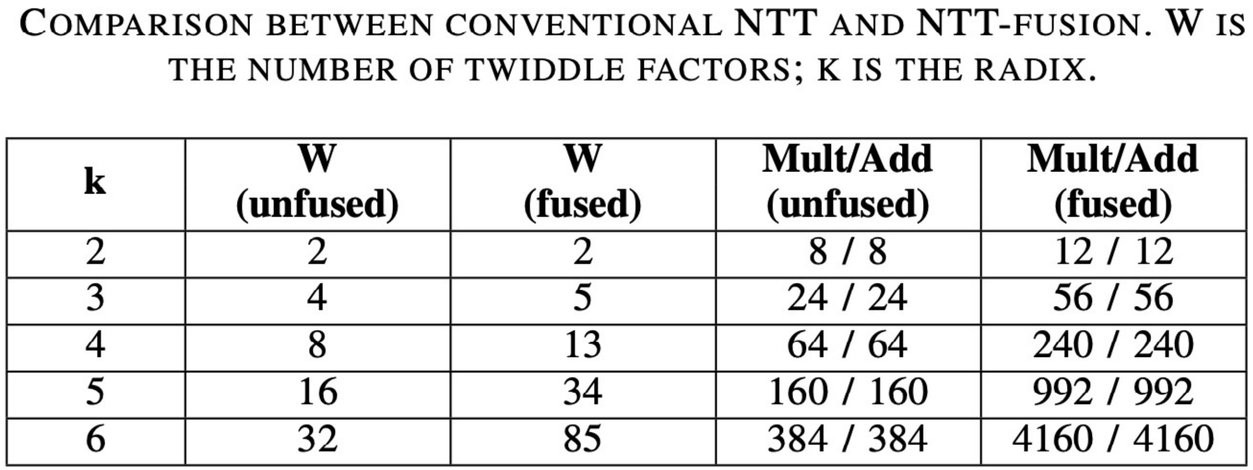
Poseidon: Practical Homomorphic Encryption Accelerator

(2023 HPCA)

Methodology

NTT-fusion



Comparison between conventional NTT and NTT-fusion. W is the number of twiddle factors; k is the radix.

k	W (unfused)	W (fused)	Mult/Add (unfused)	Mult/Add (fused)
2	2	2	8 / 8	12 / 12
3	4	5	24 / 24	56 / 56
4	8	13	64 / 64	240 / 240
5	16	34	160 / 160	992 / 992
6	32	85	384 / 384	4160 / 4160

HFAuto

FHE ACCELERATOR-POSEIDON

Overall Architecture

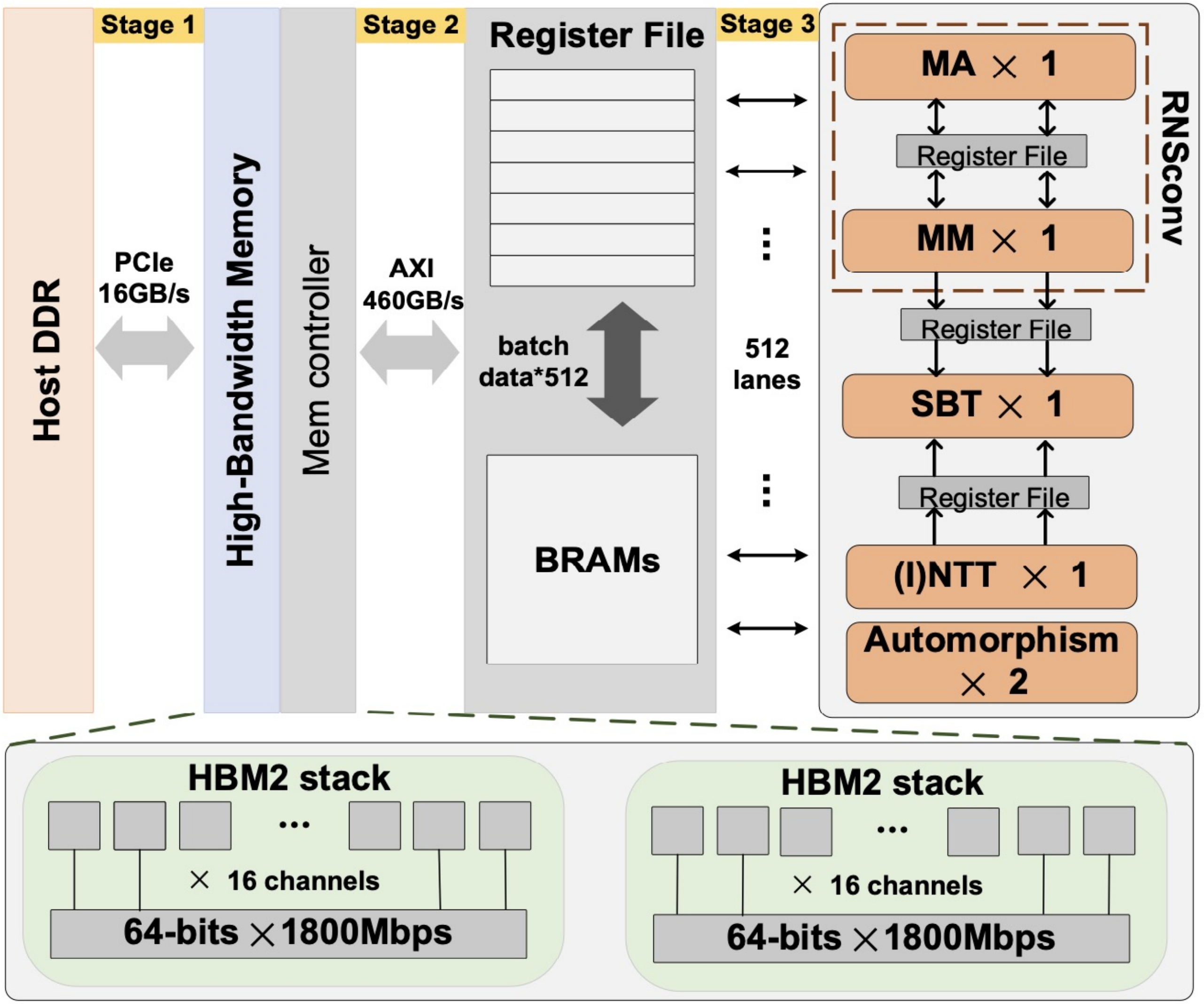


Fig. 2. Poseidon overall architecture.

