FLUTE: A LUT-Based Floating-Point Engine for Efficient LLM Acceleration Exploiting Core-Tail Hierarchical Alignment

Tielong Liu[1,2], Zeyu Zhu[1,2], Gang Li[1], Zitao Mo[1], Chen Bai[3], Junwen Si[1],
Peihuan Ni[1], Minnan Pei[1], Peisong Wang[1], Zhuoran Song[4], Xiaoyao Liang[4], Jian Cheng[1]

[1]C[2]DL, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China
[3]Fudan University, Shanghai, China
[4]Shanghai Jiao Tong University, Shanghai, China

MICRO 2026

Outline

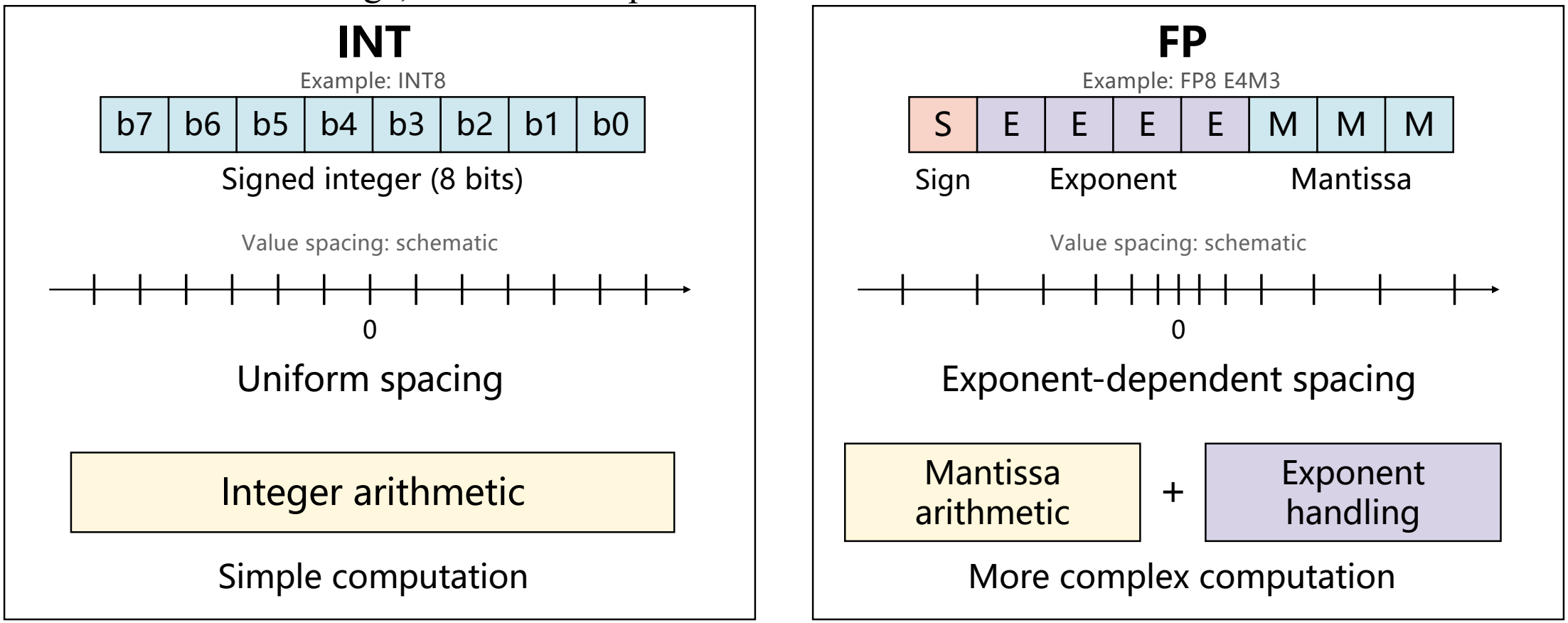
中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

- Background and Motivation
- Core-Tail Hierarchical Exponent Alignment
- The FLUTE Accelerator Architecture
- Evaluation

Background and Motivation

Weight-only Quantization: INT or FP?

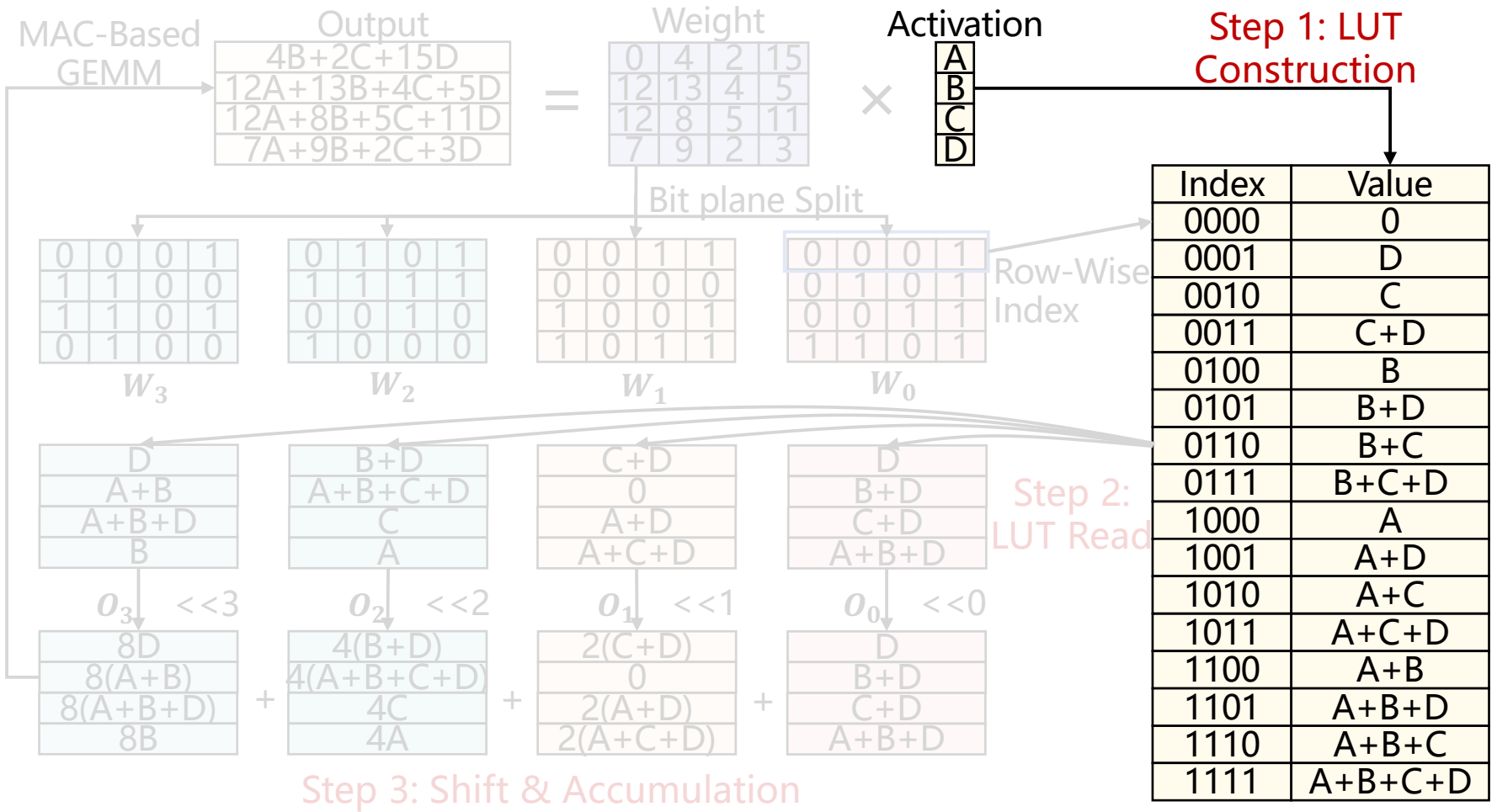
- FP: wider numeric range, but more complex arithmetic.



LUT computing is a promising approach to efficient low-bit FP inference.

