USC University of Southern California

DEVCOM

HERA: A Bandwidth-efficient Accelerator for Fully Homomorphic Encryption on HBM-enabled FPGA

Zhihan Xu, Rajgopal Kannan, Viktor Prasanna

Session VII: Applications

Presented @ ISFPGA 2026

Feb. 24, 2026

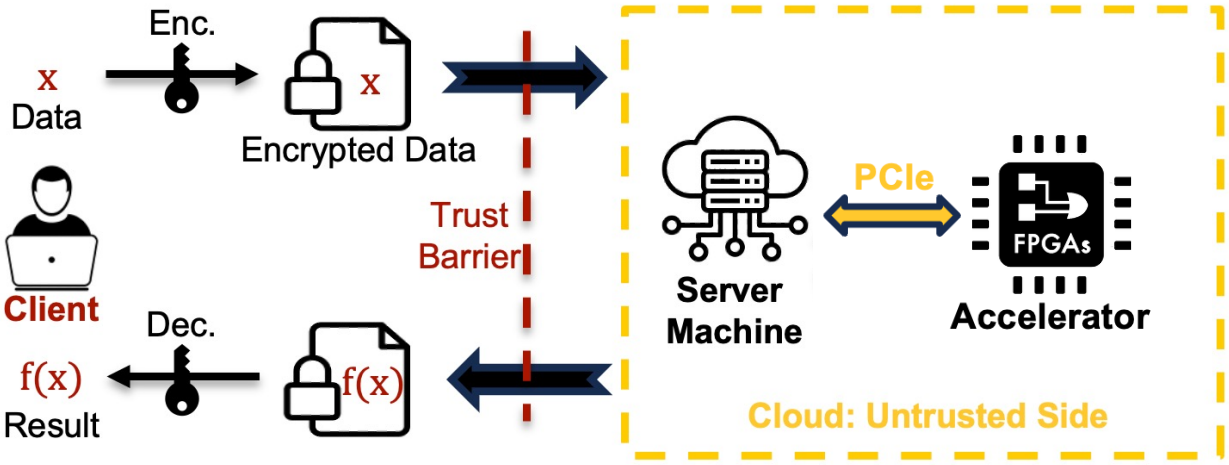
USC

Outline

- Introduction
- Motivation
- HERA - Architecture & Contribution
- Evaluation
- Conclusion

Fully Homomorphic Encryption (FHE)

- Data privacy issue in cloud computing
- Computation on encrypted data (or Ciphertext, Ct)



modular arithmetic on integer polynomials

But …

Large computational & memory overhead

- Ct expansion: $100\times - 1000\times$
- $10^4 - 10^6\times$ slower

- Complex subroutines
  - Number Theoretic Transform (NTT)
  - Automorphism (Auto)
  - …

Memory access challenges

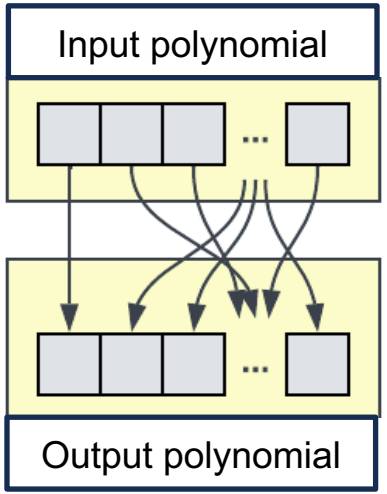
NTT

- FFT in finite field (integer modulo a prime)
- Large strided access
  - ➢ large polynomial degree $(\text{i.e.}, 2^{16})$

Auto

- Coefficient permutation
- Irregular access

Poor locality
Low bandwidth utilization

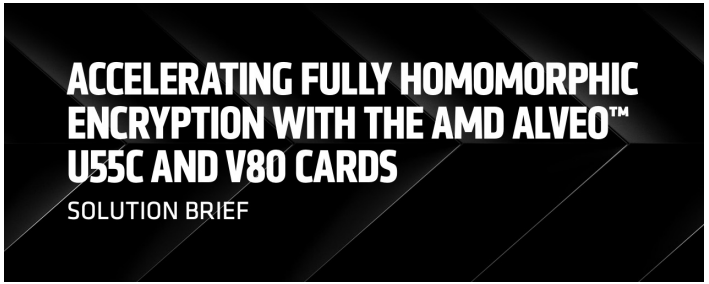


Auto permutation

Why HBM-enabled FPGA

FHE is primarily memory-bound

- U280/U55C: HBM2 (460 GB/s)
- V80: HBM2e (810 GB/s)
- DDR: 38 GB/s



AMD solution brief

A practical solution for FHE acceleration

Outline

- Introduction
- Motivation
- HERA - Architecture and Contribution
- Evaluation
- Conclusion