Computational Cores

MA/MM

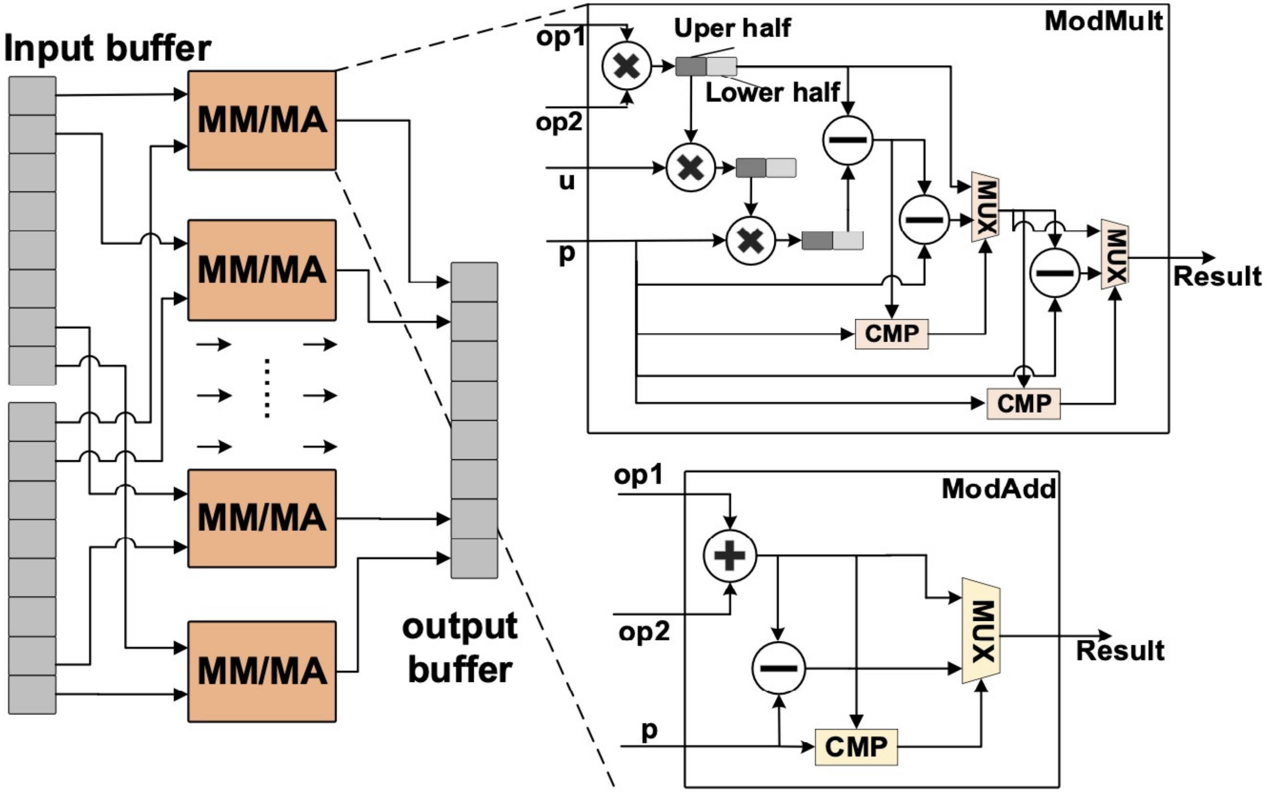


Fig. 3. MA/MM core architecture. We implement fine-grained decomposition to reduce the resource consumption of ModMult. Following Barrett Reduction algorithm, we implement a subtractor to perform ModAdd.

RNSconv

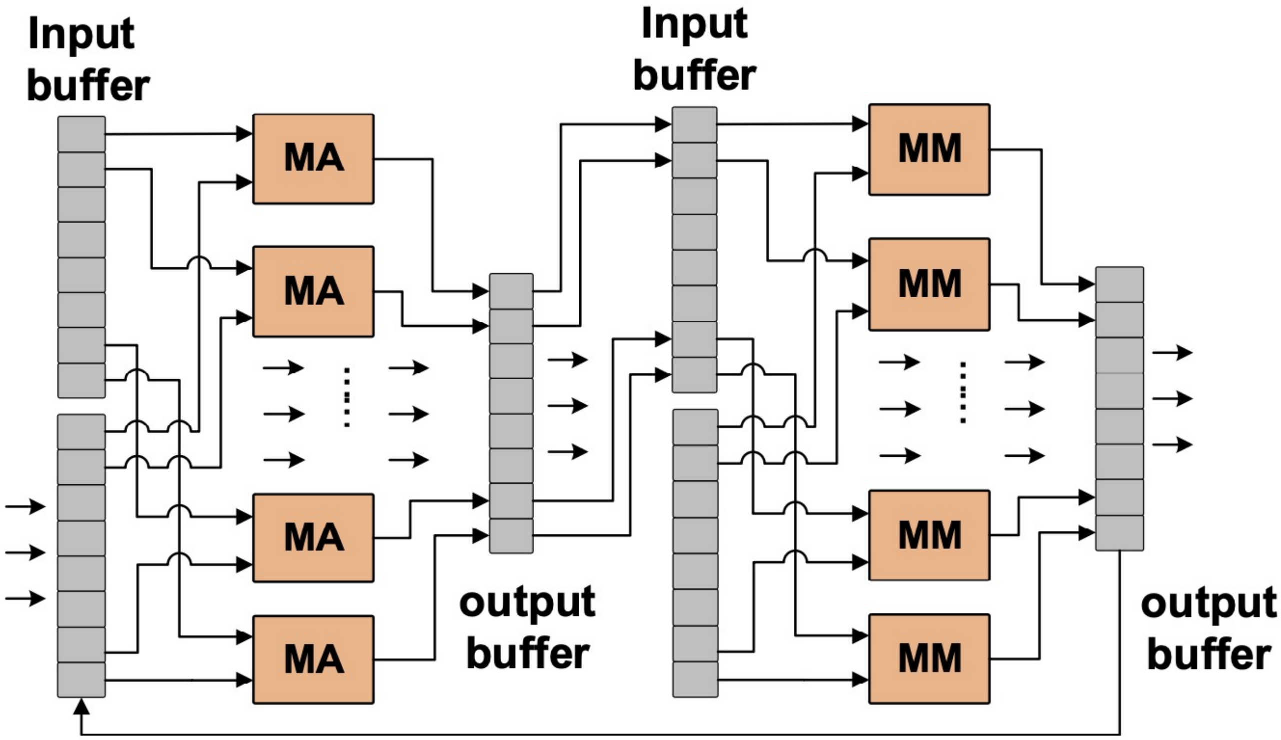
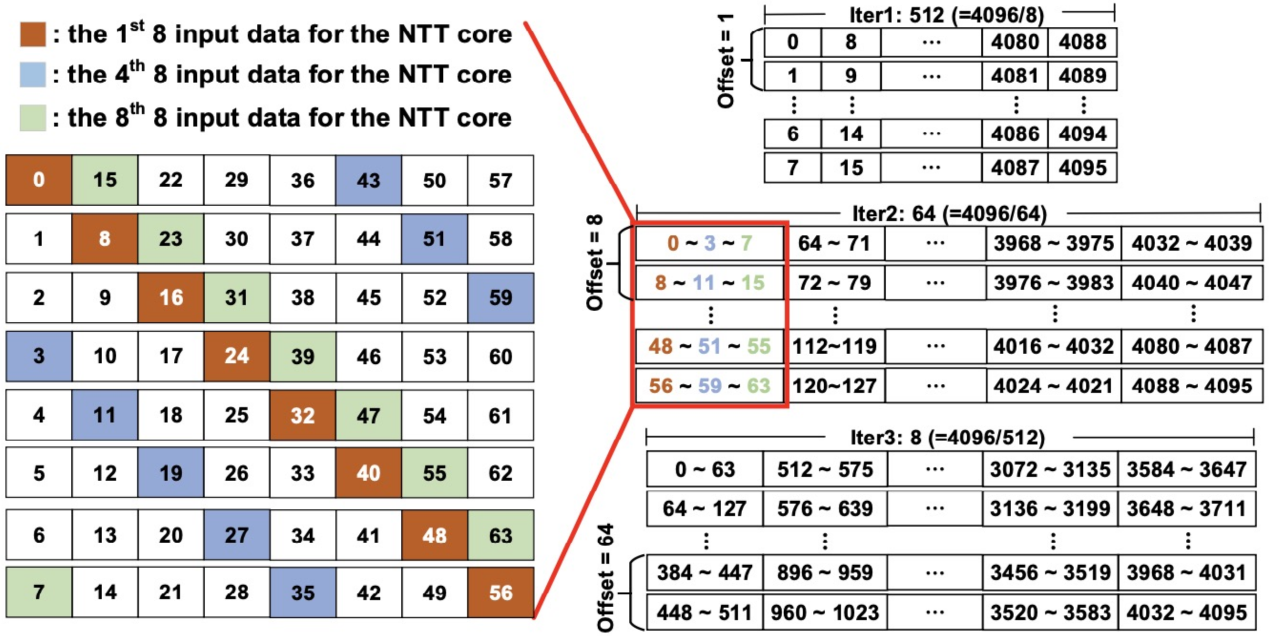


Fig. 4. RNSconv architecture, designed by cascading MA and MM core.

