GME

GPU-based Microarchitectural Extensions to Accelerate Homomorphic Encryption

Kaustubh Shivdikar[1], Yuhui Bao[1], Rashmi Agrawal[2], Michael Shen[1], Gilbert Jonatan[3], Evelio Mora[4], Alexander Ingare[1], Neal Livesay[1], José L. Abellán[5], John Kim[3], Ajay Joshi[2], David Kaeli[1]

1 Northeastern University

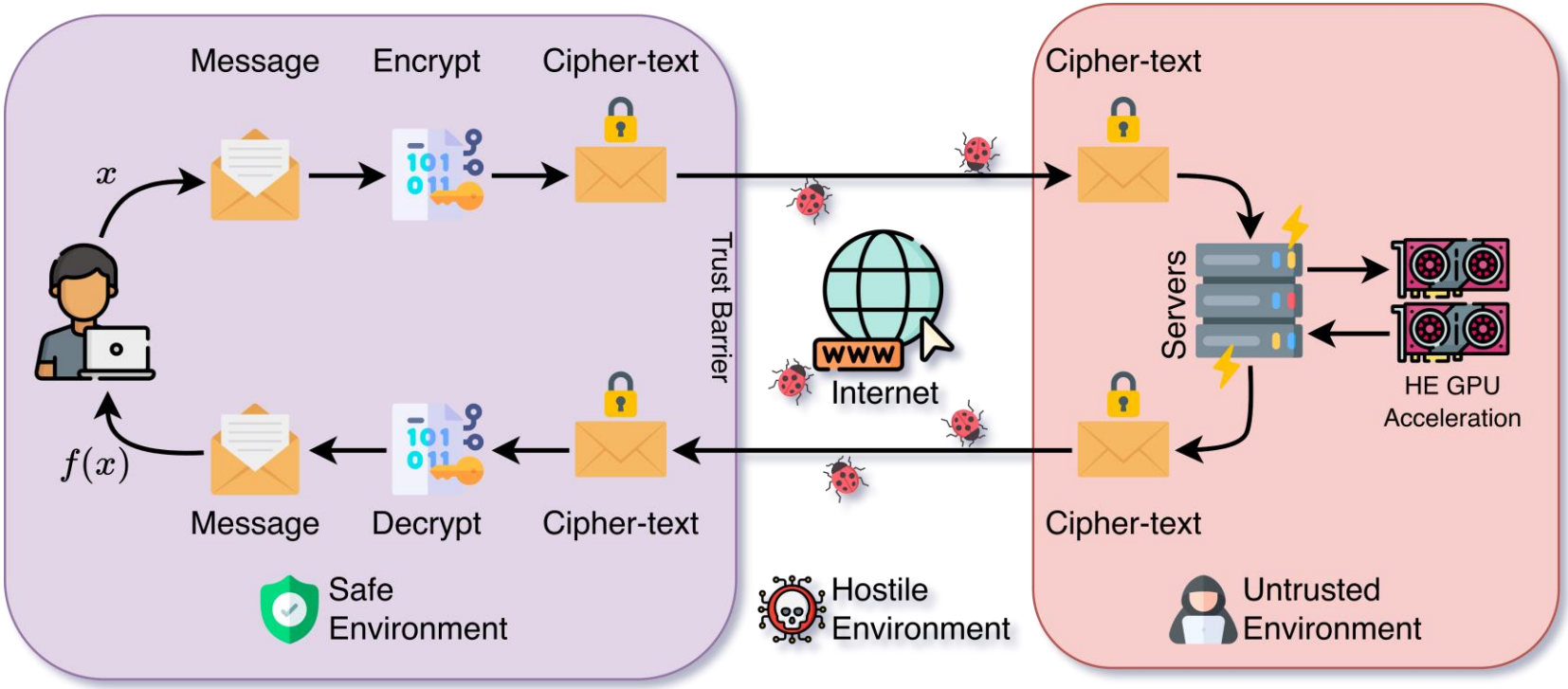
2 Boston University

3 KAIST

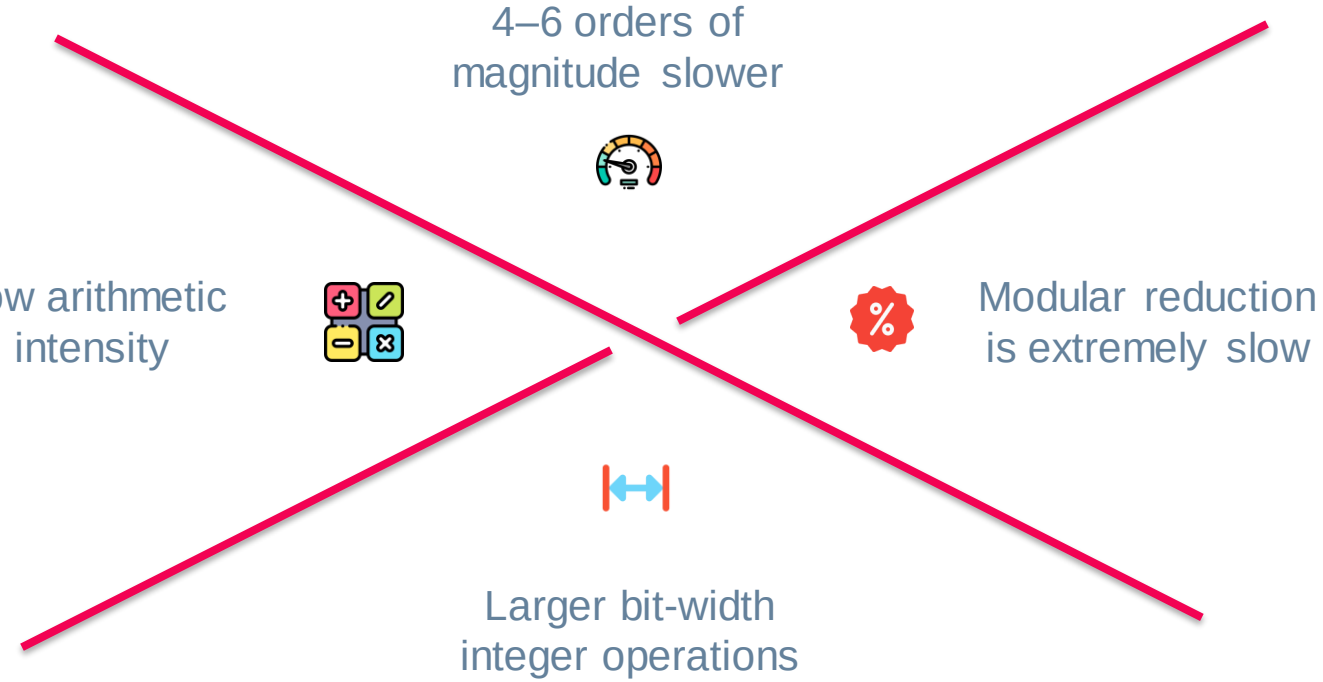
4 UCAM Universidad Católica de Murcia

5 Universidad de Murcia

What is Homomorphic Encryption?



