Sub Millisecond TFHE Bootstrapping on GPU

P. Alves, B. Barbakadze, E. Di Maria, J. Klemsa,
A. Leroy, G. Oyarzun

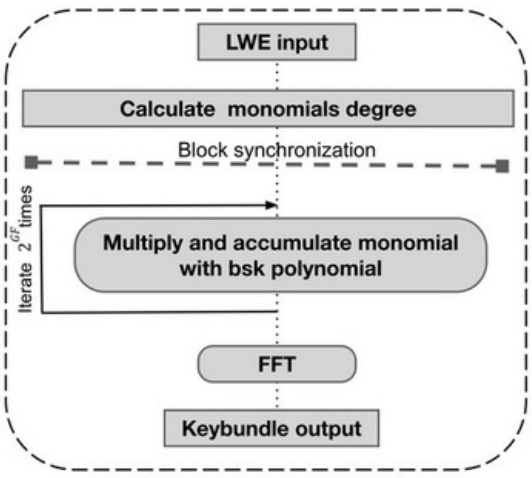
Key contributions

Sub-millisecond bootstrapping on GPU, achieved on Hopper or more recent architectures by:

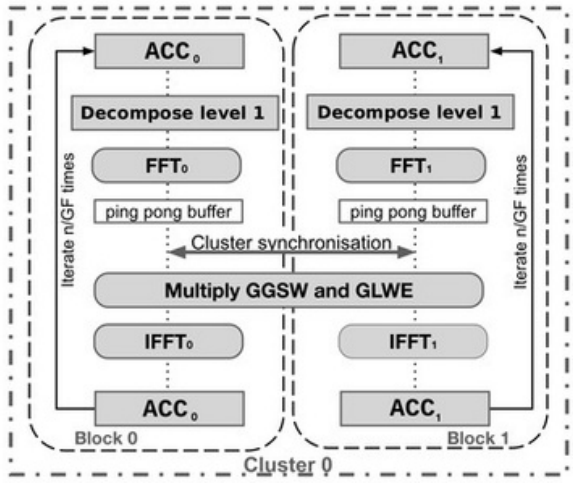
- Using a variant of the Bootstrapping algorithm, the **Multi-Bit Bootstrapping**, which exposes more parallelism
- Coordinating thread blocks using the **Thread Block Clusters** feature of Cuda
- **Maximizing the use of registers** and **minimizing the number of synchronizations** needed in the FastFourier Transform and inthe Bootstrapping itself

The architecture on GPU

Keybundle



Blind rotation



Cryptographic parameters for Classical Bootstrapping

Message size	n	N	k	ℓ	β
Booleans	837	512	4	1	23
4-bit Integers	866	2048	1	1	23

Cryptographic parameters for Multi-Bit Bootstrapping with grouping factor 4

Message size	n	N	k	ℓ	β
Booleans	736	2048	1	1	22
4-bit Integers	872	2048	1	1	22

Bootstrapping benchmark results

Comparison of Bootstrapping performance between CPU and GPU in TFHE-rs, and with VeloFHE (CHES25)

		Booleans		4-bit Integers	
Implementation CPU (classical)	**Hardware**	Latency	Throughput	Latency	Throughput
GPU (classical)	AMD EPYC 192 cores	10.3 ms	11,566 ops/s	12.3 ms	8,477 ops/s
GPU (multi-bit) (present work)	1xH100	3.1 ms	-	3.8 ms	-
GPU (multi-bit) (present work)	1xH100	**811 µs**	28,620 ops/s	**962 µs**	24,170 ops/s
VeloFHE GPU GD-I/II (CHES25)	8xH100	-	**224,450 ops/s**	-	**190,000 ops/s**
GPU (multi-bit) GD-I/II	1xRTX4090	-	11,378 ops/s	-	3,249 ops/s
	1xRTX4090	-	10,059 ops/s	-	3,972 ops/s

Integer operations and AES transciphering

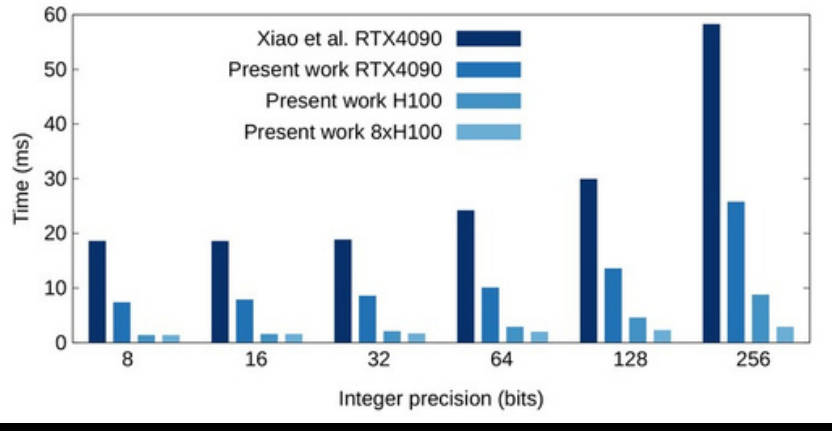
Latency of 64-bit integer operations on CPU (classical) and GPU (multi-bit) in TFHE-rs

Hardware	Add	Mul	Div	Greater than	Select
CPU AMD 192 cores	109.0 ms	402.0 ms	5800 ms	105 ms	47.8 ms
GPU 1xH100	11.4 ms	111.7 ms	946.5 ms	11.6 ms	7.6 ms
GPU 8xH100	**9 ms**	**31.9 ms**	**502 ms**	**10.6 ms**	**4.6 ms**

Comparison with FPGA-based accelerator (hello-fpga [Bel24]) on AES-128 transciphering of two inputs

Platform	Hardware	Latency
TFHE-rs CPU	AMD RYZEN 192 cores	20.9 s
`hello-fpga`	VU47P FPGA	6.6 s
TFHE-rs CUDA backend	H100	**1.7 s**

Comparison of bitand latency (ms) with Xiao et al. [XLK+ 24]



Resources
zama.org/blog
github.com/zama-ai