NTT/INTT



🟧 : the 1st 8 input data for the NTT core

🟦 : the 4th 8 input data for the NTT core

🟩 : the 8th 8 input data for the NTT core

0	15	22	29	36	43	50	57
1	8	23	30	37	44	51	58
2	9	16	31	38	45	52	59
3	10	17	24	39	46	53	60
4	11	18	25	32	47	54	61
5	12	19	26	33	40	55	62
6	13	20	27	34	41	48	63
7	14	21	28	35	42	49	56

Iter1: 512 (=4096/8)

Offset = 1

0	8	···	4080	4088
1	9	···	4081	4089
⋮	⋮		⋮	⋮
6	14	···	4086	4094
7	15	···	4087	4095

Iter2: 64 (=4096/64)

Offset = 8

0 ~ 3 ~ 7	64 ~ 71	···	3968 ~ 3975	4032 ~ 4039
8 ~ 11 ~ 15	72 ~ 79	···	3976 ~ 3983	4040 ~ 4047
⋮	⋮		⋮	⋮
48 ~ 51 ~ 55	112~119	···	4016 ~ 4032	4080 ~ 4087
56 ~ 59 ~ 63	120~127	···	4024 ~ 4021	4088 ~ 4095

Iter3: 8 (=4096/512)

Offset = 64

0 ~ 63	512 ~ 575	···	3072 ~ 3135	3584 ~ 3647
64 ~ 127	576 ~ 639	···	3136 ~ 3199	3648 ~ 3711
⋮	⋮		⋮	⋮
384 ~ 447	896 ~ 959	···	3456 ~ 3519	3968 ~ 4031
448 ~ 511	960 ~ 1023	···	3520 ~ 3583	4032 ~ 4095

Fig. 5. Data access pattern in Poseidon. We show the front three iterations (or phases) in TABLE III.

Data access pattern

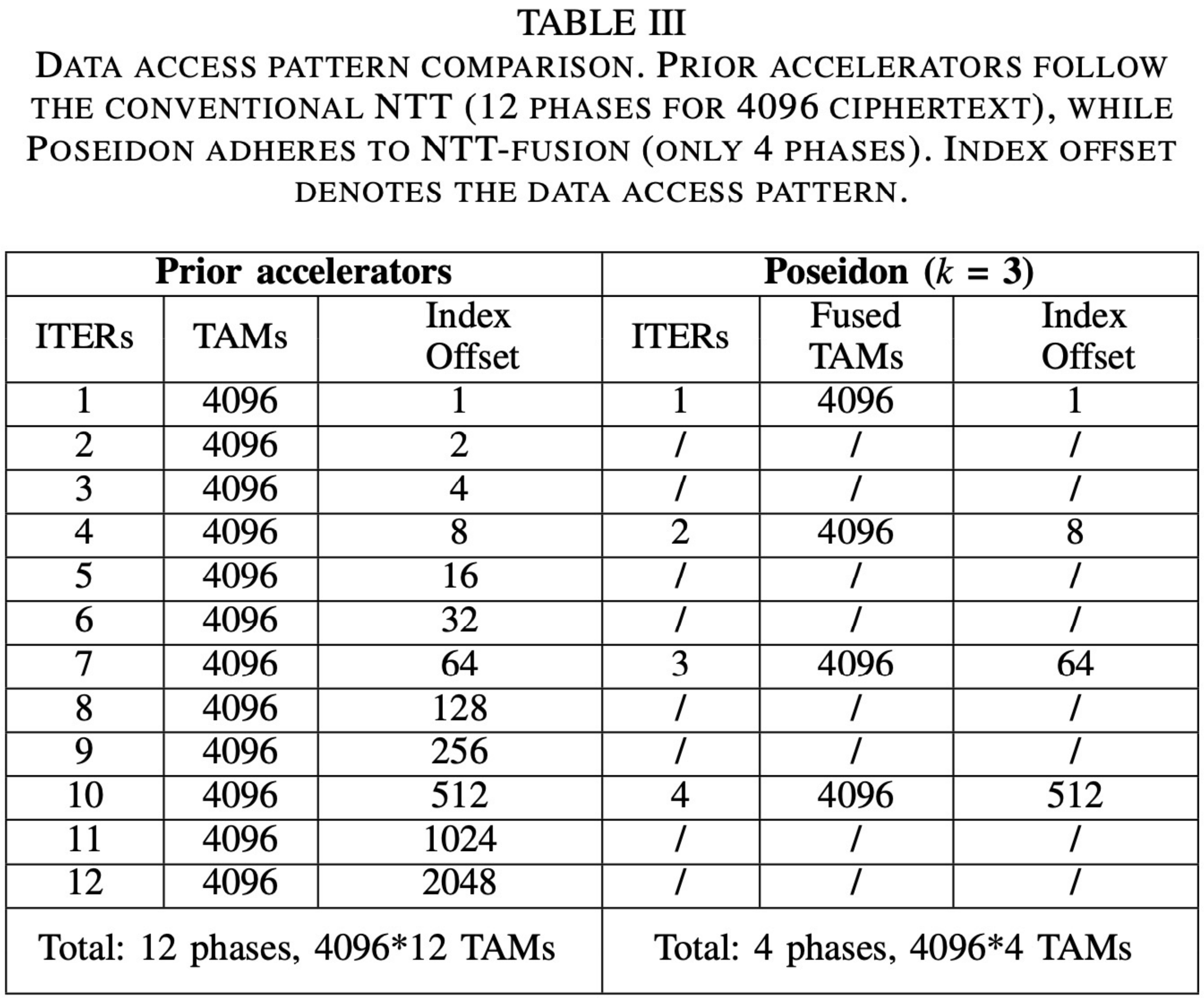
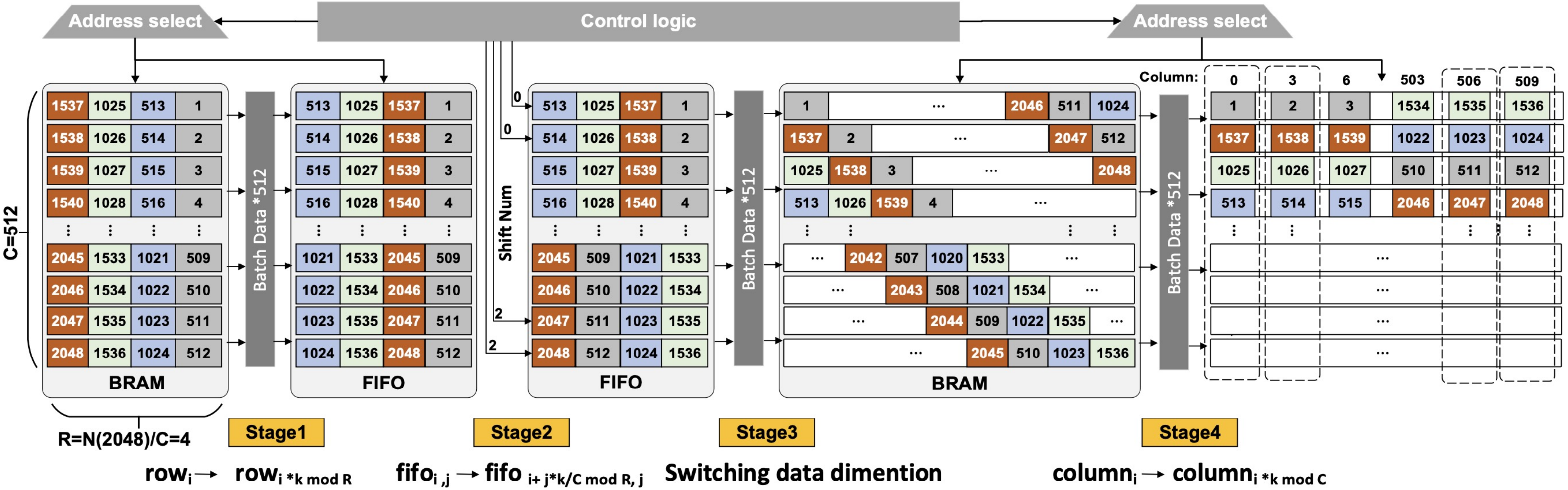


TABLE III

Data access pattern comparison. Prior accelerators follow the conventional NTT (12 phases for 4096 ciphertext), while Poseidon adheres to NTT-fusion (only 4 phases). Index offset denotes the data access pattern.