Background and Motivation

LUT Computing — Step 1: LUT Construction

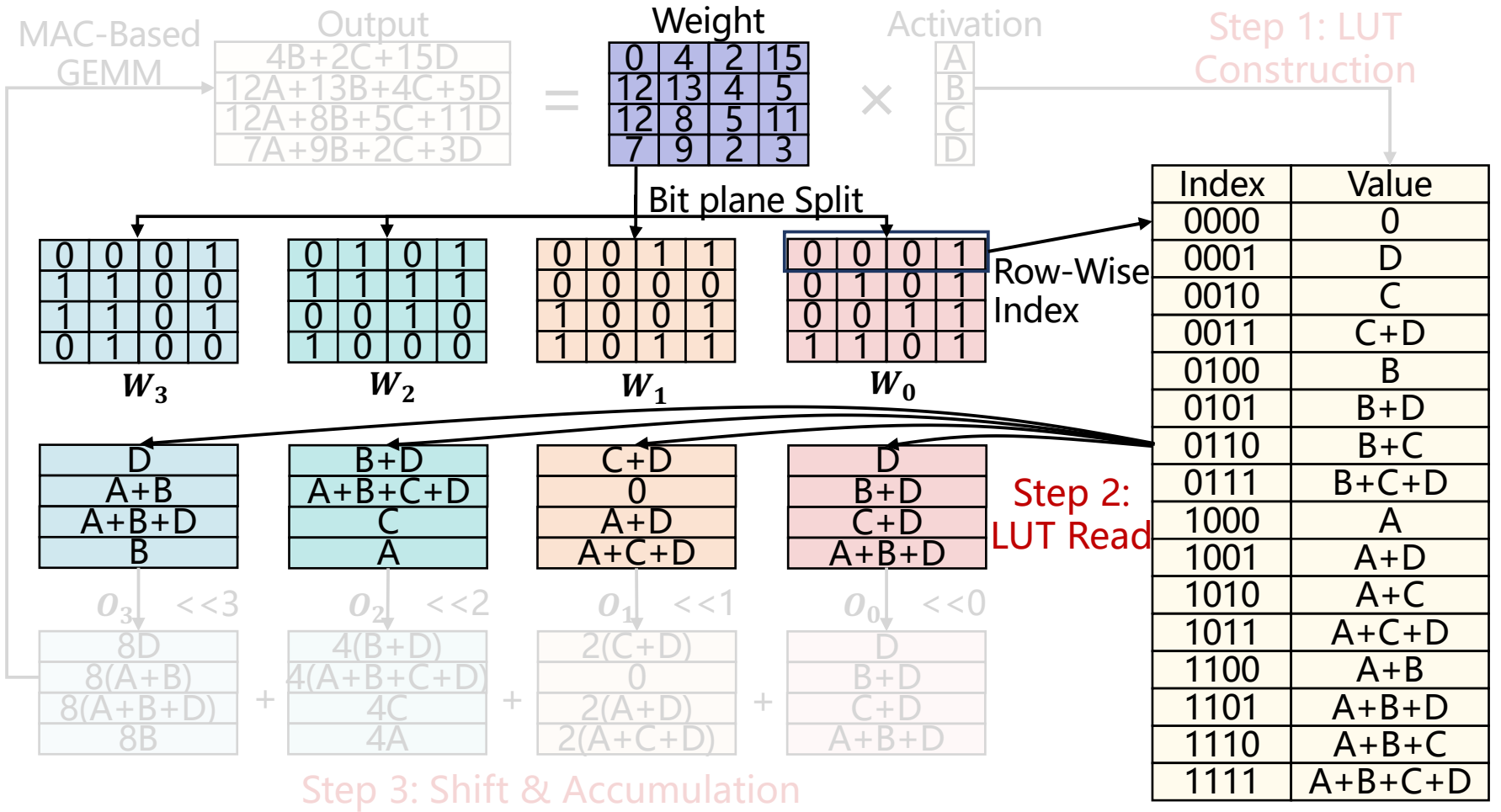


Index	Value
0000	0
0001	D
0010	C
0011	C+D
0100	B
0101	B+D
0110	B+C
0111	B+C+D
1000	A
1001	A+D
1010	A+C
1011	A+C+D
1100	A+B
1101	A+B+D
1110	A+B+C
1111	A+B+C+D

Build the activation LUT once and reuse it across weight rows.

Background and Motivation

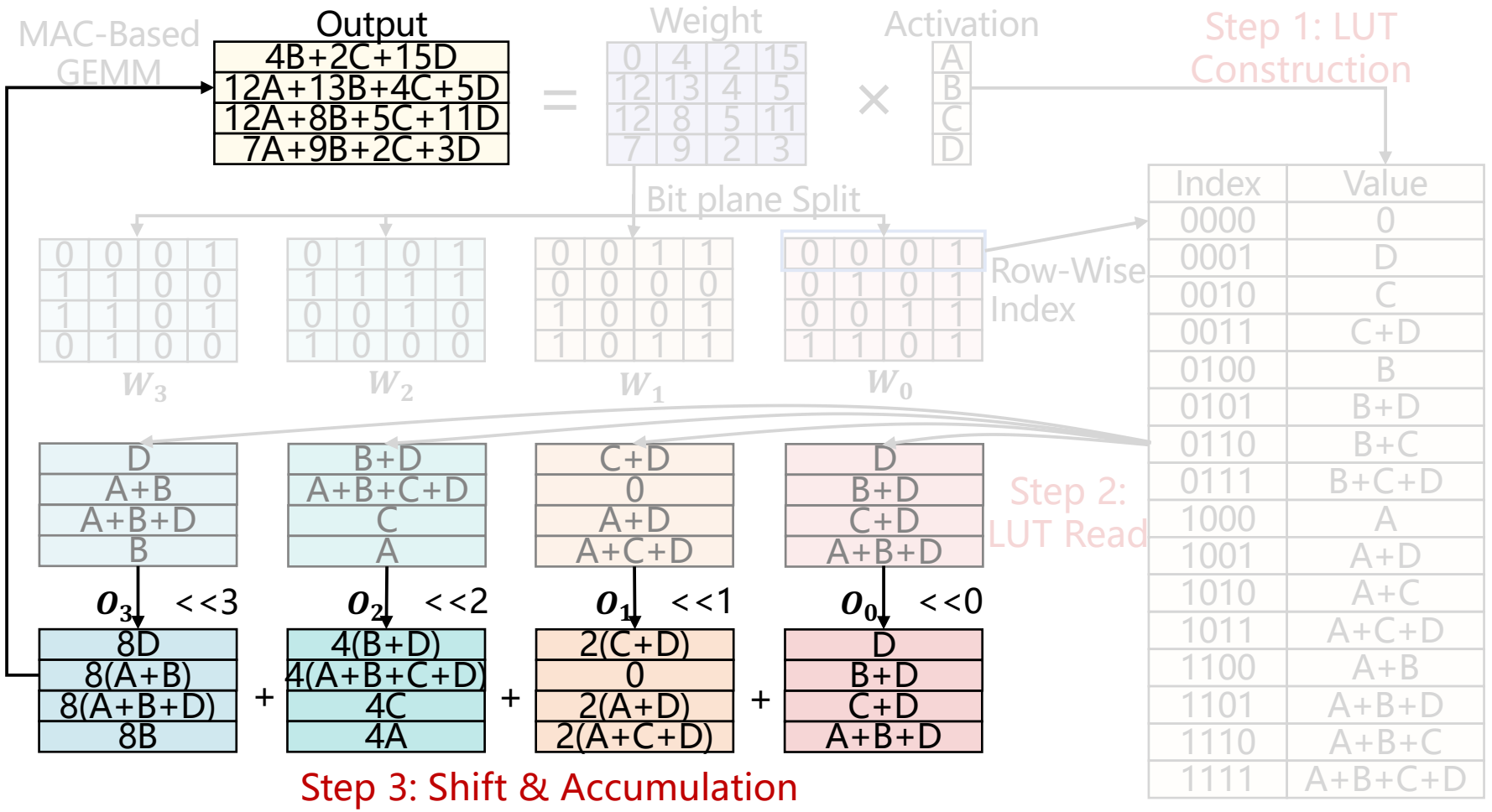
LUT Computing — Step 2: Bit-plane Lookup



Four adjacent bits in one bit plane form a shared LUT index.

