KU LEUVEN

COSIC



Hardware acceleration for FHE

Ingrid Verbauwhede
KU Leuven, COSIC

Leuven, June 2023

erc

Credits: Michiel Van Beirendonck, Furkan Turan, Robin Geelen, Sujoy Sinha Roy, Junfeng Fan, Fre Vercauteren and all PhD students

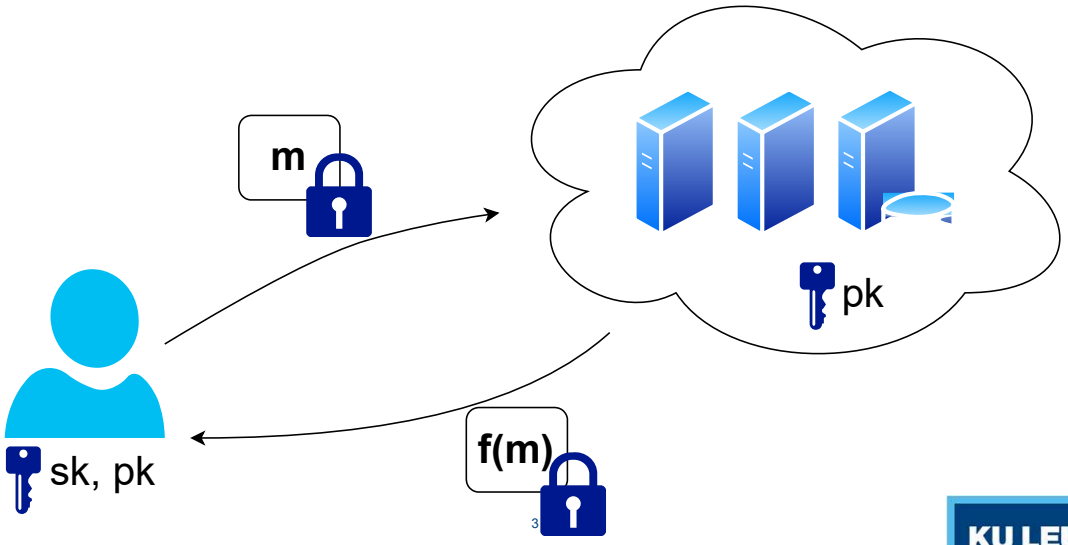
Outline

- Multiple schemes
- Multiple Hardware options…
- Try to give you an insight into the design space (zoo of options)
- Workshop in the context of the EU Twinning Project enCRYPTON