Motivation (1)

Lack of systematic analysis of HBM bandwidth utilization

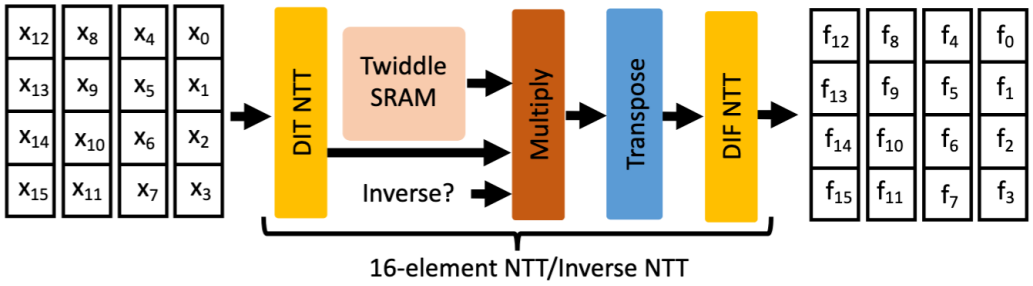
- Channel allocation
- Mapping of polynomials and operands
- AXI configuration
- Achievable bandwidth

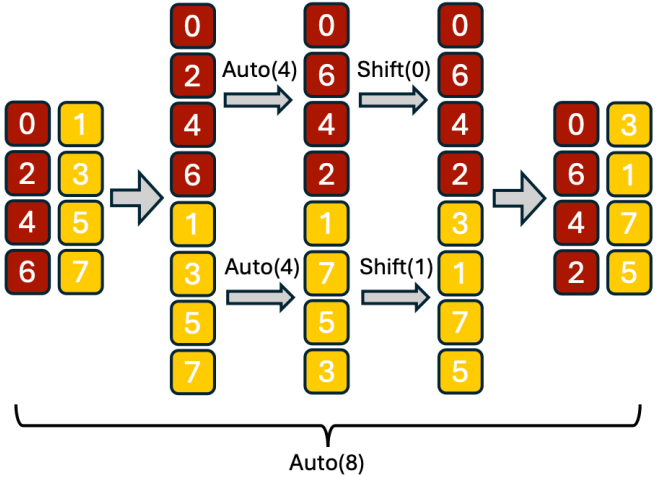
How to exploit HBM bandwidth for FHE acceleration

Motivation (2)

Kernel decomposition of NTT & Auto

- Break a large transform into multiple small independent ones
  - ➢ Improve spatial locality
  - ➢ Parallel processing
  - ➢ Bandwidth-friendly



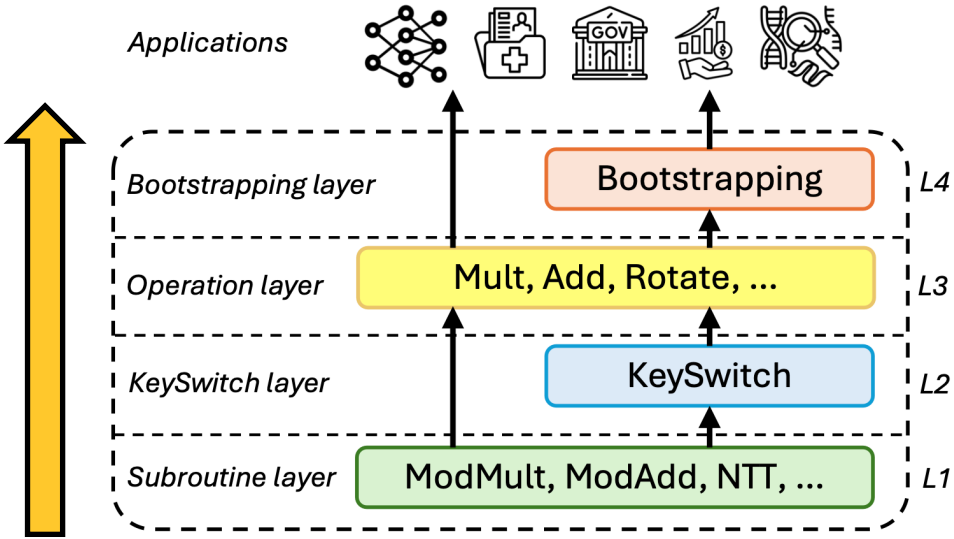


2D-NTT figure from: A. Feldmann, et al. "*F1: A Fast and Programmable Accelerator for Fully Homomorphic Encryption*", MICRO 2021