Background and Motivation

LUT Computing — Step 3: Shift and Accumulation

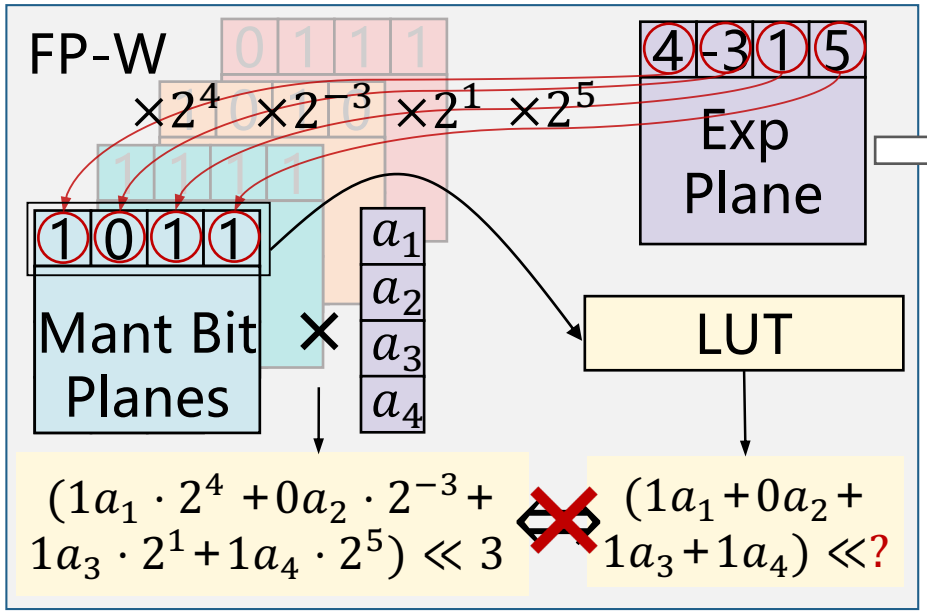


A shared bit significance allows one shift per bit plane.

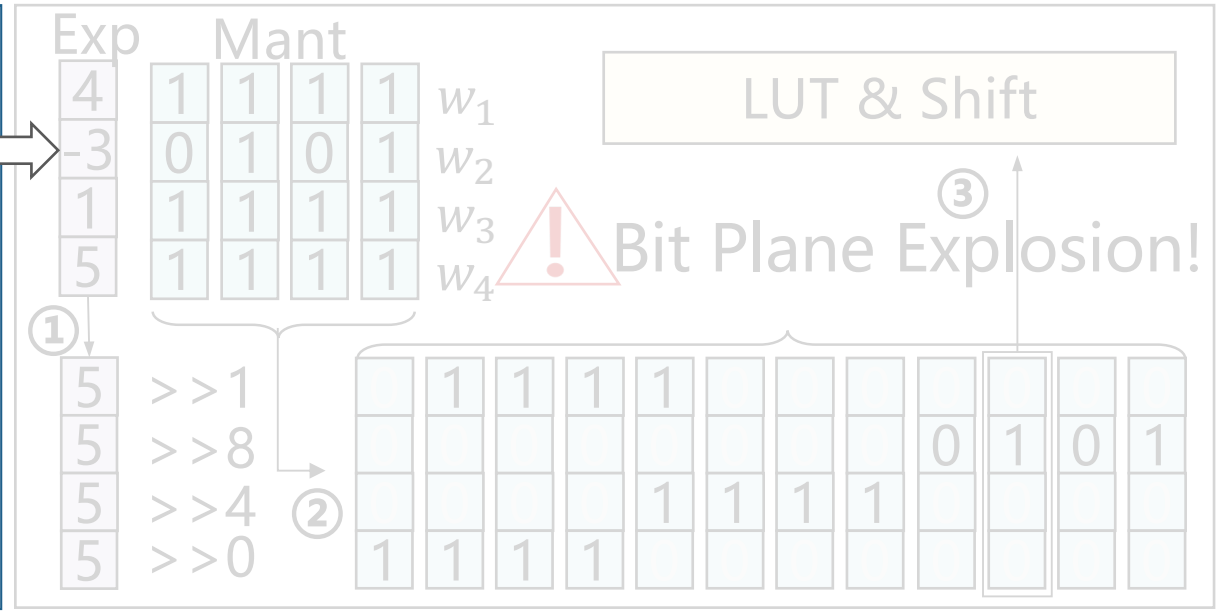
Background and Motivation

FP–LUT Mismatch

Independent exponents give the same mantissa bit different significance.



(b) Mismatch for FP Quantization.



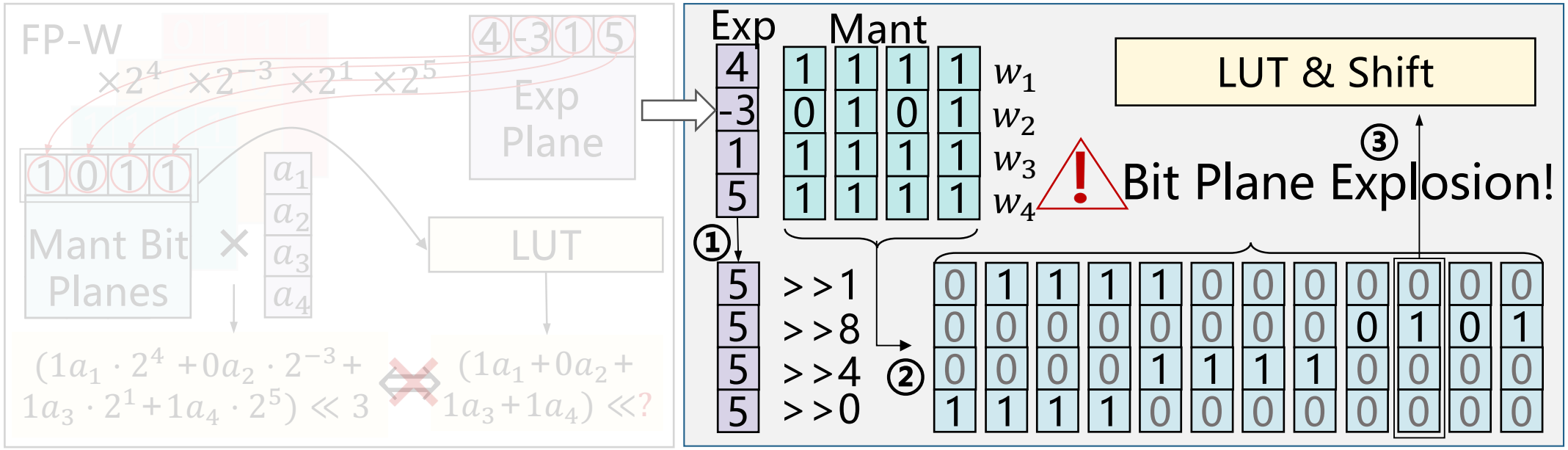
(c) Naïve Solution.

One shared shift cannot recover the FP result from a LUT lookup.