2

KU LEUVEN

Fully Homomorphic Encryption



3

KU LEUVEN

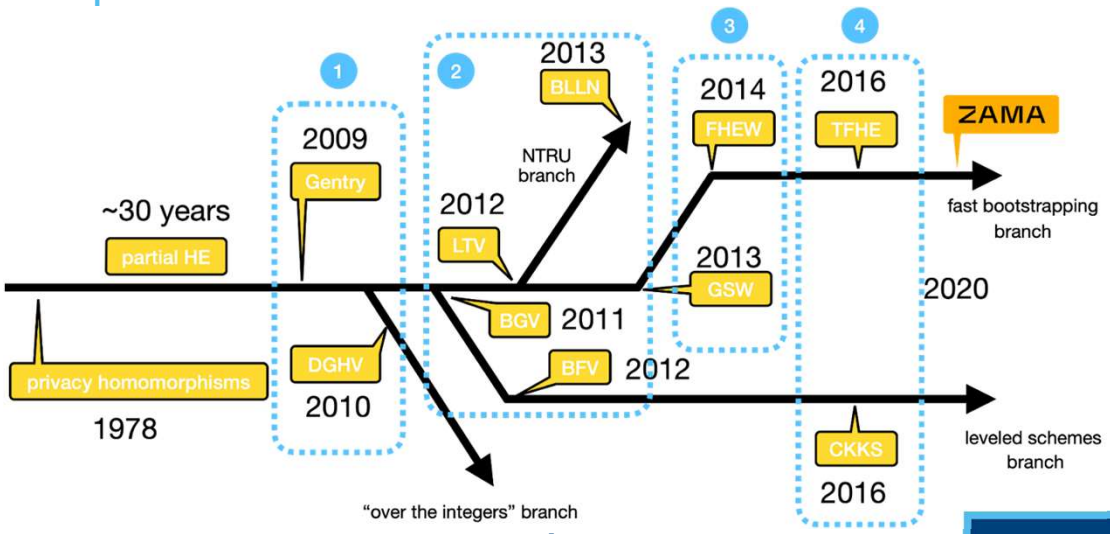
Multiple schemes

- Partially homomorphic: Paillier system
- Somewhat homomorphic:
  - Limited number of multiplications
  - Fan-Vercauteren:
- Fully Homomorphic Encryption
  - Unlimited number of multiplications
  - Requires 'bootstrapping'
- Multiple schemes:
  - BFV: Brakerski – Fan – Vercauteren
  - BGV: Brakerski – Gentry – Vaikuntanathan
  - TFHE: Torus Fully Homomorphic Encryption
  - …

4

KU LEUVEN

Multiple FHE schemes



[copied from ZAMA website]

5

KU LEUVEN

Challenge large numbers:

- Experiment 1 [CHES2015] : YASHE (now no longer used, reduced security)
  - Ciphertext size 5MB to 20MB (Polynomial size is 32768 (2^15) to 65536 (2^16), modulus 1200 to 2500 bits), could evaluate depth of Simon block cipher
- Experiment 2 [TC2018]: HEPCLOUD, FV
  - Ciphertext pair 9.2MB with parameters Polynomial size is 32768 (2^15), modulus 1128 bits, depth 36, 85 bits security level.
  - Bottleneck: I/O between FPGA and external memory
- Experiment 3 [TC2020]: HEAWS, FV
  - Cipher text pair 180KB, with parameters Polynomials size is 4096, modulus min 372 (Q), 180 (q), depth 4, more than 80 bits security.
  - Useful for small neural network applications
  - Fits on one FPGA

6

KU LEUVEN

DARPA DPRIVE program: in progress



- Dedicated ASIC acceleration of BGV
  - 150mm^2 chip in 12nm GF
  - Within 10× of plaintext computation
  - 10,000× faster than software reference
  - Parameter set for 128-bit security
    - Support *bootstrapping*
- Four teams of researchers
  - **Galois**, Duality, SRI, and Intel
- Several phases
  - **Phase 1**: design, implementation and verification of system architecture and IP blocks

Now: phase 2 running, with three teams: Galois, Duality and Intel
Now: phase 2 also includes CKKS

7

KU LEUVEN

BGV parameters in DPRIVE

Parameter	Range	Example
Security parameter	N/A	128 bits
Ring dimension N	$512 - 65536$	**65536**
Plaintext modulus p^r	$\geqslant 2$	127^3
Ciphertext packing ℓ	$1 - 65536$	64 slots
Max $\log_2(QP)$ for key switching	$20 - 1782$	**1782 bits**
Max $\log_2(Q)$ for ciphertext	$20 - 1263$	1263 bits
Max multiplicative depth L	N/A	31

Ciphertext: 21 MB, Key-switch key: 84 MB

8

KU LEUVEN

Hardware acceleration options

9

KU LEUVEN

Challenges

- Computational complexity
  - NTT/FFT acceleration
- **Memory**
  - SIZE
  - BANDWIDTH



ASIC (phase 1)

- 150 mm^2 in 12nm
- Global Foundaries
- Memory hierarchy
- 57 – 115 Watt

Cloud FPGA

- Alveo U280 (in 5nm or 7 nm)
- Included into Amazon AWS F1
- Memory hierarchy
- 225 Watt! (cooling)

10

KU LEUVEN

Three experiments – three domain specific processors

FPGA - HEAWS

- BFV – leveled HE
- 80 bit security
- Shallow depth

ASIC – DPRIVE – BASALISC

- BGV – includes Bootstrap
- 128 bit security
- DPRIVE constraints

FPGA - FPT

- TFHE
- 128/110 bit security
- Alveo U280

- NTT acceleration
- Residue Number System
- Dedicated instruction set
- No cache: compile time known