Problems with Homomorphic Encryption



Key Contributions

Introduce a new CU-side interconnect

- Mitigates redundant data transfers

Native modular reduction support

- Modulus operation is used extensively in HE kernels

Wider multiply-accumulate units

- Native support for 64-bit integer arithmetic

Locality Aware Block Scheduler

- Scaling out to multiple HE kernels

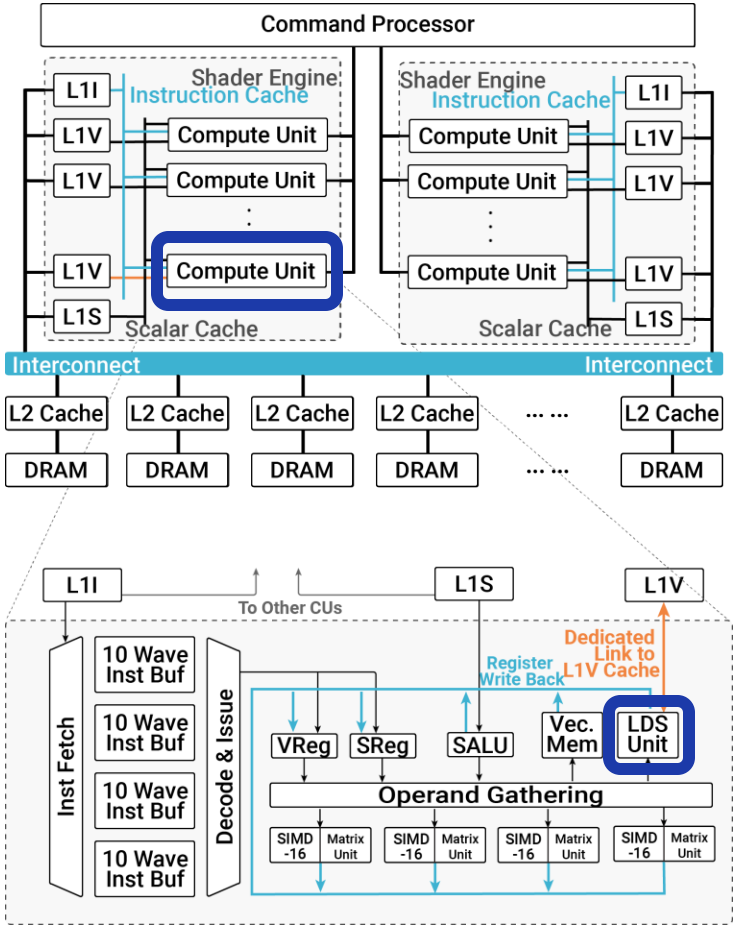
AMD CDNA Architecture

Choice of hardware platform: MI100

- Capitalize on well-established GPU ecosystems
- In-house CDNA architecture simulator

  NaviSim [1] (Open sourced)

Parameter	Value
Max Frequency	1502 MHz
Peak Performance	23.07 TFLOPs
Register File Size	15 MB
CU Count	120
L1 Vector Cache	16 KB per CU
Local Data Share	7.5 MB
Peak Bandwidth	1229 GB/s



Architecture of CDNA GPU

[1] Bao, Yuhui, et al. "NaviSim: A Highly Accurate GPU Simulator for AMD RDNA GPUs." Proceedings of the International Conference on Parallel Architectures and Compilation Techniques. PACT 2022.

CKKS HE Kernels

Kernels	Description	Working dataset size
PolyAdd	Add a plaintext to a ciphertext	14 MB
PolyMult	Multiplying a plaintext with a ciphertext	14 MB
HEAdd	Add two ciphertexts	28.3 MB
HEMult	Multiply two ciphertexts	127.3 MB
HERotate	Circular rotate elements by r slots	127.3 MB
HERescale	Restore the scale of a ciphertext	42.3 MB

MI100 GPU architectural parameters

- L1 Cache: 16 KB per CU
- Local Data Share: 7.5 MB
- Shared L2 Cache: 8 MB

$$Arithmetic\ Intensity = \frac{Number\ of\ Integer\ Operations}{Amount\ of\ Data\ Transferred\ (Bytes)}$$