Background and Motivation

Naive Alignment: Bit-Plane Explosion

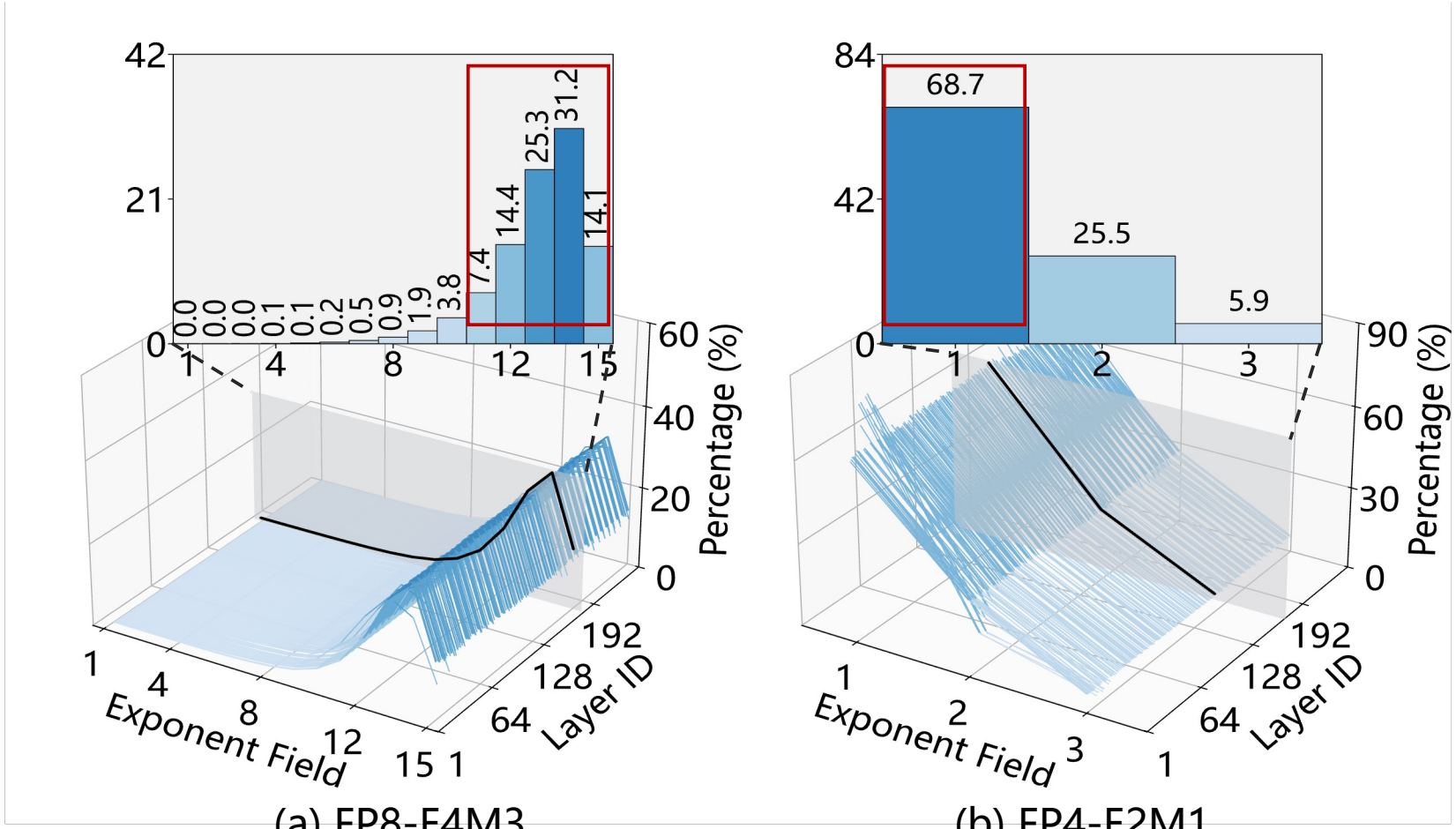
Naive FP8 alignment can require up to 19 bit planes.



How are exponents actually distributed in LLM weights?

Background and Motivation

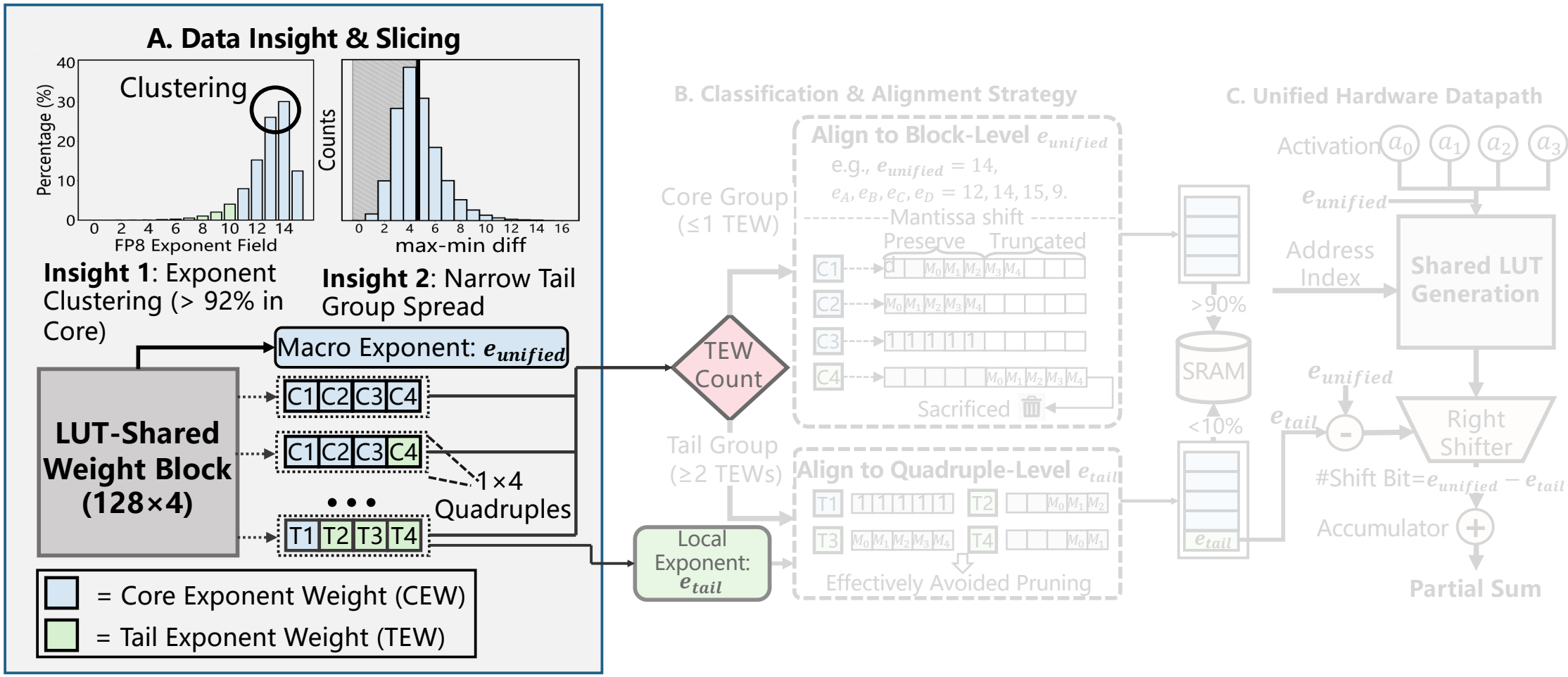
Exponent Clustering Effect



Most weights occupy a narrow core exponent range; only a small fraction forms the tail.

Core-Tail Hierarchical Exponent Alignment

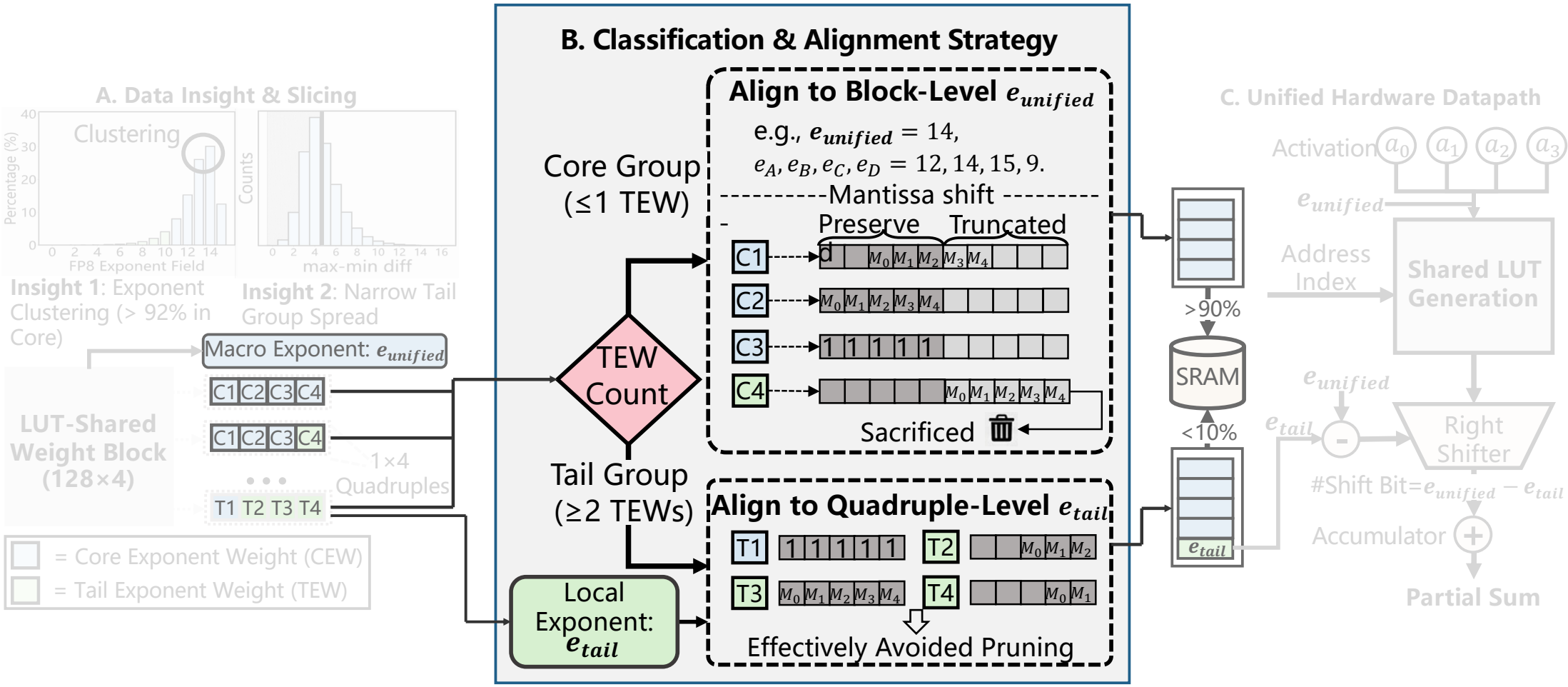
A. Data Insight and Quadruple Formation



Exploit both block-level exponent clustering and local tail structure.