Prior accelerators			**Poseidon (k = 3)**		
ITERs	TAMs	Index Offset	ITERs	Fused TAMs	Index Offset
1	4096	1	1	4096	1
2	4096	2	/	/	/
3	4096	4	/	/	/
4	4096	8	2	4096	8
5	4096	16	/	/	/
6	4096	32	/	/	/
7	4096	64	3	4096	64
8	4096	128	/	/	/
9	4096	256	/	/	/
10	4096	512	4	4096	512
11	4096	1024	/	/	/
12	4096	2048	/	/	/
Total: 12 phases, 4096*12 TAMs			Total: 4 phases, 4096*4 TAMs		

Automorphism



BRAM

1537	1025	513	1
1538	1026	514	2
1539	1027	515	3
1540	1028	516	4
⋮	⋮	⋮	⋮
2045	1533	1021	509
2046	1534	1022	510
2047	1535	1023	511
2048	1536	1024	512

FIFO

513	1025	1537	1
514	1026	1538	2
515	1027	1539	3
516	1028	1540	4
⋮	⋮	⋮	⋮
1021	1533	2045	509
1022	1534	2046	510
1023	1535	2047	511
1024	1536	2048	512

FIFO

513	1025	1537	1
514	1026	1538	2
515	1027	1539	3
516	1028	1540	4
⋮	⋮	⋮	⋮
2045	509	1021	1533
2046	510	1022	1534
2047	511	1023	1535
2048	512	1024	1536

BRAM

1 … 2046 511 1024
1537 2 … 2047 512
1025 1538 3 … 2048
513 1026 1539 4 …
⋮ ⋮ ⋮ ⋮
… 2042 507 1020 1533 …
… 2043 508 1021 1534 …
… 2044 509 1022 1535 …
… 2045 510 1023 1536

0	3	6	503	506	509
1	2	3	1534	1535	1536
1537	1538	1539	1022	1023	1024
1025	1026	1027	510	511	512
513	514	515	2046	2047	2048
⋮	⋮			⋮	⋮
		…			
		…			
		…			
		…			

Fig. 6. Automorphism architecture in Poseidon. Only one dual-port BRAM is needed. The result of Stage 4 will be written back to HBM directly.

EVALUATION

Platform

Xilinx Alveo U280(owns HBM) FPGA plugged into the PCIe slot of the mainboard
Vivado and Vitis on the host side.

Baseline

CPU (Intel Xeon Gold 6234) running at 3.3 GHz with a single thread
state-of-the-art GPU
FPGA
4 FHE accelerator ASICs

Benchmark

- Logistic regression (LR). It is the HELR algorithm implementation based on the CKKS scheme.In combination with Bootstrapping, we use the multiplication depth of L = 38 and evaluate the average performance of 10 iterations supported by two Bootstrapping operations.
- LSTM. It is the Long-Term Short-Term (LSTM) model. It requires 50 Bootstrapping operations in total during one inference.
- ResNet-20. This benchmark is the inference of an image on the ResNet-20 model implemented with FHE.
- Packed bootstrapping(packed bootstrapping algorithm). bootstrapping L = 3; multiplication depth L = 57.

Accelerator Performance

FHE Basic Operations

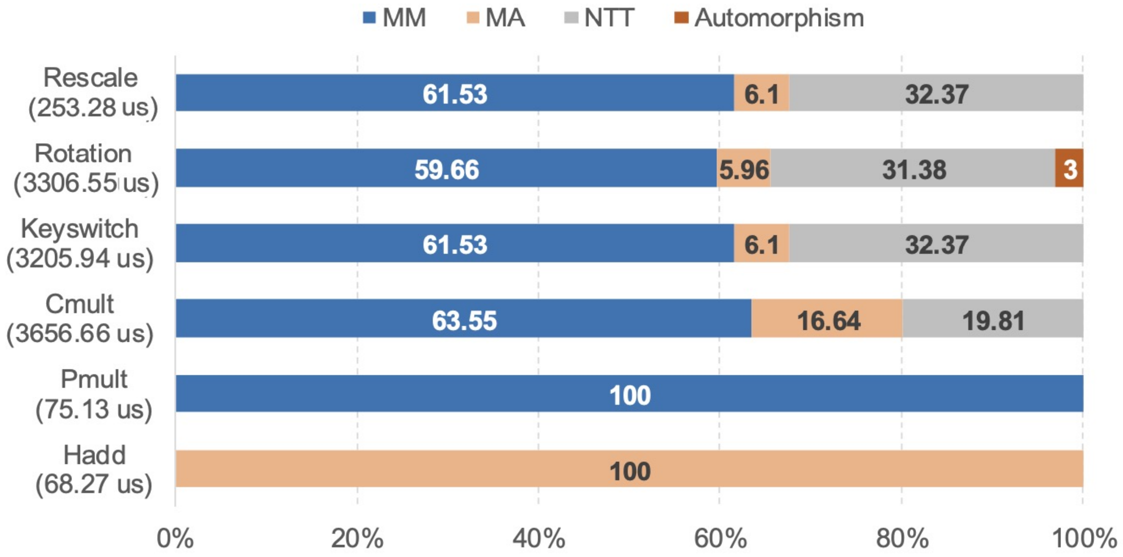
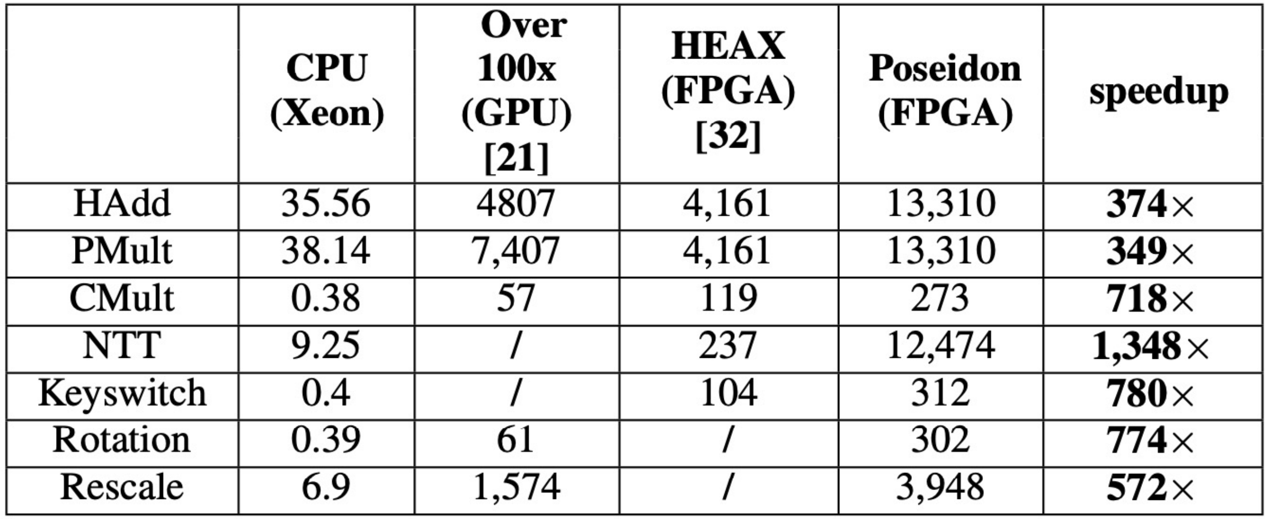


Fig. 7. Operator core analysis. The ciphertext parameters are set to $N = 2^{16}, L = 44$.