All HE operations have **Arithmetic Intensity** < 1

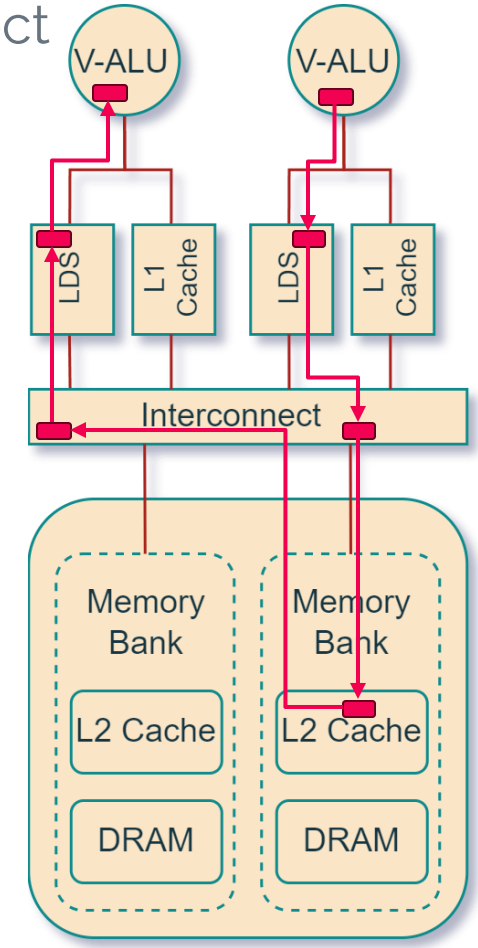
Compute Unit-side Interconnect
cNoC

Existing GPU on-chip network

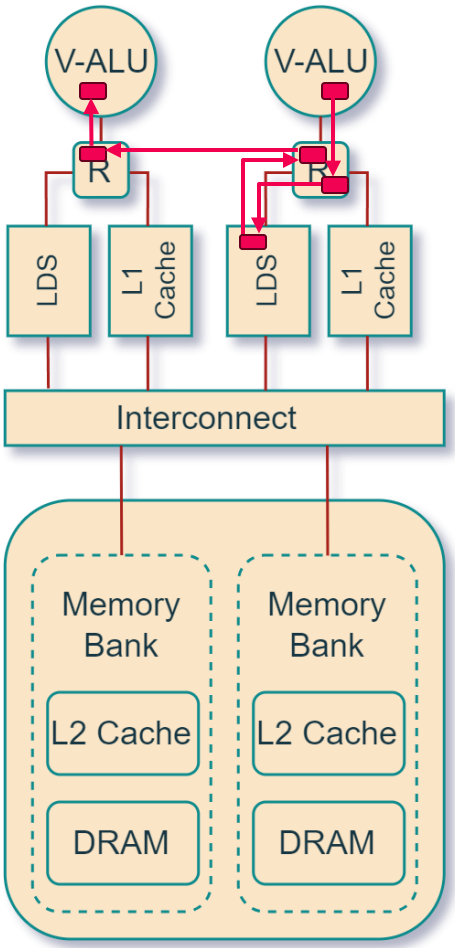
- Sharing data between CUs requires data to traverse the **entire memory stack**