Core-Tail Hierarchical Exponent Alignment

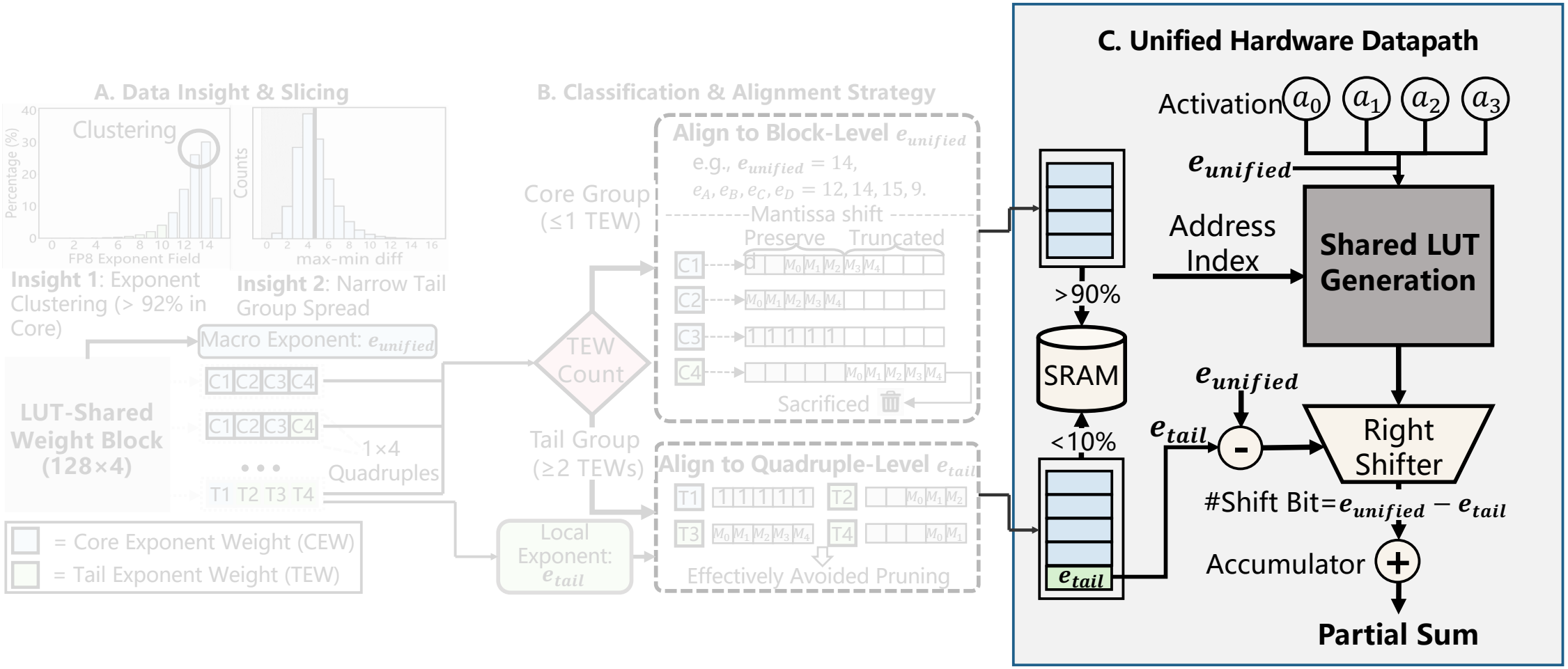
B. Classification and Alignment Strategy



Use local alignment where tail information would otherwise be lost.

Core-Tail Hierarchical Exponent Alignment

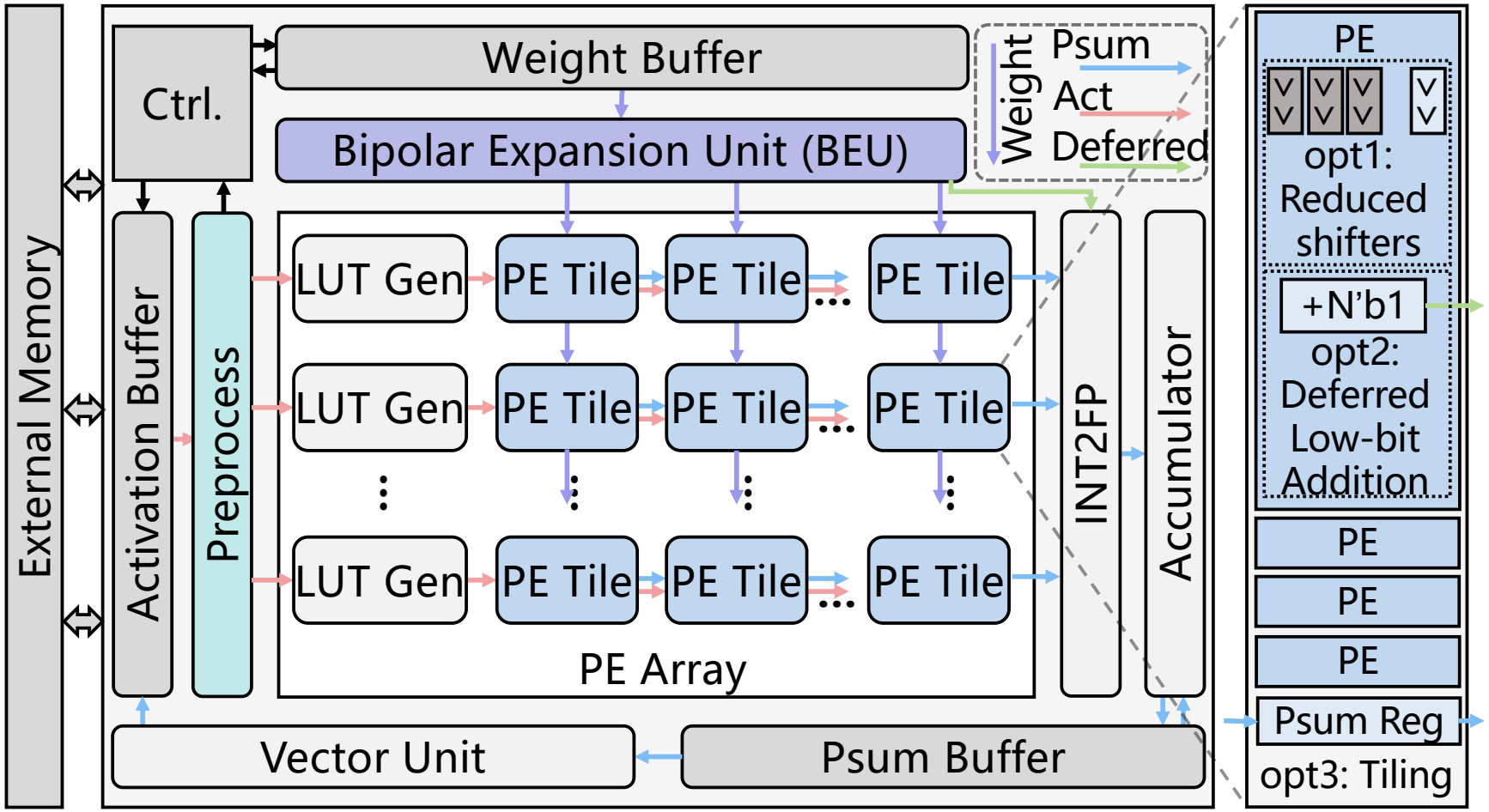
C. Unified LUT Execution



Compact core–tail alignment enables one unified LUT execution path.