TABLE IV

PERFORMANCE COMPARISON OF FHE BASIC OPERATIONS. WE USE “OPERATIONS PER SECOND” AS THE PERFORMANCE METRIC.



	CPU (Xeon)	Over 100x (GPU) [21]	HEAX (FPGA) [32]	Poseidon (FPGA)	speedup
HAdd	35.56	4807	4,161	13,310	**374**×
PMult	38.14	7,407	4,161	13,310	**349**×
CMult	0.38	57	119	273	**718**×
NTT	9.25	/	237	12,474	**1,348**×
Keyswitch	0.4	/	104	312	**780**×
Rotation	0.39	61	/	302	**774**×
Rescale	6.9	1,574	/	3,948	**572**×

Full-system performance

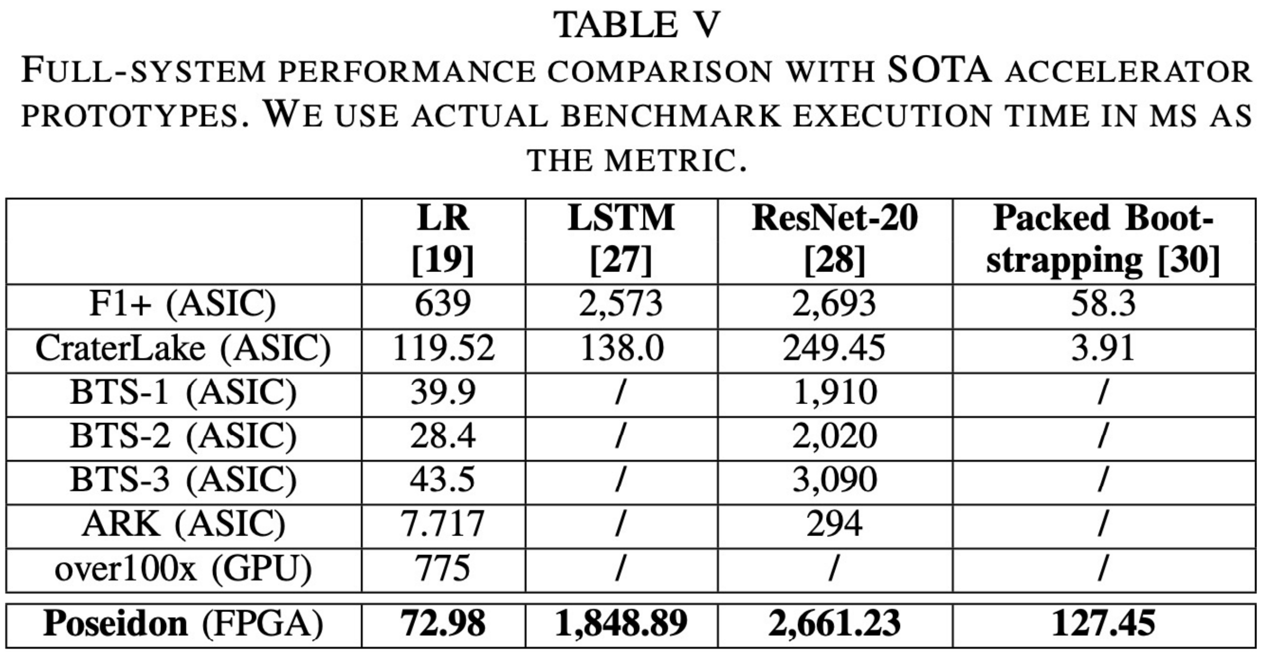


TABLE V

FULL-SYSTEM PERFORMANCE COMPARISON WITH SOTA ACCELERATOR PROTOTYPES. WE USE ACTUAL BENCHMARK EXECUTION TIME IN MS AS THE METRIC.

	LR [19]	**LSTM [27]**	**ResNet-20 [28]**	**Packed Bootstrapping [30]**
F1+ (ASIC)	639	2,573	2,693	58.3
CraterLake (ASIC)	119.52	138.0	249.45	3.91
BTS-1 (ASIC)	39.9	/	1,910	/
BTS-2 (ASIC)	28.4	/	2,020	/
BTS-3 (ASIC)	43.5	/	3,090	/
ARK (ASIC)	7.717	/	294	/
over100x (GPU)	775	/	/	/
Poseidon (FPGA)	**72.98**	**1,848.89**	**2,661.23**	**127.45**

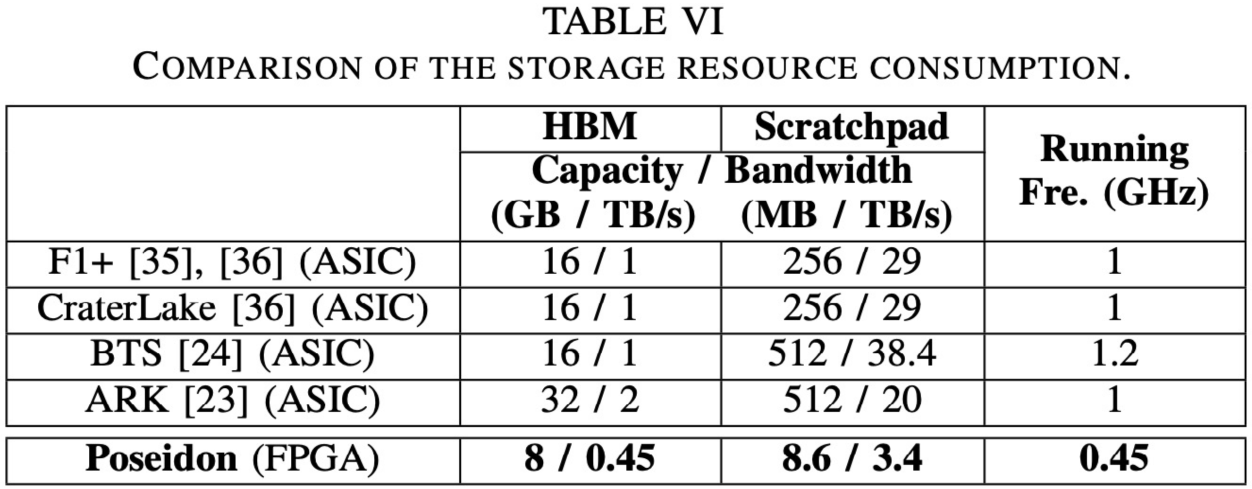


TABLE VI

COMPARISON OF THE STORAGE RESOURCE CONSUMPTION.

	HBM	**Scratchpad**	**Running Fre. (GHz)**
	Capacity / Bandwidth		
	(GB / TB/s)	**(MB / TB/s)**	
F1+ [35], [36] (ASIC)	16 / 1	256 / 29	1
CraterLake [36] (ASIC)	16 / 1	256 / 29	1
BTS [24] (ASIC)	16 / 1	512 / 38.4	1.2
ARK [23] (ASIC)	32 / 2	512 / 20	1
Poseidon (FPGA)	**8 / 0.45**	**8.6 / 3.4**	**0.45**