Motivation (3)

FPGA-aware & hierarchical optimizations

- Holistic algorithm-architecture co-design



Outline

- Introduction
- Motivation
- HERA - Architecture and Contribution
- Evaluation
- Conclusion

HERA Contribution (1)

Kernel decomposition

- Improve bandwidth efficiency
- Better align with hardware

- NTT decomposition (2-D)
  - 256-point NTT ($\sqrt{2^{16}}$)
  - matrix transpose
    - shift operations

- Auto decomposition (2-D)
  - 256-point Auto
    - a single shift

Table 2: NTT and Automorphism design comparison

Work	Decompose NTT	Decompose Auto	w/o Dedicated NTT Units	Unified Perm. Design
F1* [33]	✓	✓	✗	✗
CraterLake* [34]	✓	✓	✗	✗
BTS* [26]	✓	✓	✗	✗
ARK* [24]	✓	✓	✗	✗
SHARP* [23]	✓	✓	✗	✗
FAST-I* [14]	✓	✗	✗	✗
FAB† [3]	✗	✗	✓	✗
Poseidon† [45]	✗	✓	✗	✗
EFFACT†* [20]	✗	✓	✗	✗
FAST-II† [42]	✗	✗	✓	✓
OLA† [43]	✗	✗	✓	✓
HERA†	✓	✓	✓	✓

Note: Both [14] and [42] are named FAST, and we use I and II to distinguish between them.
*ASIC-based accelerator; †FPGA-based accelerator

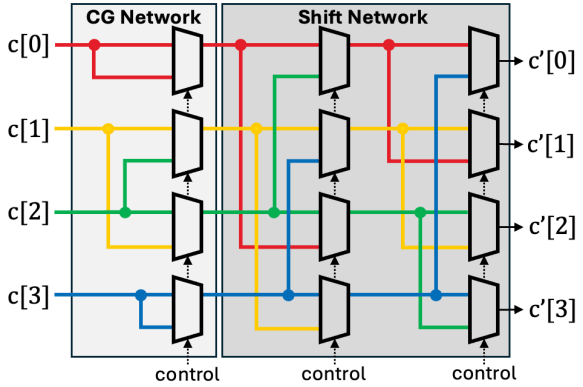
First FPGA-based accelerator with both NTT & Auto decomposition

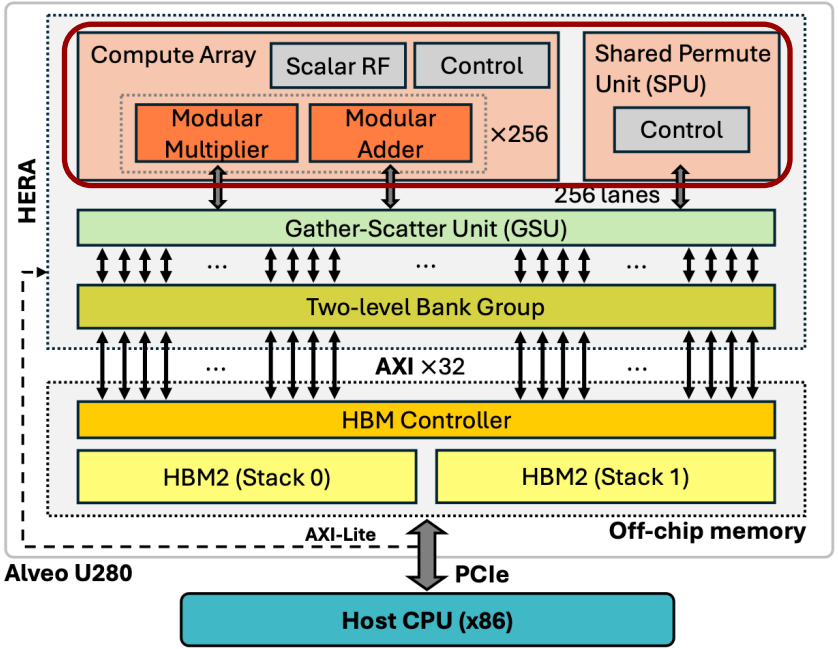
2D-NTT figure from: A. Feldmann, et al. "*F1: A Fast and Programmable Accelerator for Fully Homomorphic Encryption*", MICRO 2021