CU-side network on chip

- Allows **inter-CU communication**

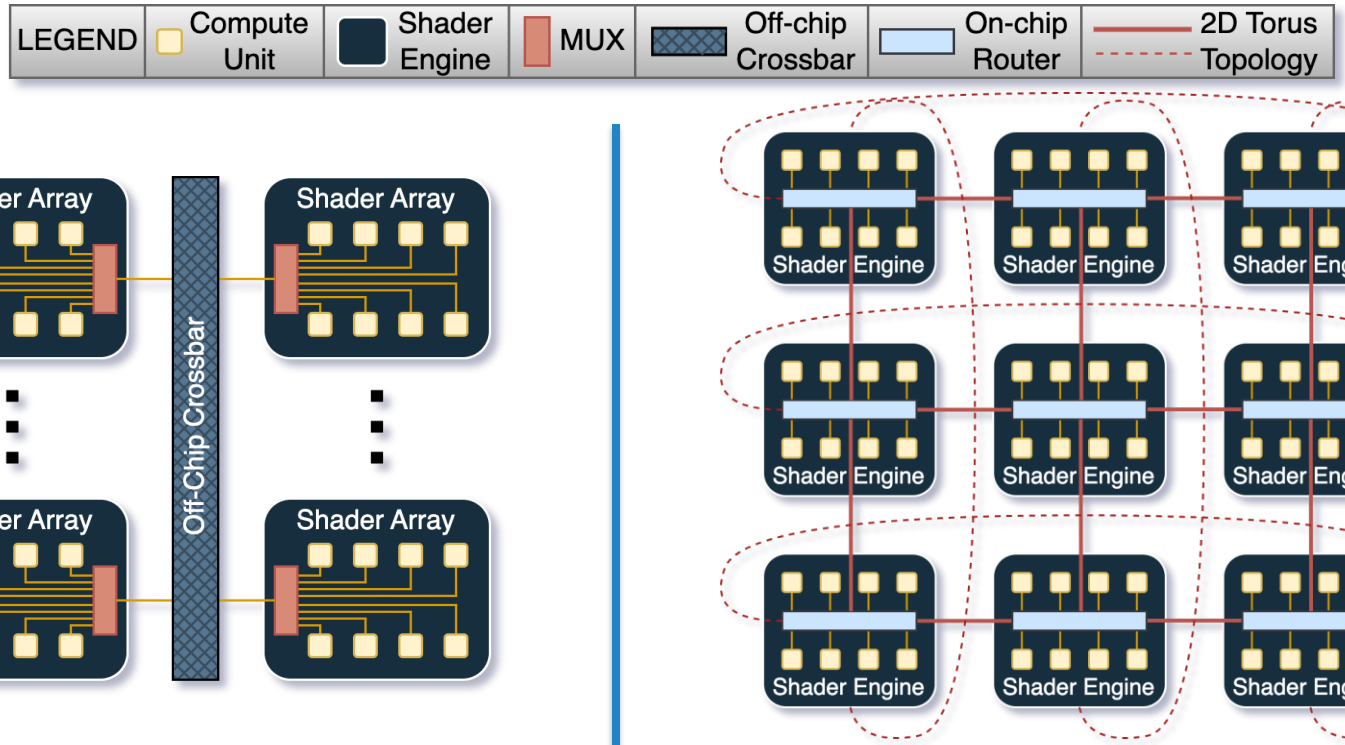


Existing GPU on-chip network



CU-side network on chip

Enhancing the on-chip network



(a) Existing GPU on-chip network limits inter-CU communication

(b) Proposed 2D torus topology for enabling inter-CU communication

Hardware Support for Modular Reduction

MOD

- Each operation in HE is followed by a **modulo** operation
- Modulo computation involves expensive **division** operation
- **Barrett's reduction** replaces division with a set of bit-shift and multiplication operations

Architecture	mod-red (cycles)	mod-add (cycles)	mod-mul (cycles)
Vanilla MI100	46	62	63
GME - MOD	26	18	38

Goal: $rem = x \,\%\, q$

$$rem = x - \left(\left\lfloor \frac{x}{q} \right\rfloor \times q \right)$$

Precomputation

$$m = len(q)$$

$$\mu = \left\lfloor \frac{2^{2m+1}}{q} \right\rfloor$$

Proposed Barrett's Reduction Algorithm[1]

$$c = x \gg (m - 2)$$
$$quot = (c \times \mu) \gg (m + 3)$$
$$rem = x - quot \times q$$

if $rem \geq q$ **then**

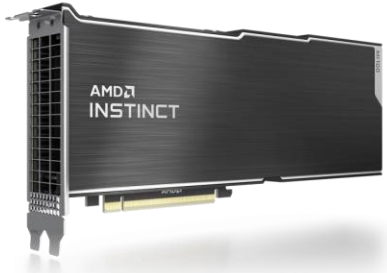
$$rem = rem - q$$

return rem

[1] Shivdikar et al., 2022. Accelerating polynomial multiplication for homomorphic encryption on GPUs. In 2022 IEEE International Symposium on Secure and Private Execution Environment Design (SEED)