Bandwidth Utilization

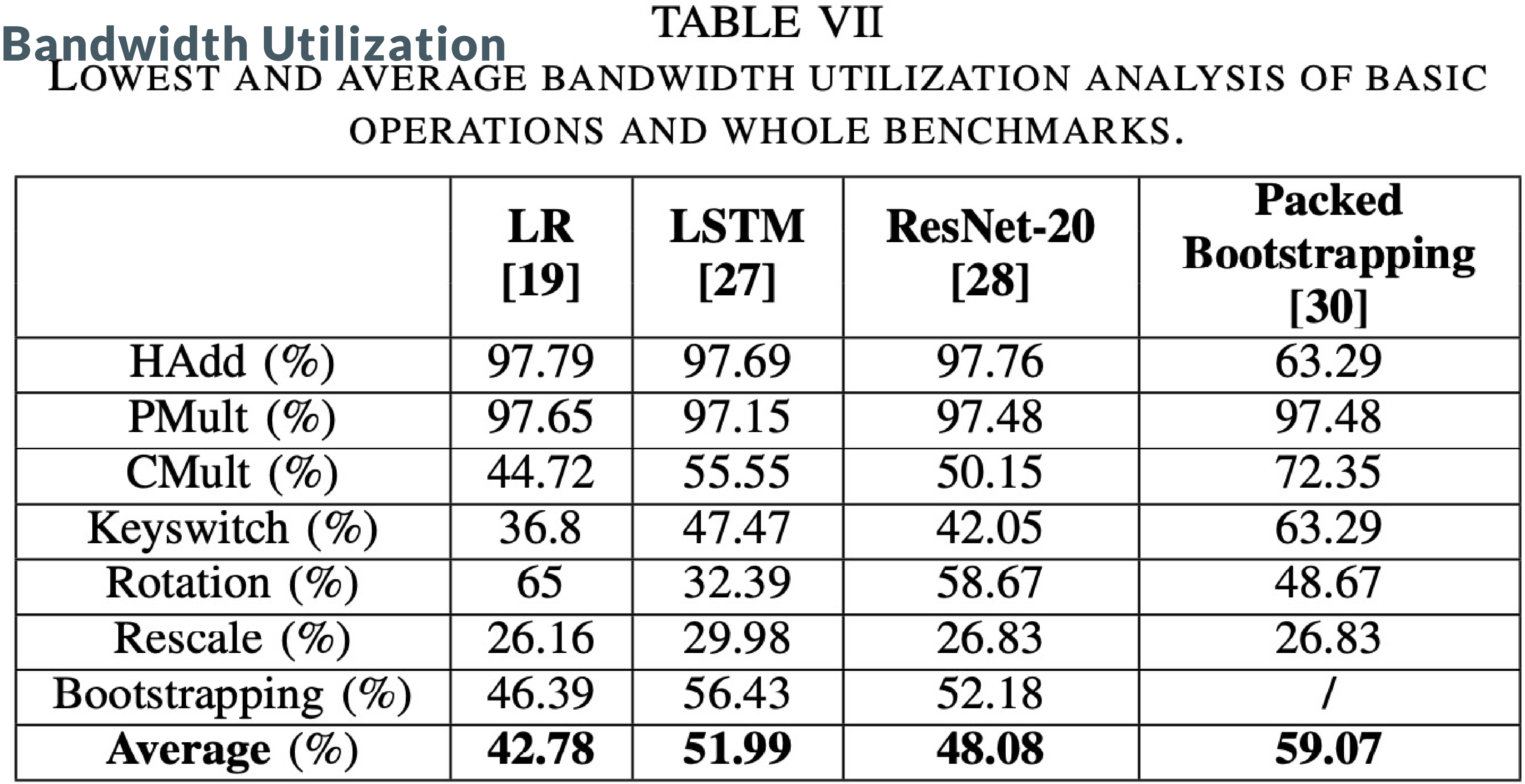


TABLE VII

LOWEST AND AVERAGE BANDWIDTH UTILIZATION ANALYSIS OF BASIC OPERATIONS AND WHOLE BENCHMARKS.

	LR [19]	**LSTM [27]**	**ResNet-20 [28]**	**Packed Bootstrapping [30]**
HAdd (%)	97.79	97.69	97.76	63.29
PMult (%)	97.65	97.15	97.48	97.48
CMult (%)	44.72	55.55	50.15	72.35
Keyswitch (%)	36.8	47.47	42.05	63.29
Rotation (%)	65	32.39	58.67	48.67
Rescale (%)	26.16	29.98	26.83	26.83
Bootstrapping (%)	46.39	56.43	52.18	/
Average (%)	**42.78**	**51.99**	**48.08**	**59.07**

Poseidon Specifics

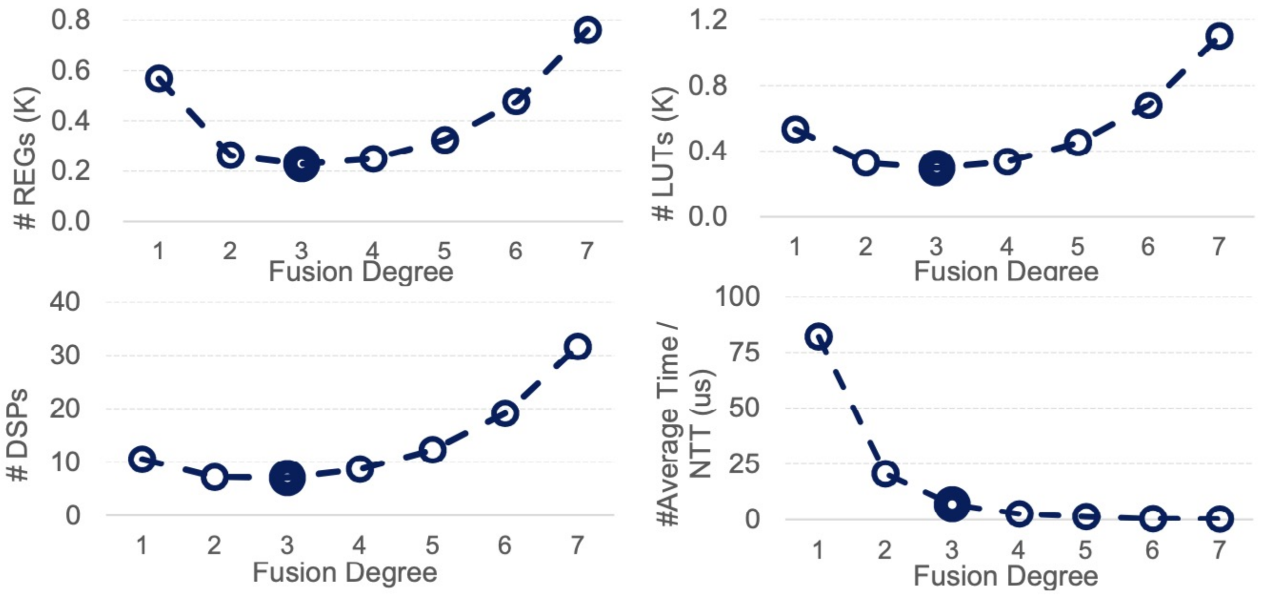


Fig. 10. Parameter Selection – *k*. We evaluated the FPGA resource usage (in actual #) and the Average Execution Time per NTT (bottom right), scaled by *k*. The optimal point emerges at $k = 3$, where it consumes the lowest resources with the highest speed.

HFAuto

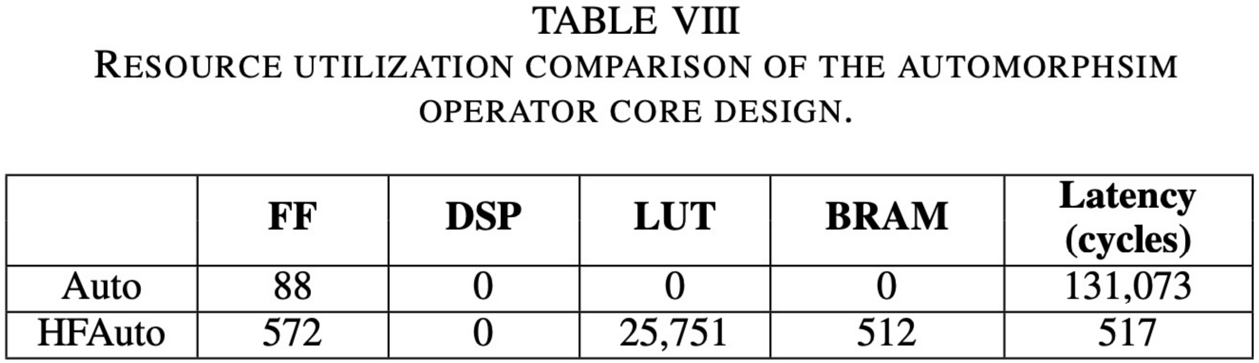


TABLE VIII
RESOURCE UTILIZATION COMPARISON OF THE AUTOMORPHSIM OPERATOR CORE DESIGN.

