YATA: Yet Another TFHE Accelerator with Key Compression and Radix-8 NTT

Kotaro Matsuoka Takashi Sato
Kyoto University, Japan
matsuoka.kotaro@gmail.com takashi@i.kyoto-u.ac.jp

TCHES 2026 Paper

Motivation

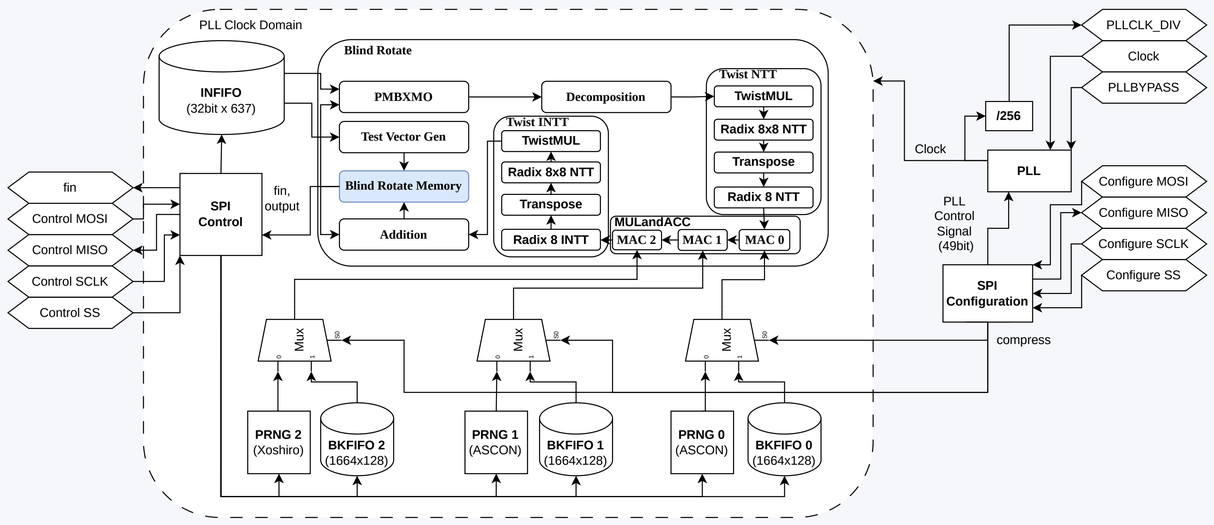
- **Bootstrapping** is the main performance bottleneck in FHE.
- In TFHE, Bootstrapping is dominated by **BlindRotate**; prior accelerators are often limited by **memory bandwidth** (large bootstrapping keys).
- To hit memory bottleneck, high performance polymul unit (NTT) is required.
- Goal: a **silicon-proven** TFHE accelerator with lowest Blind Rotate latency without requiring high speed memory.

Key Ideas / Contributions

- **Radix-8 NTT**: choose a prime modulus that makes twiddle multiplications and modular reductions hardware-friendly.
- **Key Compression**: store short **seeds** instead of nonce polynomials in the bootstrapping key; reconstruct on the fly with CSPRNG (ASCON XOF).
- **Silicon-proven TFHE**: fabricated in TSMC 22 nm ULP, 2×3 mm^2 die, **0.32 ms BlindRotate** latency. (ISSCC2026 25.2 is a concurrent work.)

Architecture Overview

- **Streaming Architecture:** Dedicated computation modules in a data-flow path.
- Equipped with SPI for communication and PLL for 100 MHz order clocks.
- *Note: PoC chip uses Xoshiro128 for non-nonce parts for the benchmark.*



Radix-8 NTT With Specialized Prime

- Choose $P = K^{r/2} \cdot 2^{(r/2)\cdot o} + 1$ (even radix r) so $(K \cdot 2^o)^r \equiv 1 \bmod P$.
- In YATA: K=5 reduces multiplication by powers of K to a few shift/add operations: K=5=2^2+1, K^2=25=2^4+2^3+1, K^3=125=2^7−2^2+1.
- (Signed) Montgomery friendly: $-(P-2)P \equiv 1 \bmod 2^{ro}$
- *Lazy* reduction (delay till after radix-r operations) to cut reduction overhead.