HERA Contribution (2)

Resource-shared architecture

- Compute Array
- Shared permutation circuit
  - ➢ one stage of CG
  - ➢ log(256) stages of shift network
    - ▪ any distance across lanes





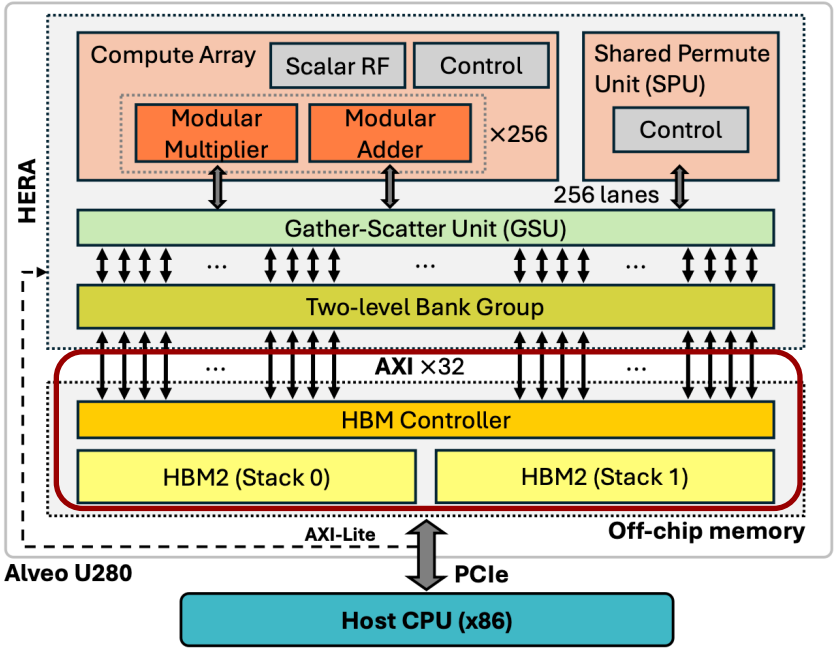
HERA Contribution (3)

Uniform data layout & access (1)

- Uniform data abstraction
- Block-wise mapping
  - ➢ Block size: 256 coefficients
  - ➢ each block: evenly across all PCs

- one PC ↔ one bank w. one dedicated AXI

Independent access of each PC, avoiding overheads due to crossbar



HBM floorplan figure from: K. Li, et al. "*ScalaBFS2: A High-performance BFS Accelerator on an HBM-enhanced FPGA Chip*", TRETS 2024

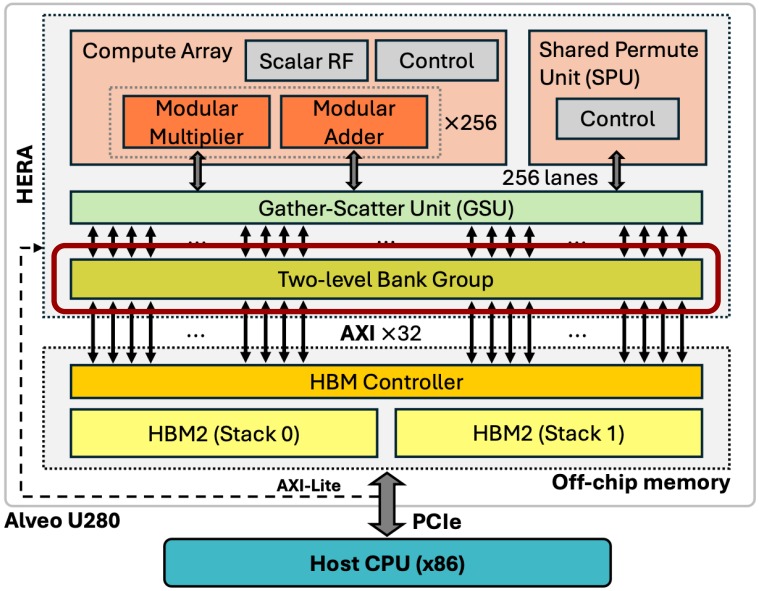
HERA Contribution (3)

Uniform data layout & access (2)

- Multi-bank on-chip scratchpad
  - ➢ 32 banks (per each PC)