	FF	**DSP**	**LUT**	**BRAM**	**Latency (cycles)**
Auto	88	0	0	0	131,073
HFAuto	572	0	25,751	512	517

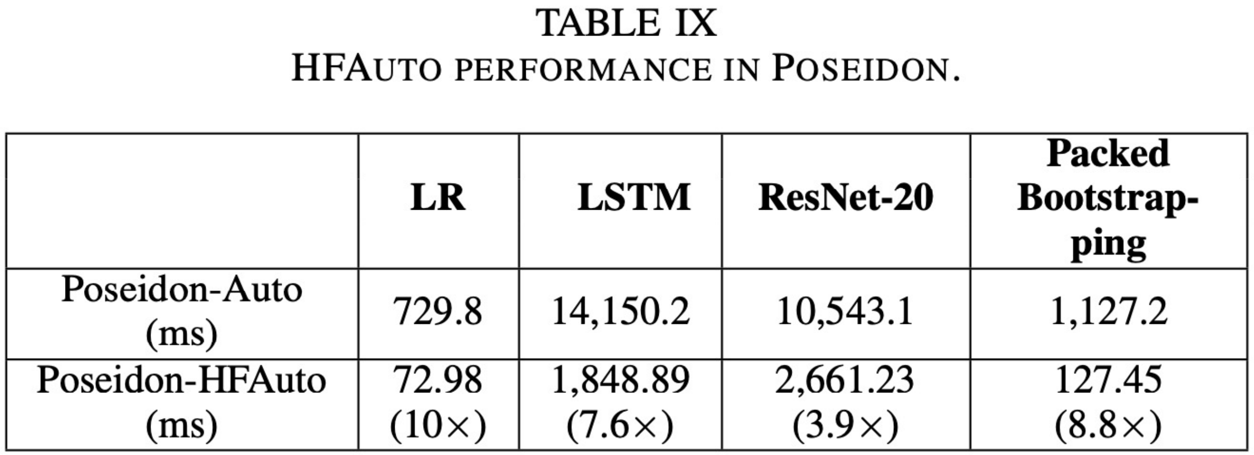


TABLE IX
HFAUTO PERFORMANCE IN POSEIDON.

	LR	**LSTM**	**ResNet-20**	**Packed Bootstrapping**
Poseidon-Auto (ms)	729.8	14,150.2	10,543.1	1,127.2
Poseidon-HFAuto (ms)	72.98 (10×)	1,848.89 (7.6×)	2,661.23 (3.9×)	127.45 (8.8×)

Scalability

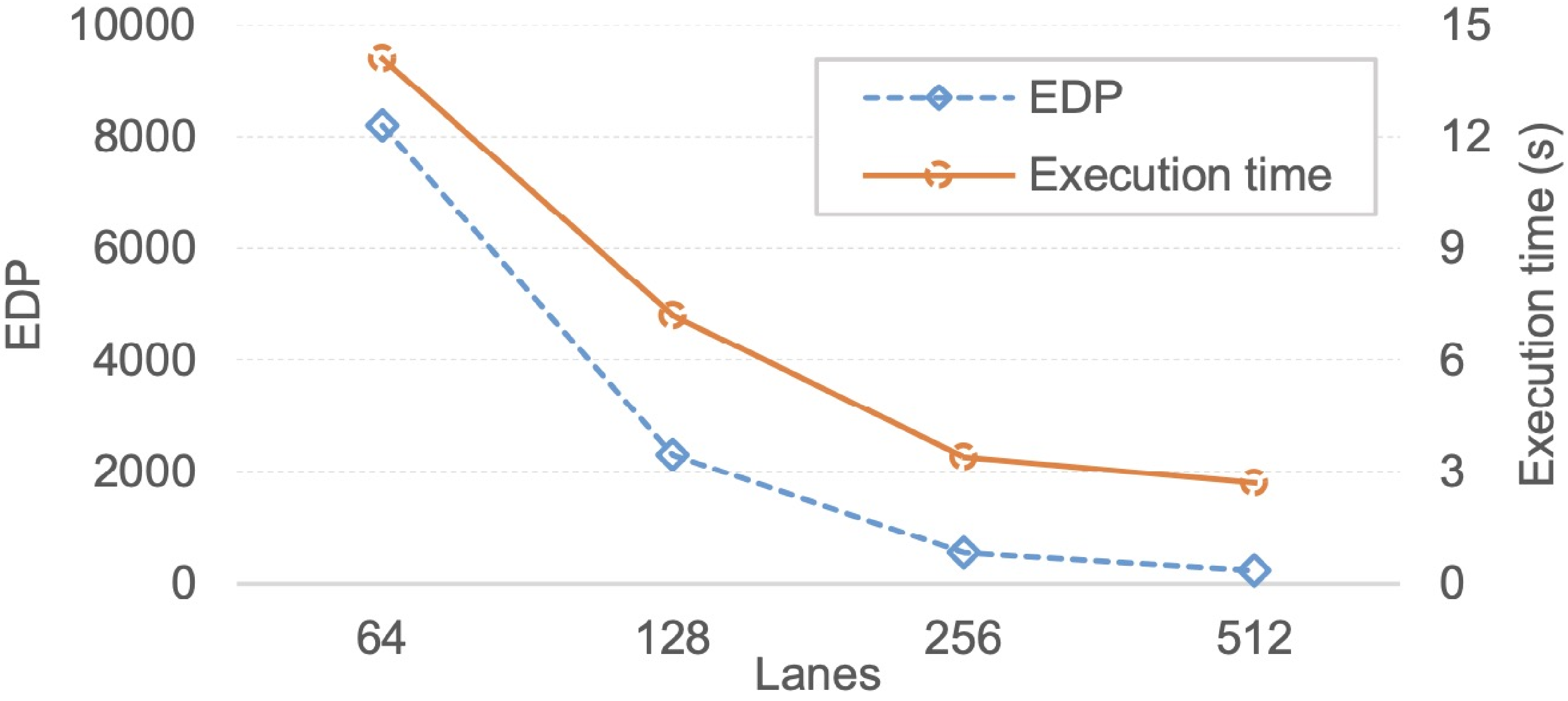


Fig. 11. Sensitivity of the lanes.

Energy

Energy Consumption and Breakdown

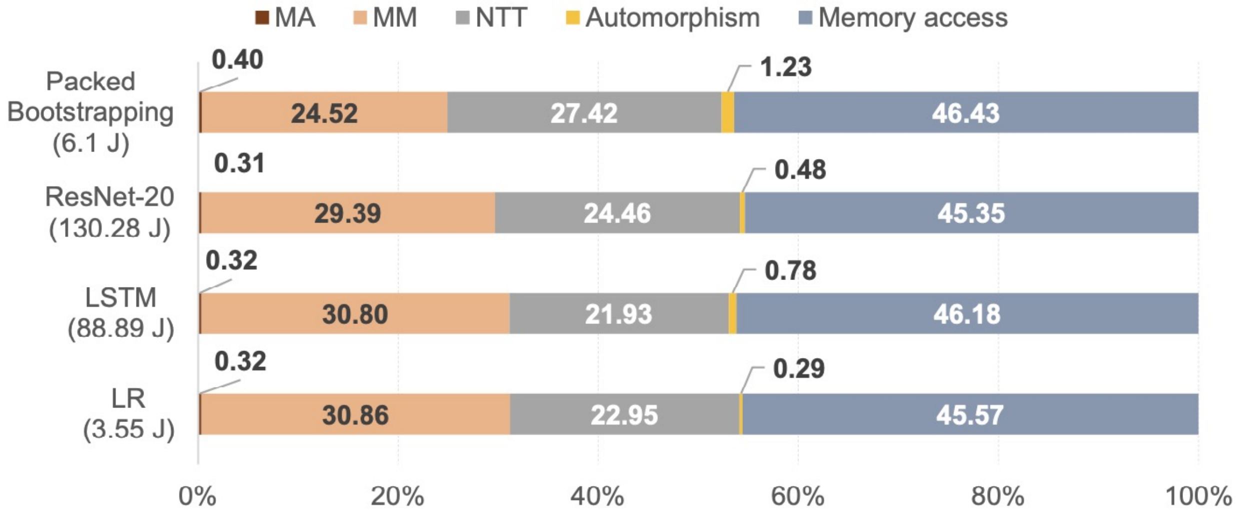


Fig. 12. Energy consumption and breakdown. MM and NTT operator cores take the major proportion besides the memory access.