Experimental Platforms

AMD CDNA **MI100** GPU



NaviSim: Cycle-level GPU simulator

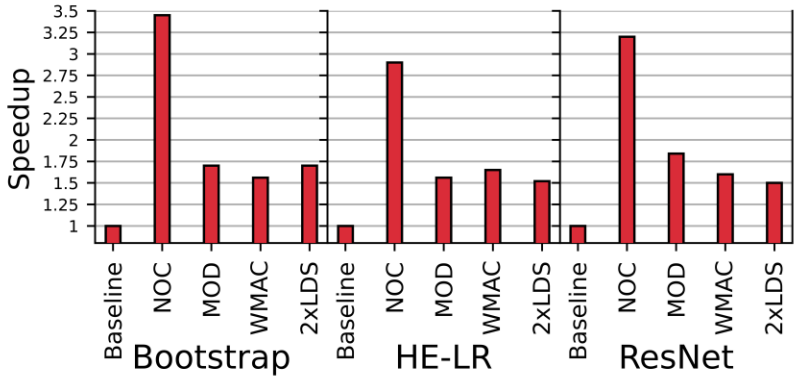
NaviSim

BlockSim: Extend NaviSim with multi-block support

Sim

Baseline → GME → Scaling Out

Performance Evaluation



Speedup achieved from each microarchitectural extension.

Accelerator	Arch.	$T_{A.S.}$ (ns)	Boot (ms)	HE-LR (ms)	ResNet-20 (ms)
Lattigo [1]	CPU	8.8e4	3.9e4	23293	-
HyPHEN [2]	CPU	2110	2.1e4	-	3.7e4
F1 [3]	ASIC	2.6e5	Yes*	1024	-
BTS [4]	ASIC	45	58.9	28.4	1910
CraterLake [5]	ASIC	17	4.5	15.2	321
ARK [6]	ASIC	4	3.7	7.42	125
FAB [7]	FPGA	470	92.4	103	-
100x [8]	V100	740	528	775	-
HyPHEN [2]	V100	-	830	-	1400
T-FHE [9]	A100	404	157	178	3793
Baseline	MI100	863	413	658	9989
GME	MI100+	**74.5**	**33.63**	**54.5**	**982**

Speedup over CPU: 796x GPU: 14.2x FPGA: 2.3x

[1] Vincent et al., **Lattigo v4**.
[2] Park et al., **HyPHEN**: A Hybrid Packing Method and Optimizations for Homomorphic Encryption-Based Neural Network.
[3] Samardzic et al., **F1**: A fast and programmable accelerator for fully homomorphic encryption.
[4] Kim et al., **BTS**: An accelerator for bootstrappable fully homomorphic encryption
[5] Samardzic et al., **Craterlake**: a hardware accelerator for efficient unbounded computation on encrypted data.
[6] Kim et al., **ARK**: Fully homomorphic encryption accelerator with runtime data generation and inter-operation key reuse.
[7] Agrawal et al., **FAB**: An FPGA-based accelerator for bootstrappable fully homomorphic encryption. HPCA 2023
[8] Jung et al., Over **100x** faster bootstrapping in fully homomorphic encryption through memory-centric optimization with GPUs.
[9] Fan et al., **Tensorfhe**: Achieving practical computation on encrypted data using gpgpu.