The FLUTE Accelerator Architecture

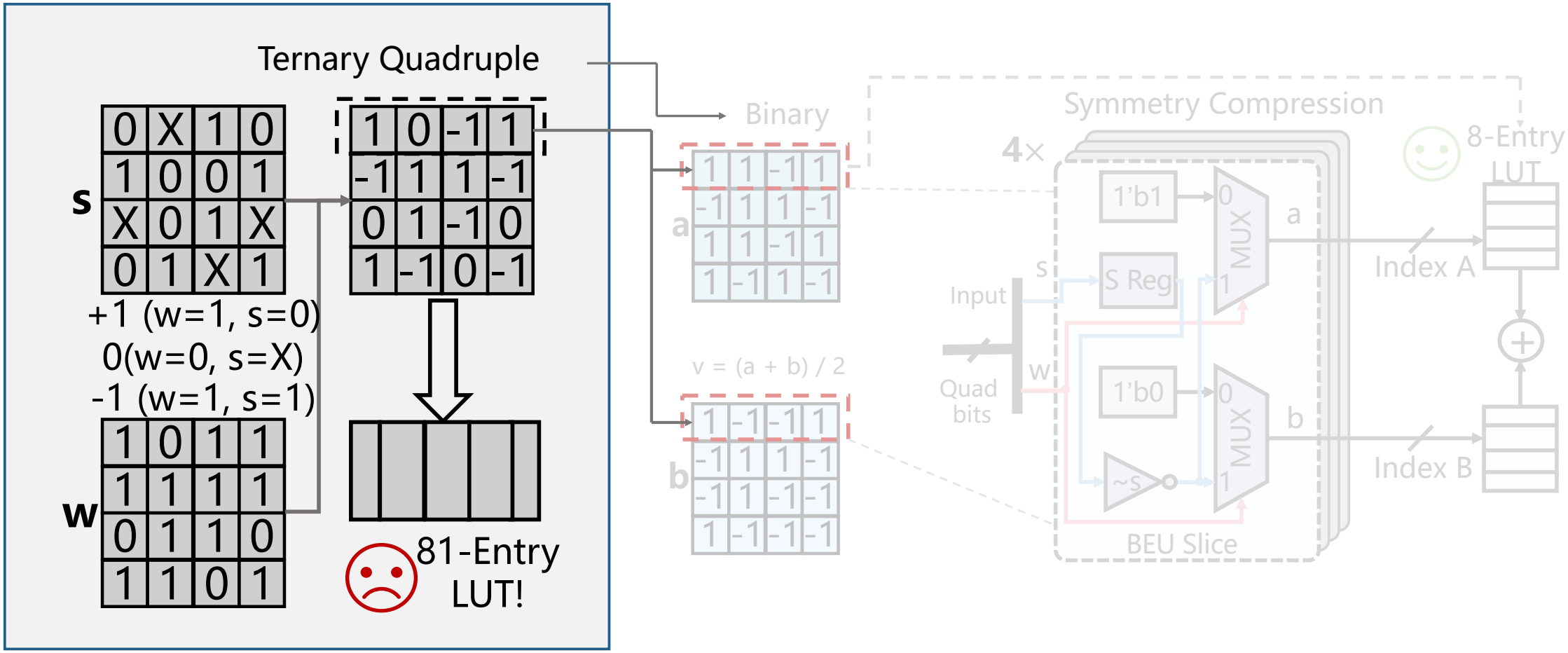
Architecture Overview



Activation LUTs and BEU-generated indices meet in the PE array.

The FLUTE Accelerator Architecture

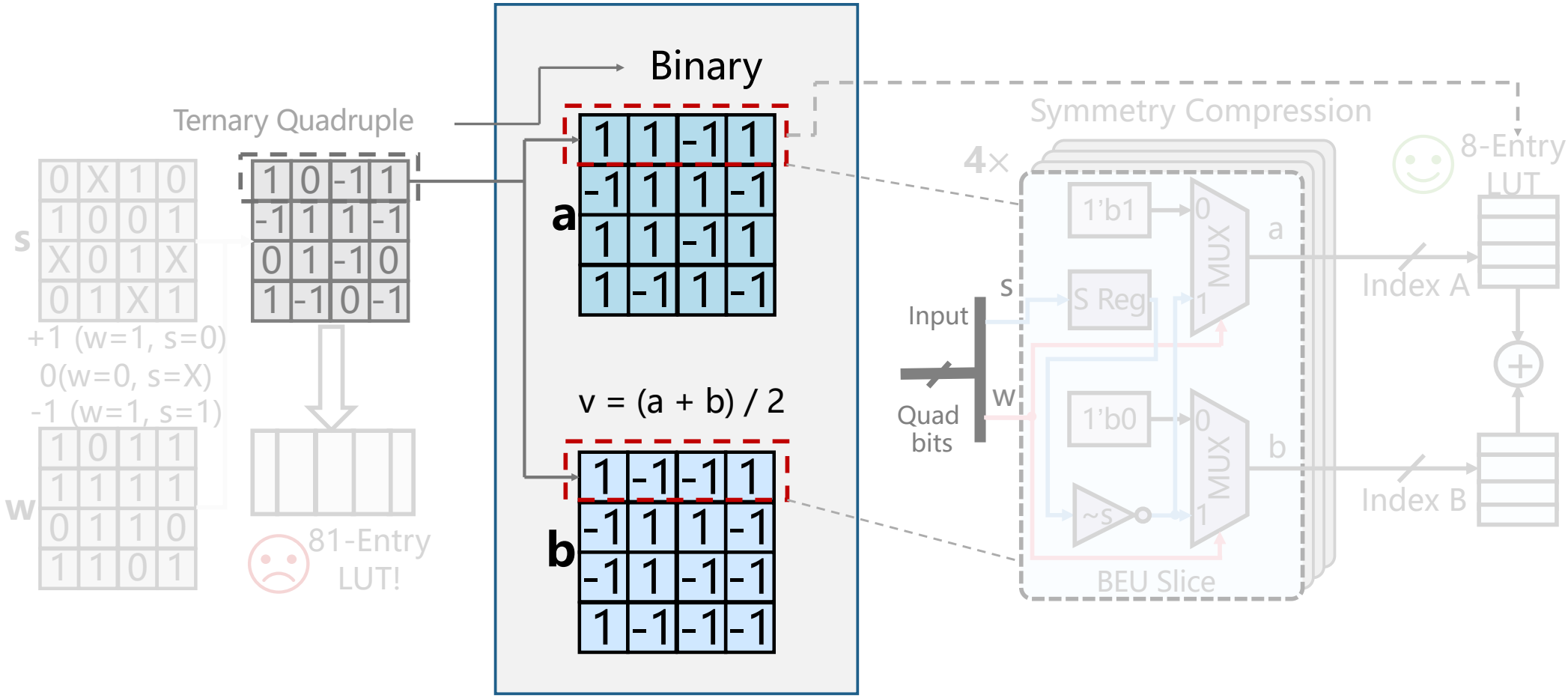
BEU Motivation: The Ternary Indexing Challenge



Direct ternary indexing undermines the benefit of a small LUT.