- FFT acceleration
- Streaming bootstrap

IEEE TC 2020

IACR 2022/657

IACR 2022/1635

11

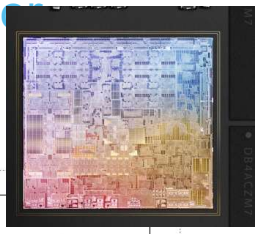
KU LEUVEN

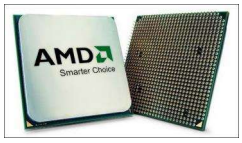
Option 1: CPU - **General** purpose processor

Programmable
Standardized instructions set
Software design

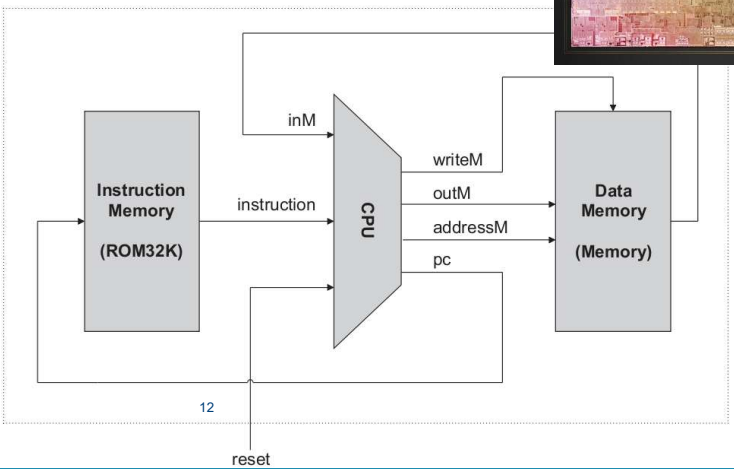


Apple M2







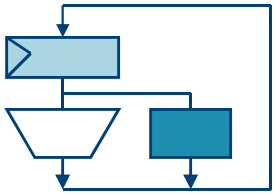


12

UVEN

Option 2: **Domain** specific processing

- Tightly couple: instruction set extension
  - Register mapped
  - *Reuse* CPU infrastructure
    - Reuse decode, registers, cache, bus network, etc.
  - Example AES instructions



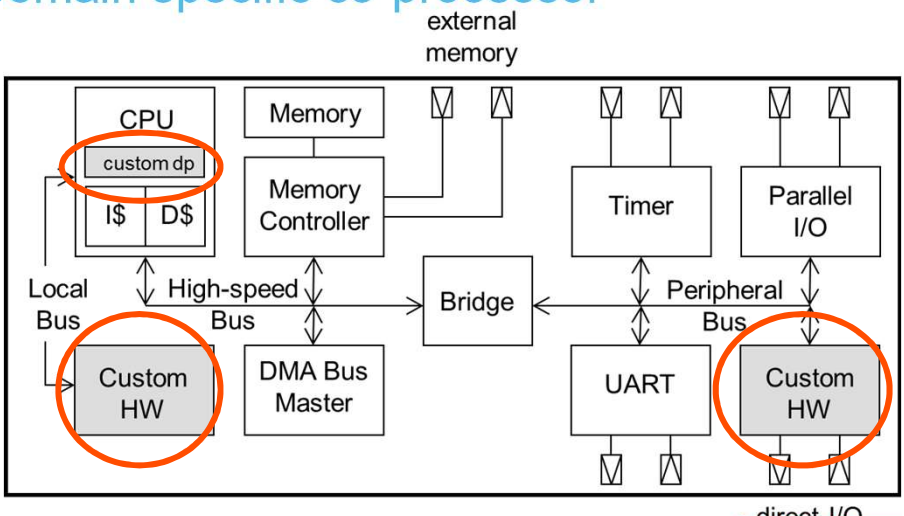
- Here: FHE specific operations
  - Leads to DOMAIN SPECIFIC PROCESSORS

13

KU LEUVEN

Option 3: Domain specific co-processor

- Inside CPU
  = custom ISA
- Local bus
  = tight coupled
- Peripheral
  = loosely coupled



[Picture: P. Schaumont, "A practical introduction to Hardware/Software Codesign", 2nd ed