Impact of μ-arch extensions

cNoC

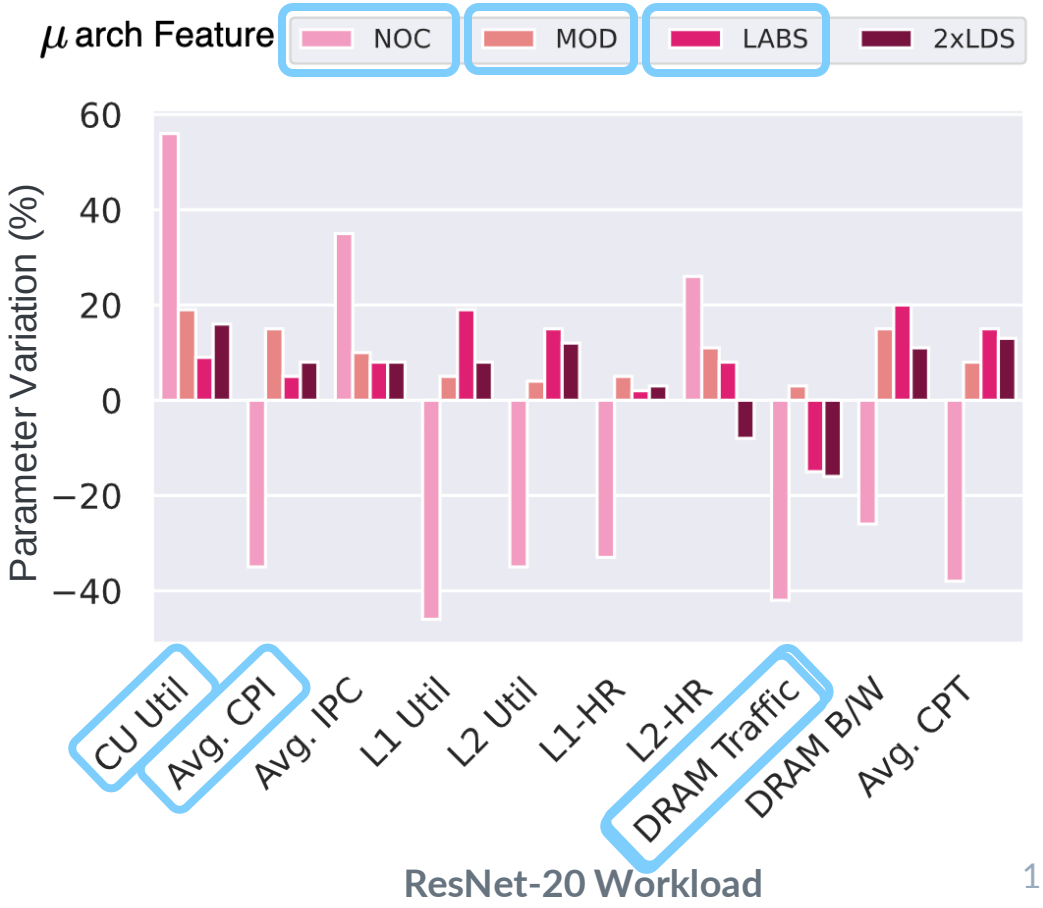
- Increases CU-Utilization
- Decreases DRAM traffic

MOD + WMAC

- MOD introduces complex instructions
- Increases avg. CPI

LABS

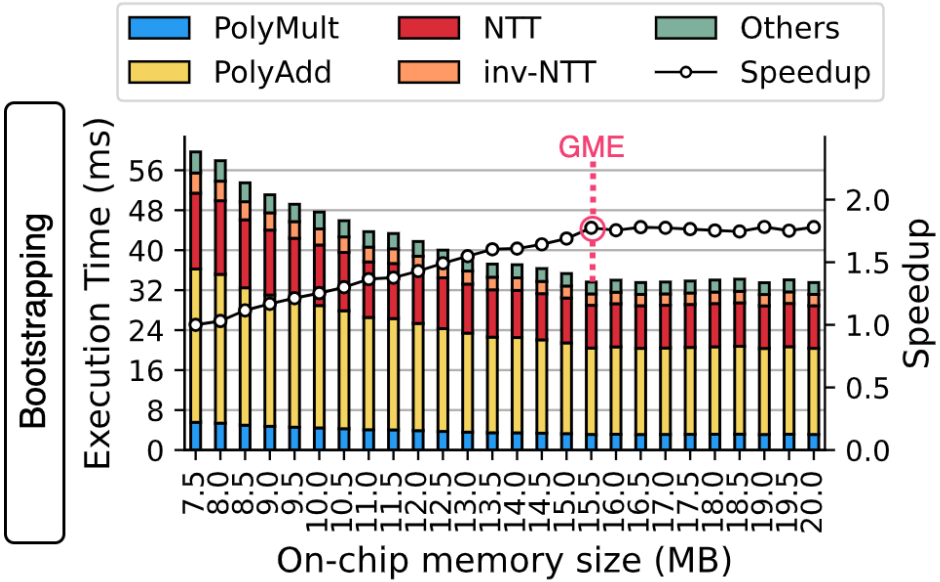
- Shared blocks are scheduled together reducing DRAM Traffic
- Avg. speedup of 1.5X