Key Compression (ASCON XOF)

- Bootstrapping key nonce polynomials dominate bandwidth
  $\Rightarrow$ Replace them with **seeds** and generate coefficients on-chip.
- BK size $\approx$ 18 MiB; at best-chip latency, required BK throughput $\approx$ 56.8 GiB/s.
- With Key Compression, effective throughput demand reduces to about **1/3** ($\approx$ 19 GB/s), and saves substantial SRAM area otherwise needed for nonce storage.
- Impl.: $k \cdot \texttt{nttsize} = 128$ parallel ASCON-XOF instances with rejection sampling. The NTT processes *nttsize* = 64 elements at once. (datapath width)
- The chip receives the post-absorption 320-bit ASCON state from the host; 64-bit XOF outputs are split into two 26-bit candidates and rejected if $\geq P$.
- ASCON modules are unrolled three times (two candidates per four cycles).

TFHE Parameters

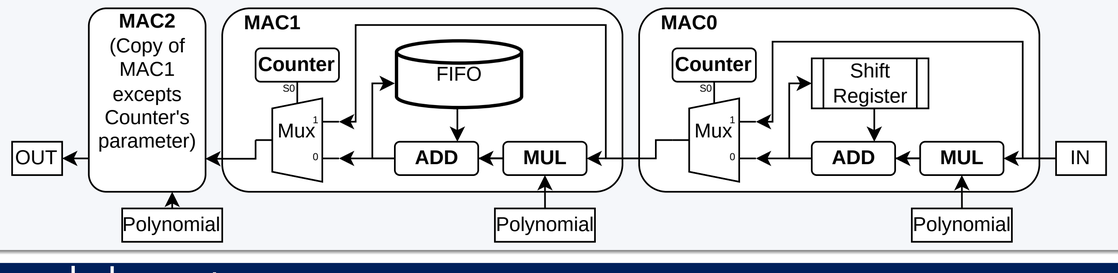
Supports only binary gates with $< 2^{-40}$ error rate.

Parameter	Value	Parameter	Value
Security estimate	$\lambda = 128$	TRLWE dimension	$k = 2$
BlindRotate loops	$n = 636$	Decomp base / levels	$Bg = 2^8$, $l = 2$
Poly degree	$N = 512$	NTT prime	$P = 5^4 \cdot 2^{16} + 1$

Note: parameters are hardcoded for a compact, debuggable tape-out; supporting flexible parameters would require more generic arithmetic/control and larger area.

MULandACC: Serial Architecture

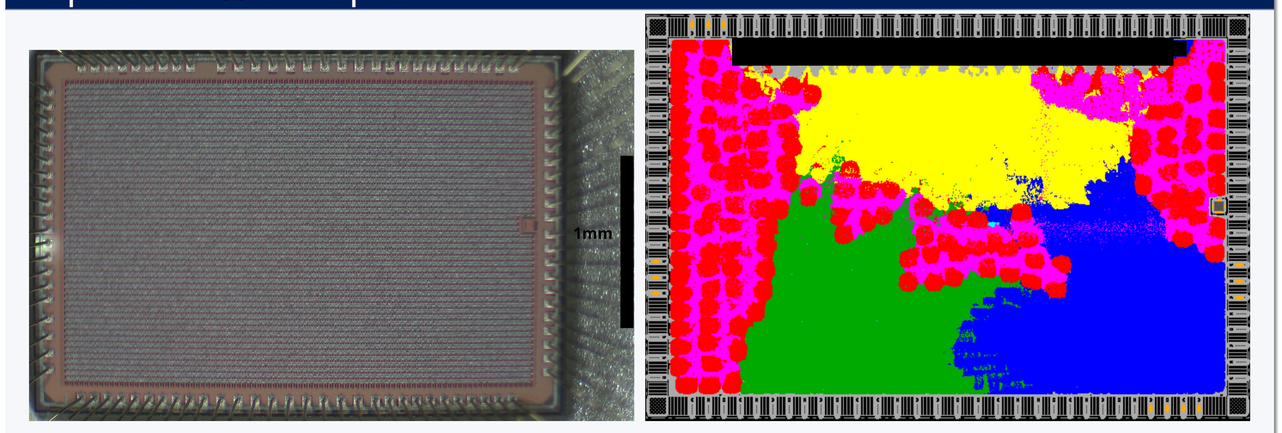
- k+1 MACs in **series** (not parallel as in FPT/HOGE): No parallel-to-serial needed.
- Multiplexed bus: pass-through NTT data and transmit results on the same wires.
- MAC 0 uses shift registers; MAC 1–2 use FIFO-based register files.



Acknowledgment

This work was supported by JST CREST Grant No. JPMJCR19K5 and in part by JSPS KAKENHI Grant No. 23KJ1319. We thank Prof. Kenichi Okada, Dr. Waleed Madany (Institute of Science, Tokyo), and Dr. Quan Cheng (Kyoto University) for the initial PLL design. The P&R was conducted by TOPPAN Inc.