Efficiency and Utilization

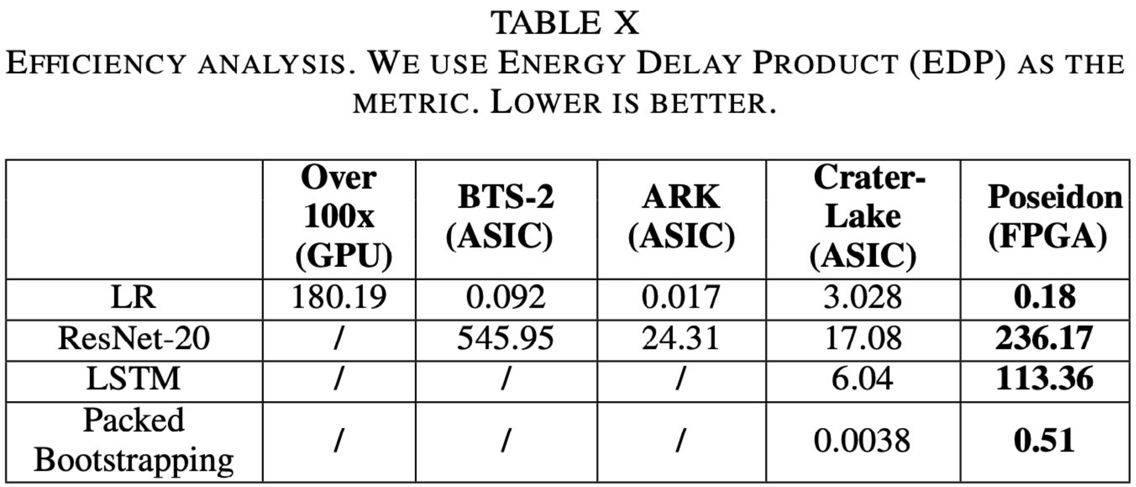


TABLE X
EFFICIENCY ANALYSIS. WE USE ENERGY DELAY PRODUCT (EDP) AS THE METRIC. LOWER IS BETTER.

	Over 100x (GPU)	BTS-2 (ASIC)	ARK (ASIC)	Crater-Lake (ASIC)	Poseidon (FPGA)
LR	180.19	0.092	0.017	3.028	**0.18**
ResNet-20	/	545.95	24.31	17.08	**236.17**
LSTM	/	/	/	6.04	**113.36**
Packed Bootstrapping	/	/	/	0.0038	**0.51**

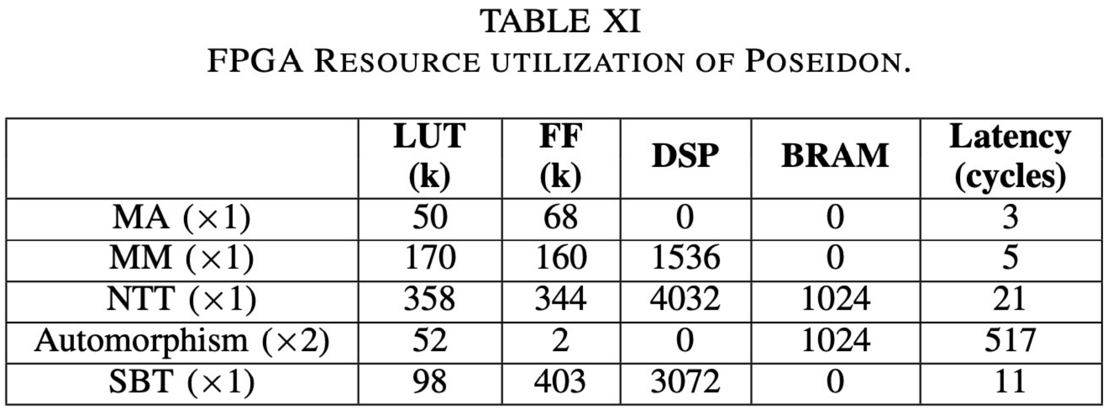


TABLE XI
FPGA RESOURCE UTILIZATION OF POSEIDON.

	LUT (k)	FF (k)	DSP	BRAM	Latency (cycles)
MA (×1)	50	68	0	0	3
MM (×1)	170	160	1536	0	5
NTT (×1)	358	344	4032	1024	21
Automorphism (×2)	52	2	0	1024	517
SBT (×1)	98	403	3072	0	11

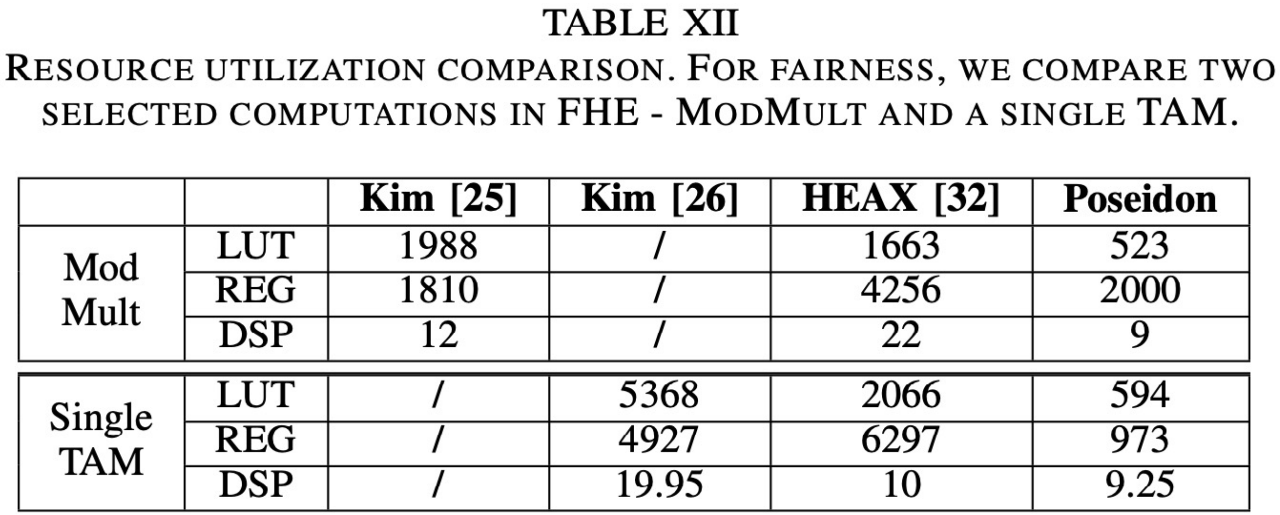


TABLE XII
RESOURCE UTILIZATION COMPARISON. FOR FAIRNESS, WE COMPARE TWO SELECTED COMPUTATIONS IN FHE - MODMULT AND A SINGLE TAM.

		Kim [25]	Kim [26]	HEAX [32]	Poseidon
Mod Mult	LUT	1988	/	1663	523
	REG	1810	/	4256	2000
	DSP	12	/	22	9
Single TAM	LUT	/	5368	2066	594
	REG	/	4927	6297	973
	DSP	/	19.95	10	9.25