- Data storage and access
  - ➢ Operand type [FAB, HPCA '23]
  - ➢ Poly. moduli [Medha, IACR TCHES '23]
  - ➢ **Uniform layout w. blocked access**

Maximize the potential of constrained SRAM

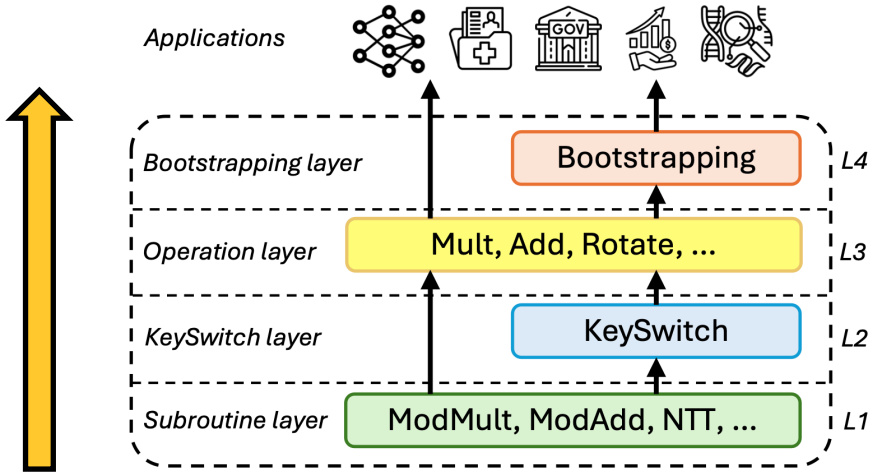


R. Agrawal, et al. *"FAB: An FPGA-based accelerator for bootstrappable fully homomorphic encryption"*, HPCA 2023
A. Mert, et al. *"Medha: Microcoded Hardware Accelerator for computing on Encrypted Data"*, IACR TCHES 2023

HERA Contribution (4)

Hierarchical Memory-centric Optimizations (HMOs)

Layer	Optimization(s)
Bootstrapping	⑥ Hoisting [2]
Operation	⑤ merge Rescale into ModDown during Mult [1]
KeySwitch	③ exhaustively fuse subroutines with single-limb dataflow ④ selective k-limb computation
Subroutine	① perform modulus switch on-chip ② fuse BConv and NTT through a single-limb dataflow



Maximize on-chip data reuse with constrained SRAM

[1] R. Agrawal, et al. *"FAB: An FPGA-based accelerator for bootstrappable fully homomorphic encryption"*, HPCA 2023
[2] M. Riaz, et al. *"HEAX: An architecture for computing on encrypted data"*, ASPLOS 2020

Outline

- Introduction
- Motivation
- HERA - Architecture and Contribution
- Evaluation
- Conclusion

Bandwidth and Resource

Alveo U280

- 300 MHz
- ~13 GB/s per AXI
- ~410 GB/s (90%)

AXI configurations

- 32 Rd/Wr FIFOs
- 512-bit port width
- burst length: 16
- # of outstanding access: 16

Resource utilization

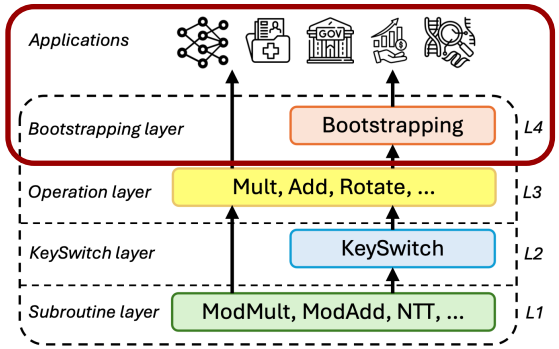
- > 80% memory
- Efficient DSP packing
  - ➢ 50-bit moduli

Resource	Available	Utilized	% Utilization
LUTs	1,304K	865K	66.4
FFs	2,607K	1,046K	40.1
DSP	9,024	4,352	48.2
BRAM	4,032	3,795	94.1
URAM	962	800	83.2

Results

Bootstrapping & Applications

- Bootstrapping
  - 8× speedup over the GPU
  - fastest vs. prior FPGAs $(1.20 - 4.13\times)$
  - benefits from reduced HBM traffic
- HELR
  - outperforms FAB, Poseidon, and EFFACT $(1.09 - 1.73\times)$
- ResNet-20
  - outperforms Poseidon, OLA, and EFFACT $(1.00 - 1.52\times)$