Chip Photo & Floorplan



Area & Power Breakdown

Area breakdown

Module	Area%
I/O pad ring	30.5
Others (incl. NTT)	15.4
MULandACC	13.7
PRNGs (Key Compression)	13.4
INTT	13.2
SPI to PRNGs	9.5
SRAM IP	4.3

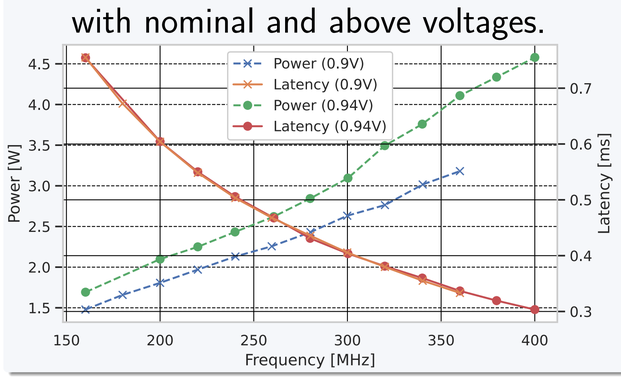
Simulated power @ 0.9 V / 500 MHz:

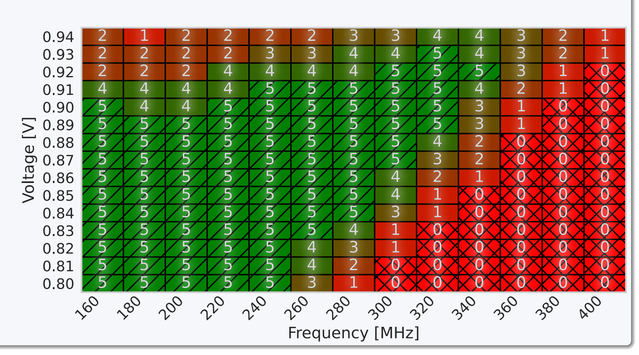
Module	Power [W]	Share
PRNGs	1.264	46%
Others (incl. NTT)	0.435	16%
MULandACC	0.374	14%
SPI to PRNGs	0.357	12%
INTT	0.319	12%
Total	2.749	100%

- Key Compression shifts the bottleneck from off-chip bandwidth to on-chip PRNG logic (power-dominant in simulation and non-negligible area).

Measured Trends

- The best result is achieved by only one of the five measured chips.
- Non-negligible on-chip IR drop in the minimal die: external 1.13 V needed to maintain 0.94 V internally (remote sense used).
- All 5 chips: $\geq$340 MHz @ 0.92 V; hold-time violations observed at lower frequencies with nominal and above voltages.





Comparison with Prior Work

- Reported point: 400 MHz @ 0.94 V
- 2× lower latency than best FPGA (FPT) without BKU
- Lowest latency as a silicon-proven ASIC
- 7–33× smaller area than simulation-based ASICs

Work	Platform	Clock [MHz]	Latency w/o BKU [ms]	Power [W]	PDP [mJ]	Area [mm^2]
FPGA implementations						
YKP	FPGA	180	7.53	50	376.5	—
ALT	FPGA	115	0.84	23	19.3	—
HOGE	FPGA	297	1.00	38	38.0	—
FPT	FPGA	200	0.66	99	65.3	—
[LLW+24]	FPGA	—	1.16	—	—	—
Simulation-based ASICs						
MATCHA	16 nm	2000	0.60	40	24.0	37.0
Morphling	28 nm	1200	0.11	53.0	5.83	74.8
Strix	28 nm	1200	0.16	77.1	12.3	141.4
UFC	7 nm	1000	0.03	76.9	2.31	197.7
Silicon-proven ASICs						
ISSCC 2026 25.2	28 nm	390	> 1	1.1	> 0.48	2.25
YATA	22 nm	400	**0.32**	4.58	1.47	5.93

Conclusion / Takeaways

- Radix-8 NTT with specialized primes enables shift/add-friendly twiddles and cheaper modular reduction.
- Key Compression reduces the dominant TFHE bandwidth bottleneck by expanding nonce polynomials from seeds.
- Silicon results show a compact 5.93 mm^2 BlindRotate engine at 0.32 ms latency.

Future work: Fast Communications; Batch Bootstrapping (pipelining) support.
Cost: $\approx$ 100K USD for design & fabrication (academic price)
Open-source RTL: `https://github.com/virtualsecureplatform/YATA`

References

[LLW+24] Zhihao Li, Xianhui Lu, Zhiwei Wang, et al. Faster NTRU-based Bootstrapping in less than 4 ms. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2024(3):418–451, July 2024. Number: 3.