14

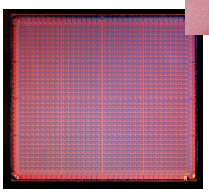
KU LEUVEN

Hardware technology: FPGA versus ASIC

FPGA

- Field Programmable Gate Array



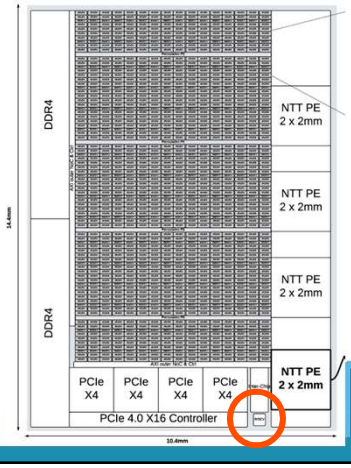




Xilinx XC4000ex (OLD!)

ASIC

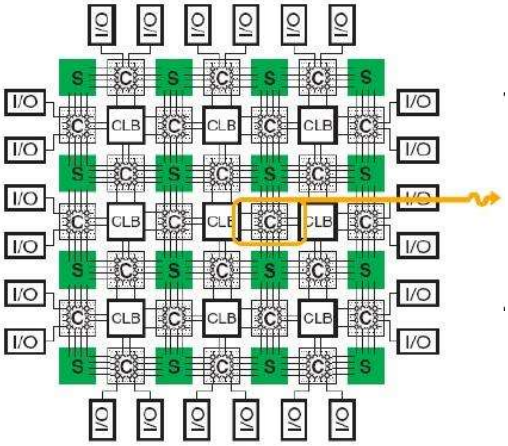
- Application Specific Integrated Circuit

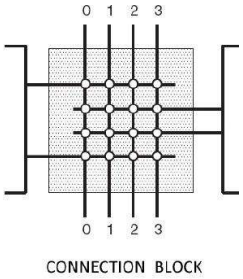


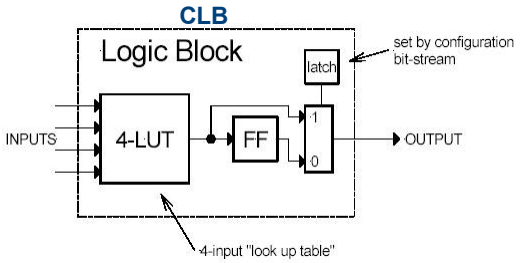
15

KU LEUVEN

FPGA: Program look-up tables and interconnect







Buy FPGA, write VHDL or Verilog
Synthesis, place and route
load onto FPGA (in the cloud)

16

First experiment: FPGA Acceleration of BFV on Amazon cloud

17

KU LEUVEN

FPGA Memory Resources (Alveo U280)



- **BRAM – 9 MB**
- **URAM – 33 MB**

LOCAL ON CHIP

- **HBM – 8 GB**

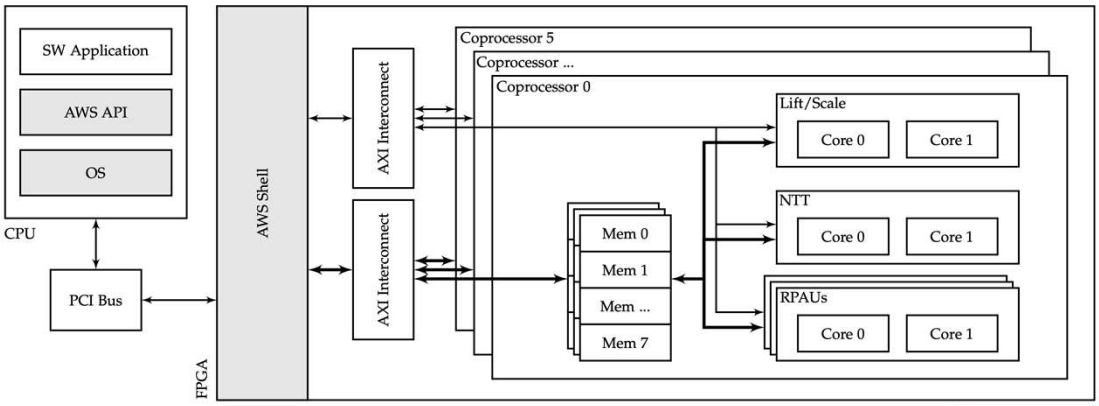
IN PACKAGE, 3D

- **DDR – 32 GB**

ON BOARD

KU LEUVEN

Amazon F1 Instance Implementation [TC20]



KU LEUVEN

Performance of Homomorphic Multiplication

- Each multiplication takes 4.34 ms.
- The overhead of a ciphertext transfer is 0.11 ms.