The FLUTE Accelerator Architecture

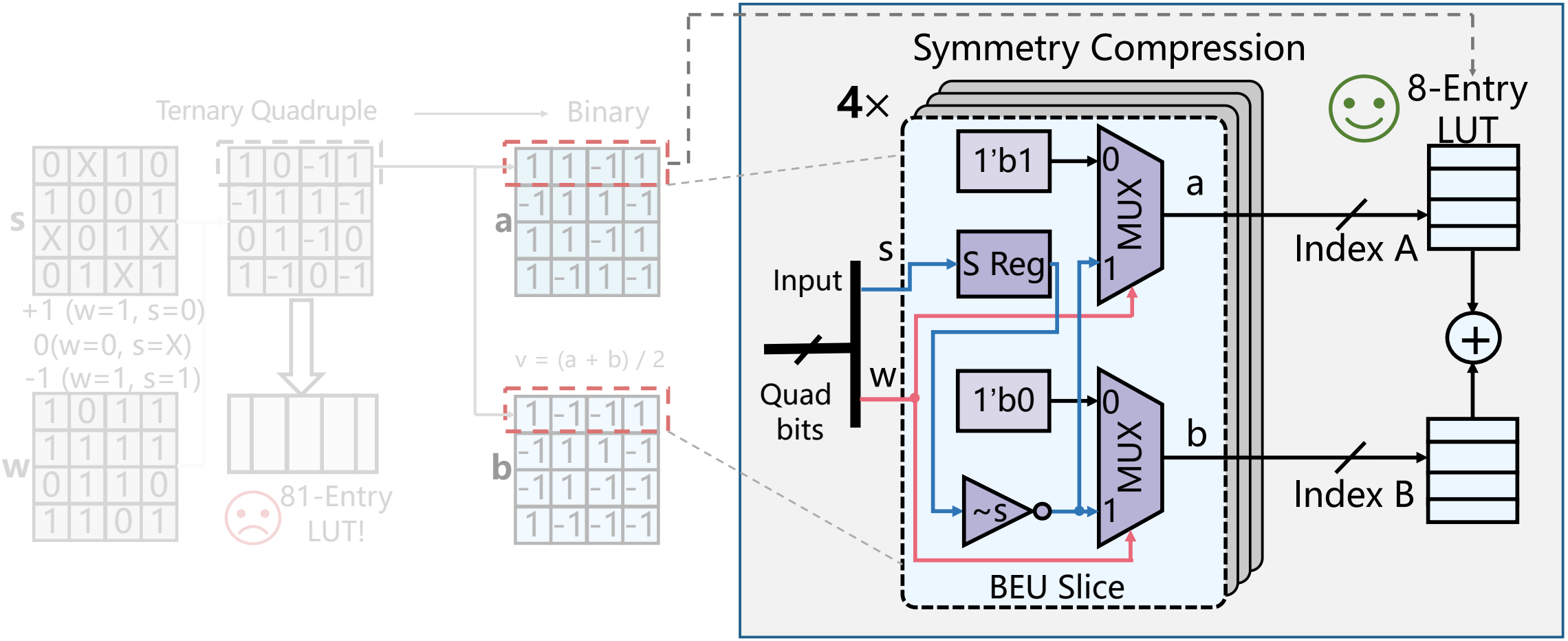
BEU: Bipolar Expansion



Two bipolar lookups reconstruct the original signed contribution.

The FLUTE Accelerator Architecture

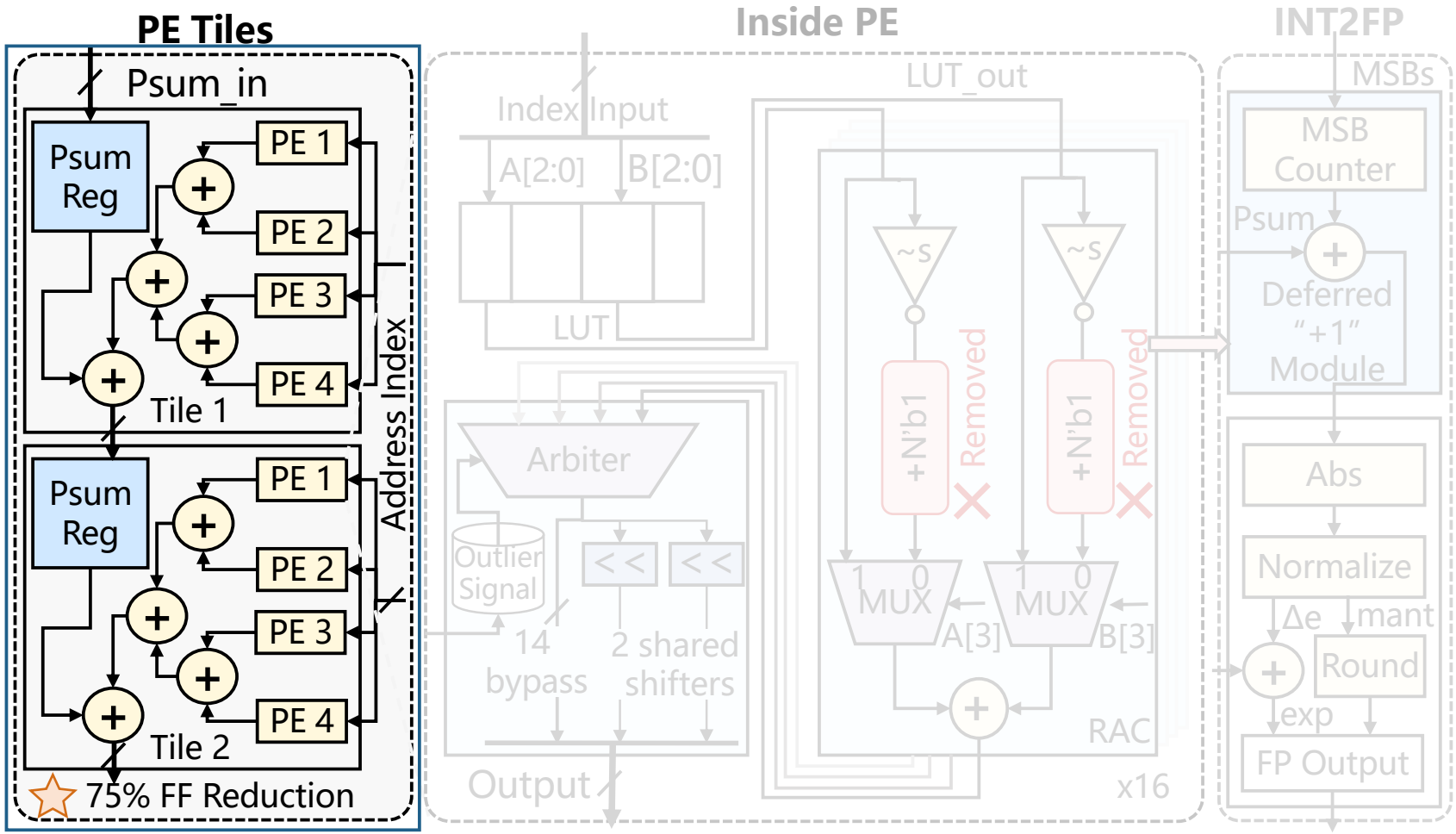
BEU: Symmetry Compression and 8-entry LUTs



BEU enables an 8-entry LUT implementation for the signed representation.

