElAtali et al. *BliMe: Verifiably Secure Outsourced Computation with Hardware-Enforced Taint Tracking*. NDSS 2024

Paper

Project

Me!