- A single coprocessor achieves 230 multiplications per second.
- **Six** coprocessors running in parallel achieves 613 multiplications.

KU LEUVEN

Comparison

- Achieve 613 homomorphic multiplications per second
- Compared to CPU
  - 13x speedup w.r.t. a highly optimized software on Intel i5 processor, 1.8 GHz
- To GPU on Amazon cloud, **5 times more work for half price and lower power!**

Compute: Amazon EC2 Instances:

	Description	Instances	Usage	Type	Billing Option	Monthly Cost
⛔	1 FPGA -> 2000 Mult	1	100 % Utilized/Mo	Linux on f1.2xlarge	On-Demand (No Co	$ 1207.80
⛔	1 GPU -> 388 Mult	1	100 % Utilized/Mo	Linux on p3.2xlarge	On-Demand (No Co	$ 2239.92

14.02.2019 - AWS Simple Monthly Calculator: https://calculator.s3.amazonaws.com/index.html
17.06.2020

21

KU LEUVEN



Second experiment: ASIC Acceleration of BGV

Darpa DPRIVE Basalisc project

22

KU LEUVEN

BASALISC Memory Hierarchy



MAC Acc – 8 KB

MAC Register File – 128 KB

Cipher Text Buffer – 64 MB

DDR – 256 GB

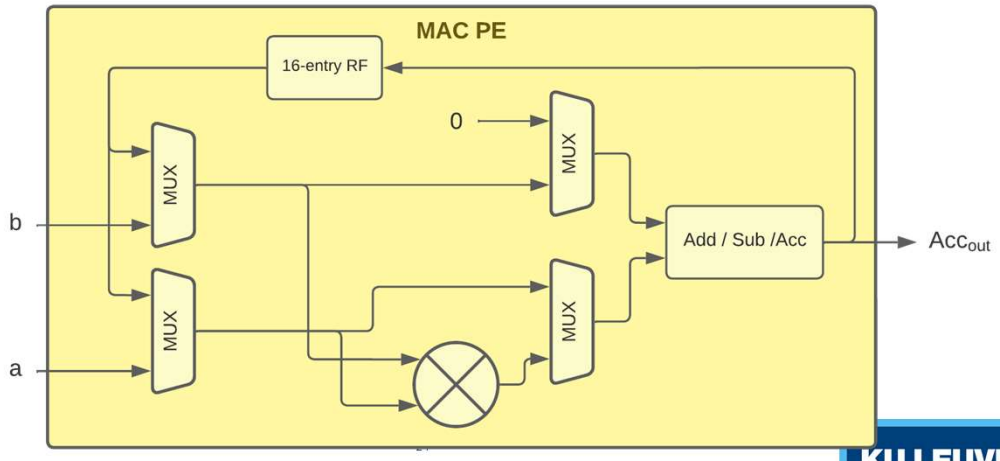
ON CHIP, Cipher Text Buffer
CTB fits 3 ciphertext pairs