	Lattigo (CPU) [24, 31]	Over 100× (GPU) [21]	BTS (ASIC) [26]	ARK (ASIC) [24]	SHARP (ASIC) [23]	FAST-I (ASIC) [14]	FAB (FPGA) [3]	Poseidon (FPGA) [45]	EFFACT (FPGA) [20]	FAST-II (FPGA) [42]	OLA (FPGA) [43]	**HERA** (FPGA) -
Frequency (GHz)	3.4	1.2	1.0	1.0	1.0	1.0	0.3	0.45	0.3	0.3	0.3	0.3
# of lanes (dp)	-	-	2048	1024	1024	1024	256	512	256	512	512	256
# of multipliers	-	-	8192	20480	35776	48144	256	512	512	512	512	256
HBM bandwidth (GB/s)	-	900	1024	1024	1024	1024	460	460	460	460	460	460
SRAM capacity (MB)	43	36	512	588	198	281	42.2	8.6	7.6	33.8	22.5	35.6
T_{boot} (ms)	3.90e4	328	22.9	3.52	3.12	1.38	92.4	127	-	48.1	166	**40.2**
$T_{A.S.}$ (ns)	88.0K	716	45.7	14.3	12.8	5.40	477	-	566	81.0	-	**176**
HELR (ms)	37.1K	775	28.4	7.42	2.53	1.33	103	73.0	64.6	36.0	-	**59.4**
ResNet-20 (s)	1377	-	2.020	0.125	0.099	0.060	-	2.661	2.175	1.859	3.300	**2.170**

R. Agrawal, et al. *"FAB: An FPGA-based accelerator for bootstrappable fully homomorphic encryption"*, HPCA 2023
Y. Yang, et al. "*Poseidon: Practical Homomorphic Encryption Accelerator*", HPCA 2023
Y. Huang, et al. *"EFFACT: A Highly Efficient Full-Stack FHE Acceleration Platform"*, HPCA 2025
Y. Yang, et al. *"OLA: An FPGA-based Overlay Accelerator for Privacy Preserving Machine Learning with Homomorphic Encryption"*, ISFPGA 2025

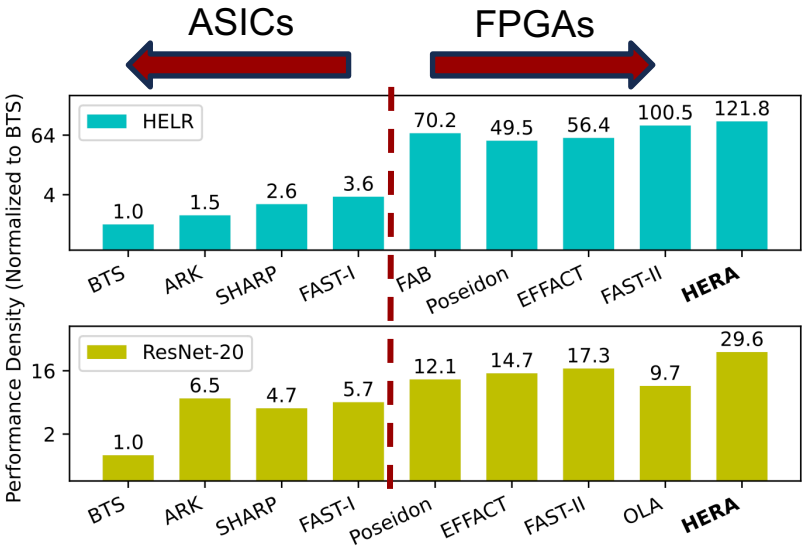
Results

Performance density

- Throughput per modular multiplier (# lanes)

- $1.71 - 3.05\times$ better than ASICs
- $1.21 - 2.46\times$ better than FPGAs

FPGA designs generally have better resource efficiency vs. ASICs.



Conclusion

- HBM - enabled FPGA is a practical platform for FHE acceleration
- Bandwidth usage and efficiency is essential
- Resource - shared architecture is suitable for FPGA with constrained resources
- Holistic memory-centric optimizations are necessary for FPGA-based acceleration
- May not meet real-time requirements (~2s per CIFAR-10 classification)
- Scalability to multi-FPGA system research is undergoing

USC University of Southern California

DEVCOM

Fun fact: **HERA** is also short for
Homomorphic **E**ncryption on
Reconfigurable **A**rray

Thank you

Q & A

{zhihanxu, prasanna}@usc.edu
fpga.usc.edu

USC