On-chip Memory Size Exploration
2xLDS

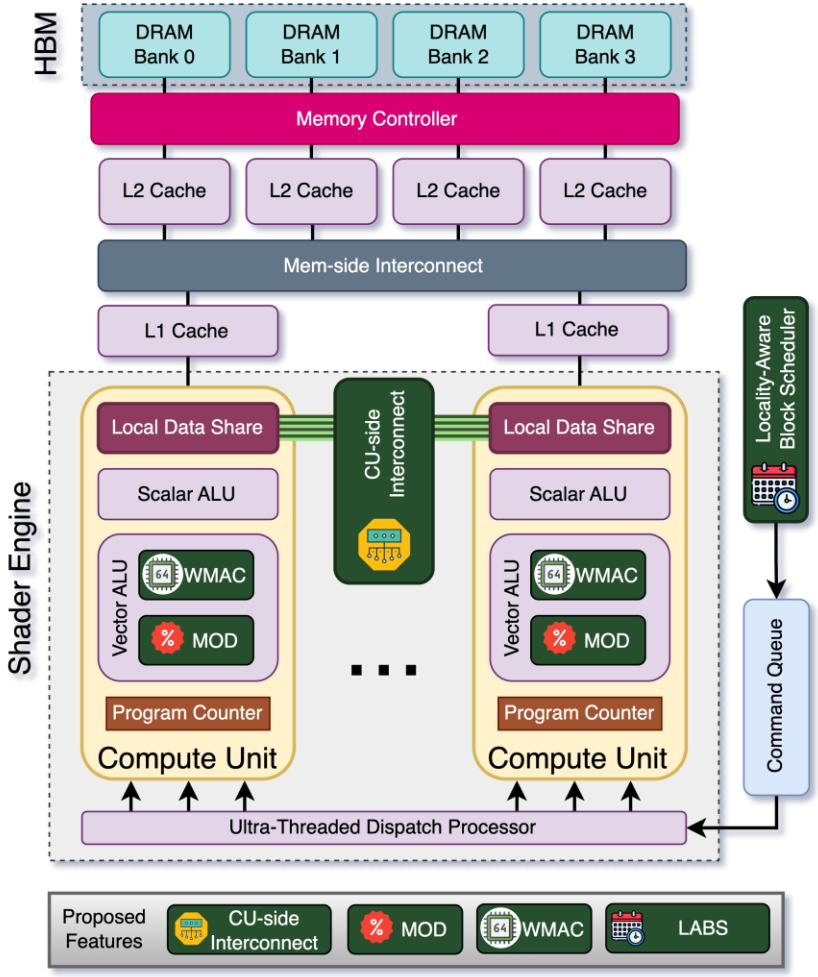
- Increasing the LDS size from **7.5MB** to **15.5MB** produces significant speedup

Workload	Speedup
Bootstrapping	1.74x
HE – LR	1.53x
ResNet – 20	1.51x



Evaluated μ-Arch Extensions

- **cNoC**: 2D Torus on-chip network
- **MOD**: Natively supported modular-reduction
- **WMAC**: 64-bit integer pipeline
- **LABS**: Sharing data across blocks



Concluding Remarks

HE has emerged as the “holy grail” of data privacy against rapidly evolving threats in the **quantum era**

Identified key **bottlenecks** in HE

- Redundant data transfers
- Expensive modular reduction operations

Presented four µ-Arch Extensions

GME speedup over

- CPU: **796x**
- GPU: **14.2x**
- FPGA: **2.3x**

Future Work

- Incorporate **Processing-in-Memory** techniques
- Integrate with **open-source** platforms

OpenFHE

Thank you!

Any questions?

Paper and Slides: wiki.kaustubh.us

Support by:
- NSF IUCRC Center for Hardware and Embedded Systems Security and Trust (CHEST) NSF CNS Award #2312275
- NSF CNS Award #2312276
- Samsung Advanced Institute of Technology, Samsung Electronics Co., Ltd
- Grant RYC2021-031966-I funded by MCIN/AEI/10.13039/501100011033
- “European Union NextGenerationEU/PRTR.”