One Key switch 84MB does not fit

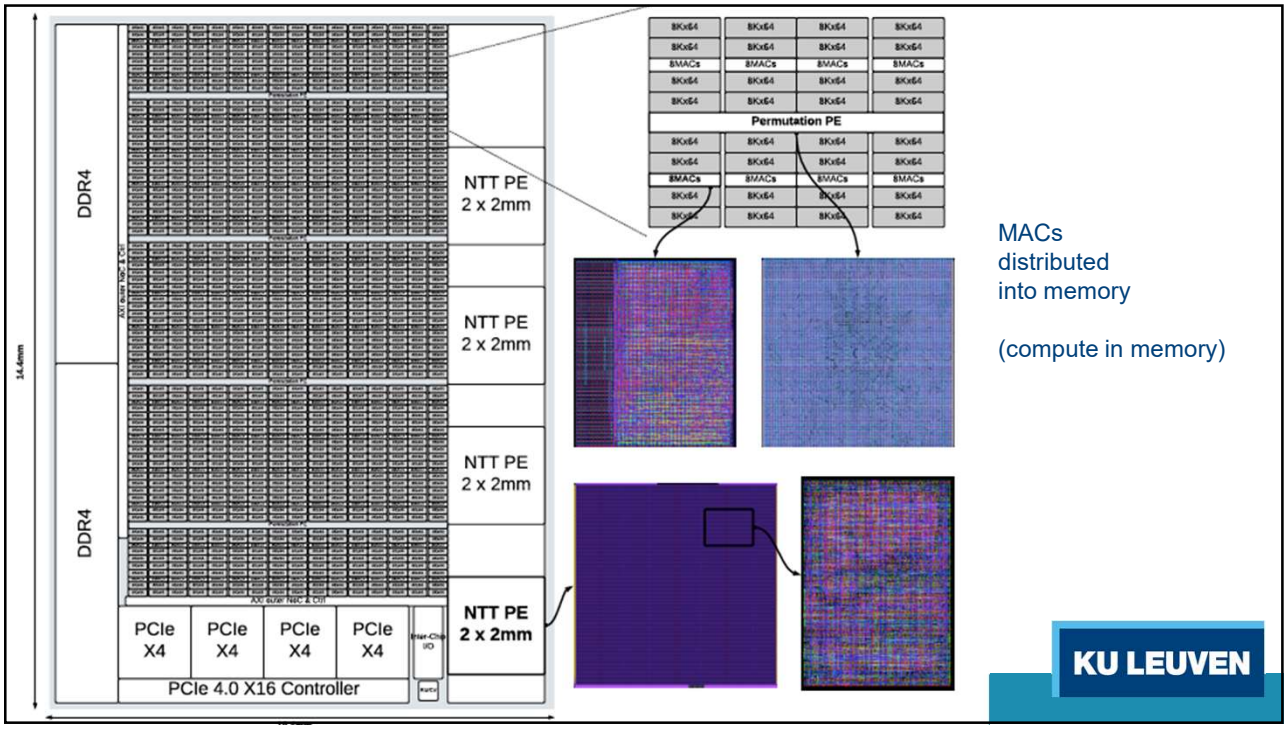
ON BOARD

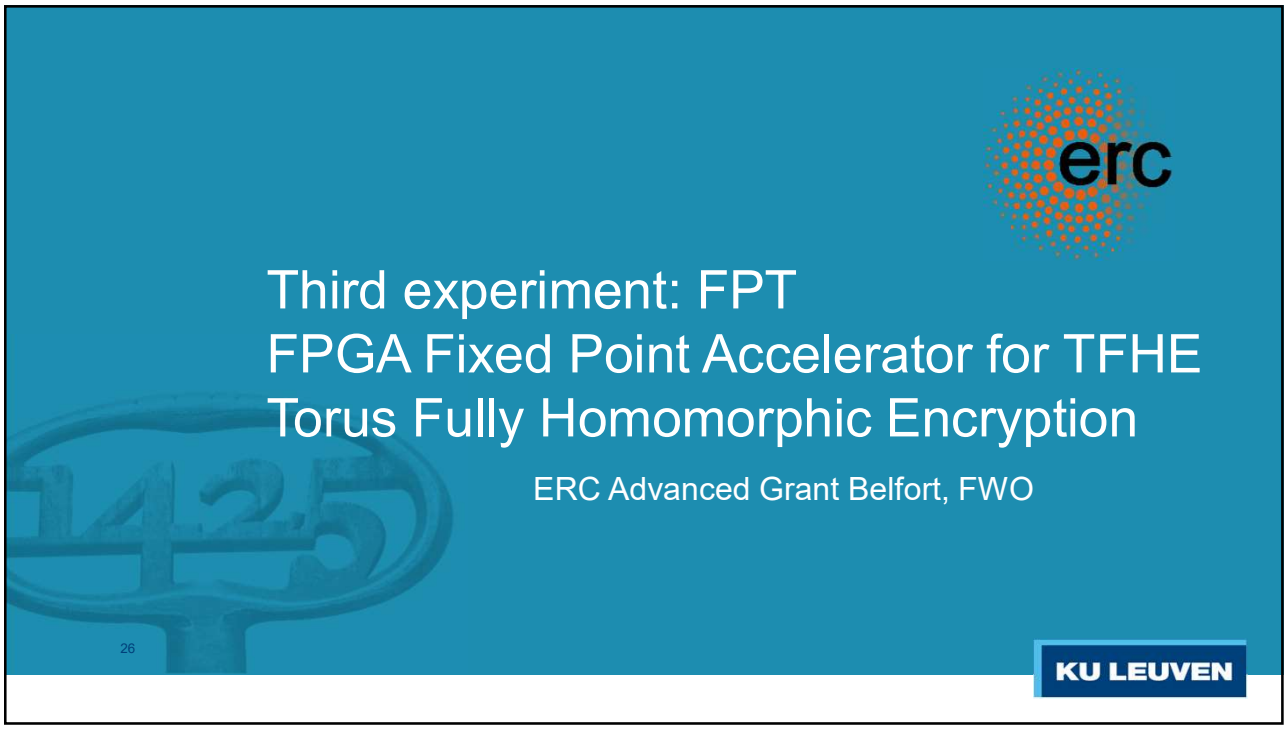
KU LEUVEN

2048 x 32 bits Modular Multiply – Accumulate unit PE

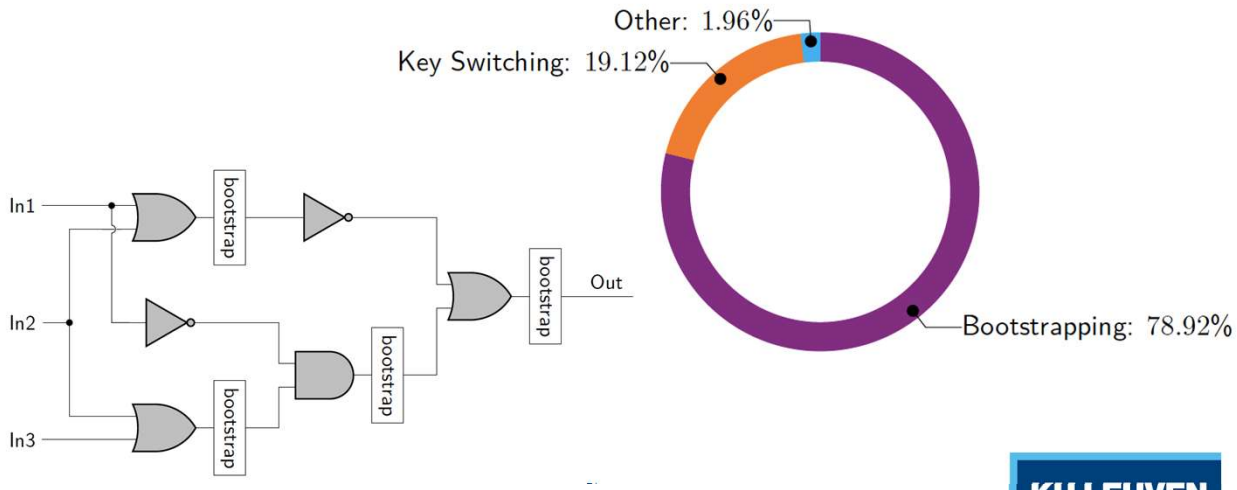


KU LEUVEN


DDR4
DDR4
NTT PE
2 x 2mm
NTT PE
2 x 2mm
NTT PE
2 x 2mm
NTT PE
2 x 2mm
PCIe X4
PCIe X4
PCIe X4
PCIe X4
PCIe 4.0 X16 Controller
Permutation PE
MACs
distributed
into memory
(compute in memory)
KU LEUVEN


erc
Third experiment: FPT
FPGA Fixed Point Accelerator for TFHE
Torus Fully Homomorphic Encryption
ERC Advanced Grant Belfort, FWO
26
KU LEUVEN

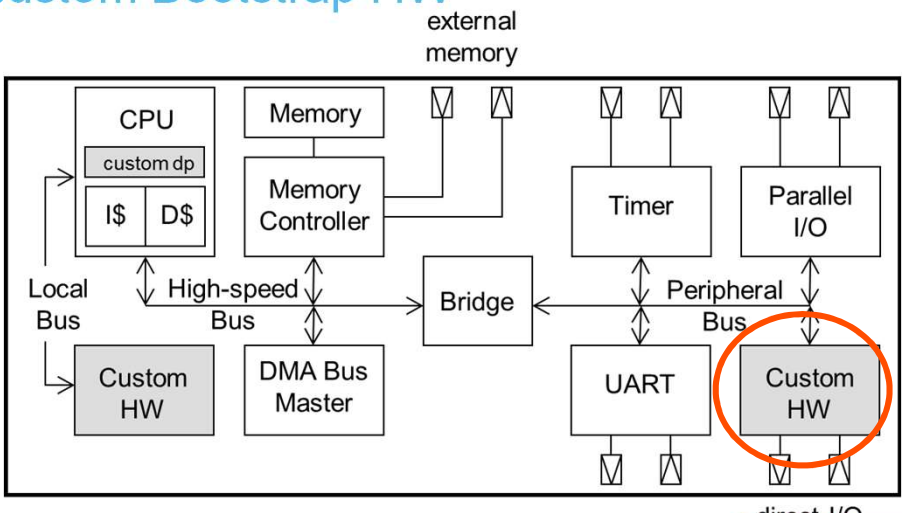
Challenge: Bootstrap acceleration



KU LEUVEN

Option 3: Custom Bootstrap HW

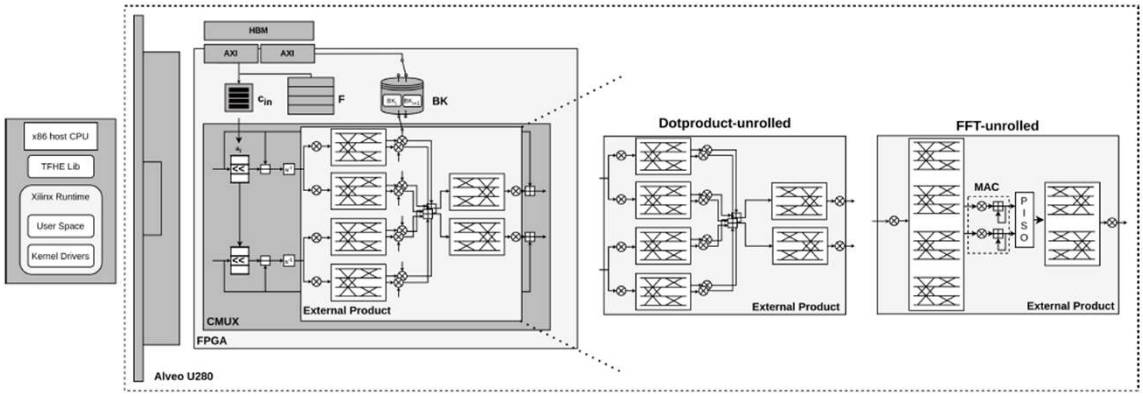
- Peripheral
  = loosely coupled



[Picture: P. Schaumont, "A practical introduction to Hardware/Software Codesign", 2nd ed

KU LEUVEN

FPGA: Bootstrap FFT accelerator



29

KU LEUVEN

Results

- FPGA
- ASIC
- CPU
- GPU

	LUT / FFs / DSP / BRAM	f (MHz)	l (ms)	TP (PBS/ms)
FPT	595K / 1024K / 5980 / 14.5Mb	200	0.58	**25.0**
YKP	842K / 662K / 7202 / 338Mb	180	3.76	3.5
	442K / 342K / 6910 / 409Mb	180	1.88	2.7
MATCHA	36.96mm^2 16nm PTM	2000	0.2	10
CONCRETE	Intel Xeon Silver 4208	2100	32	0.03
cuFHE	NVIDIA GeForce RTX 3090	1700	9.34	9.6

Clock frequency

Latency

Throughput

30

IACR 2022/1635

KU LEUVEN

Conclusions – lessons learned

- Different FHE schemes
- All operating on large data sizes
- Different hardware architectures
  - Dedicated domain-specific co-processor
  - Basalisc: programmable processor supporting BGV (and CKKS)
  - FPT: dedicated accelerator for TFHE programmable bootstrap

31

KU LEUVEN