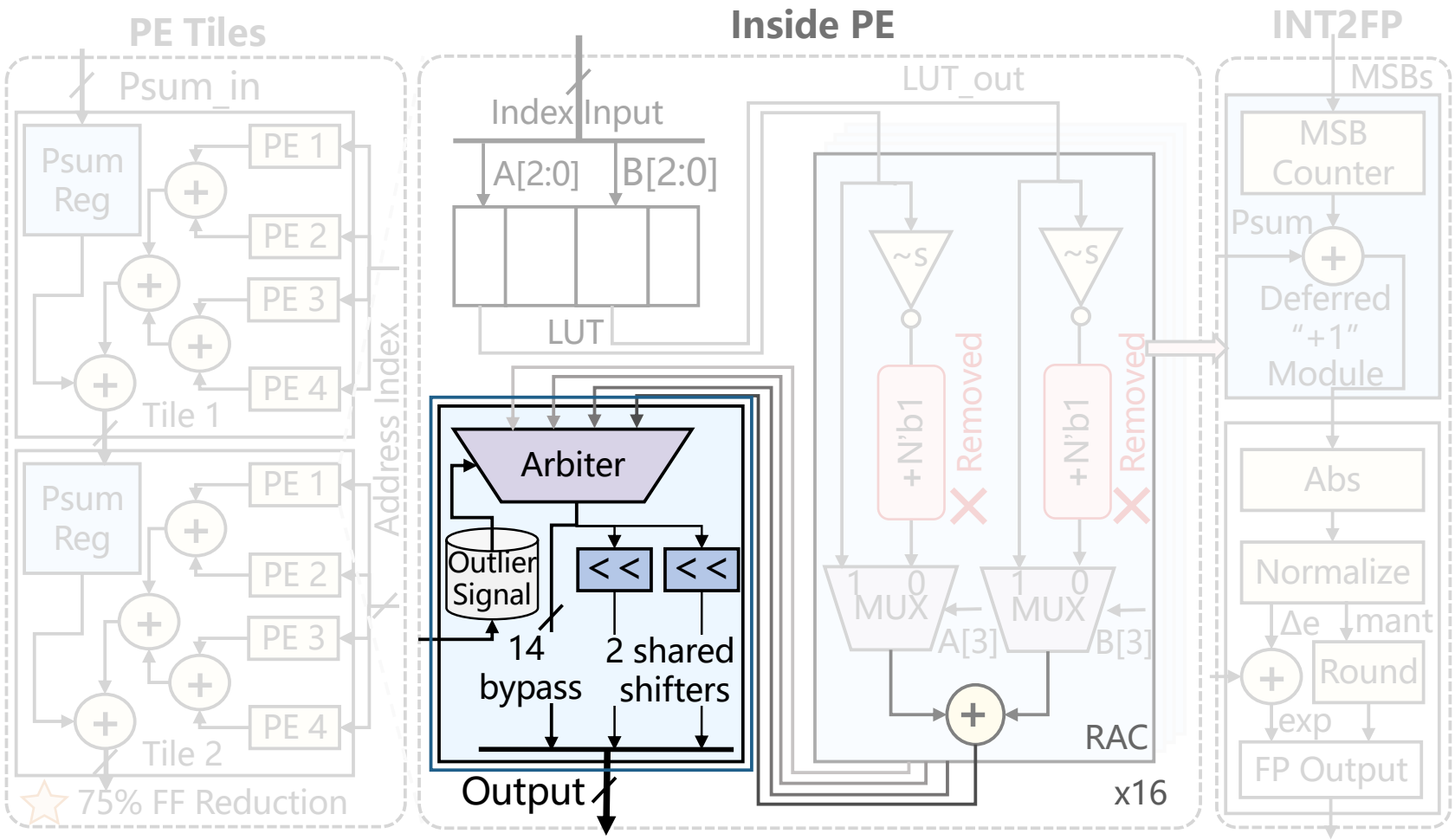
The FLUTE Accelerator Architecture

Spatial Accumulation: 4 PEs Share One Partial-Sum Register



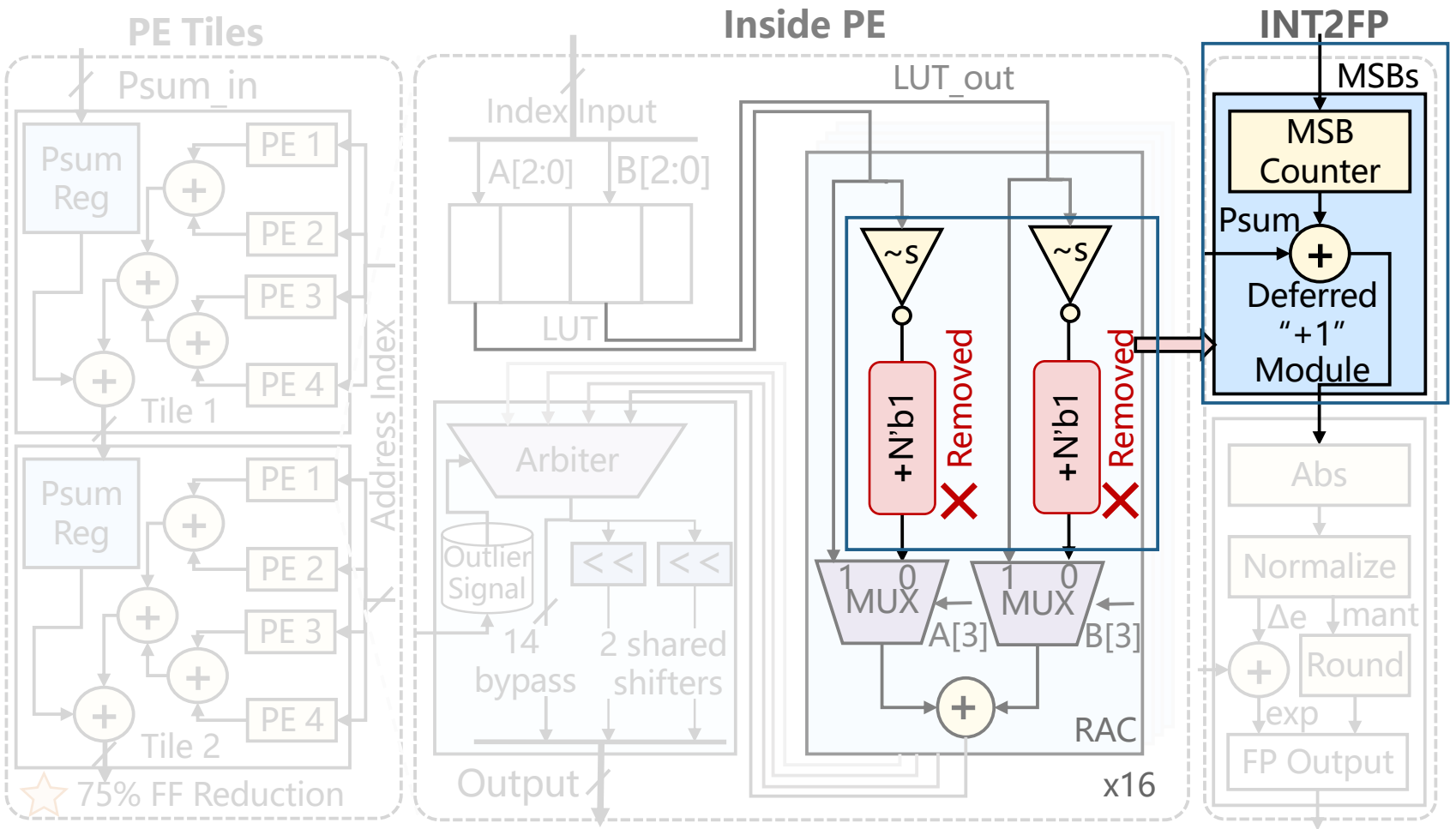
The FLUTE Accelerator Architecture

Shared Shifters: 16 RACs Share 2 Shifters for Sparse Tails



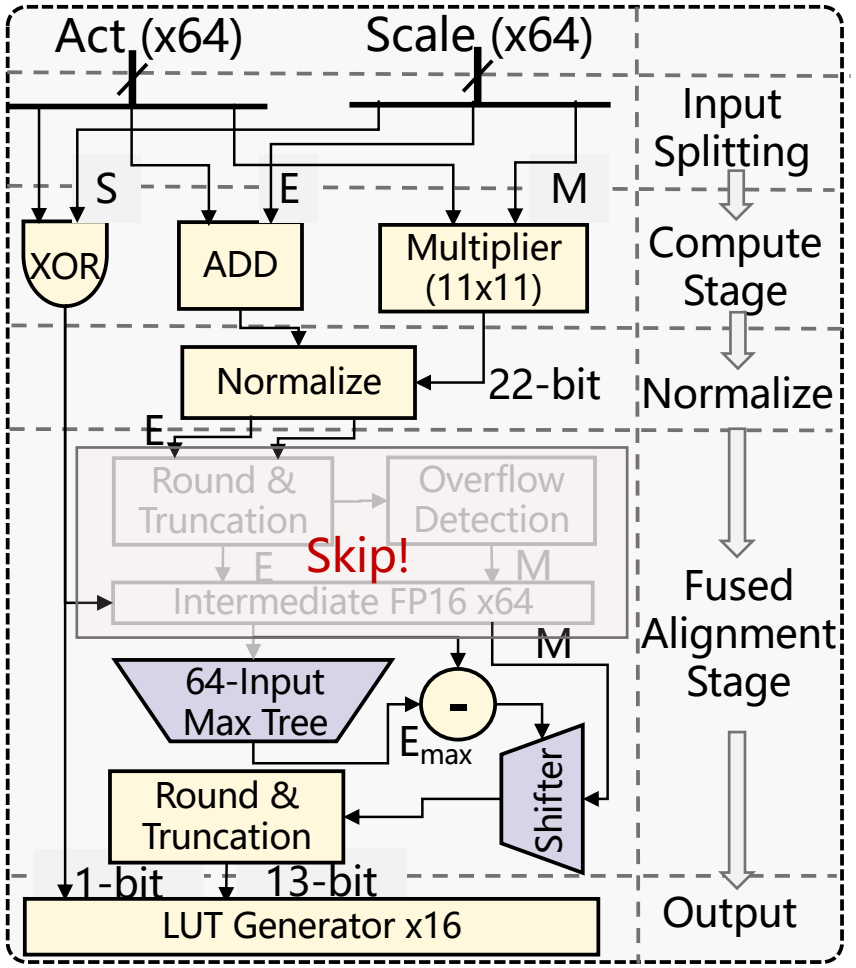
The FLUTE Accelerator Architecture

Deferred +1: Apply Sign Correction at the Row Output



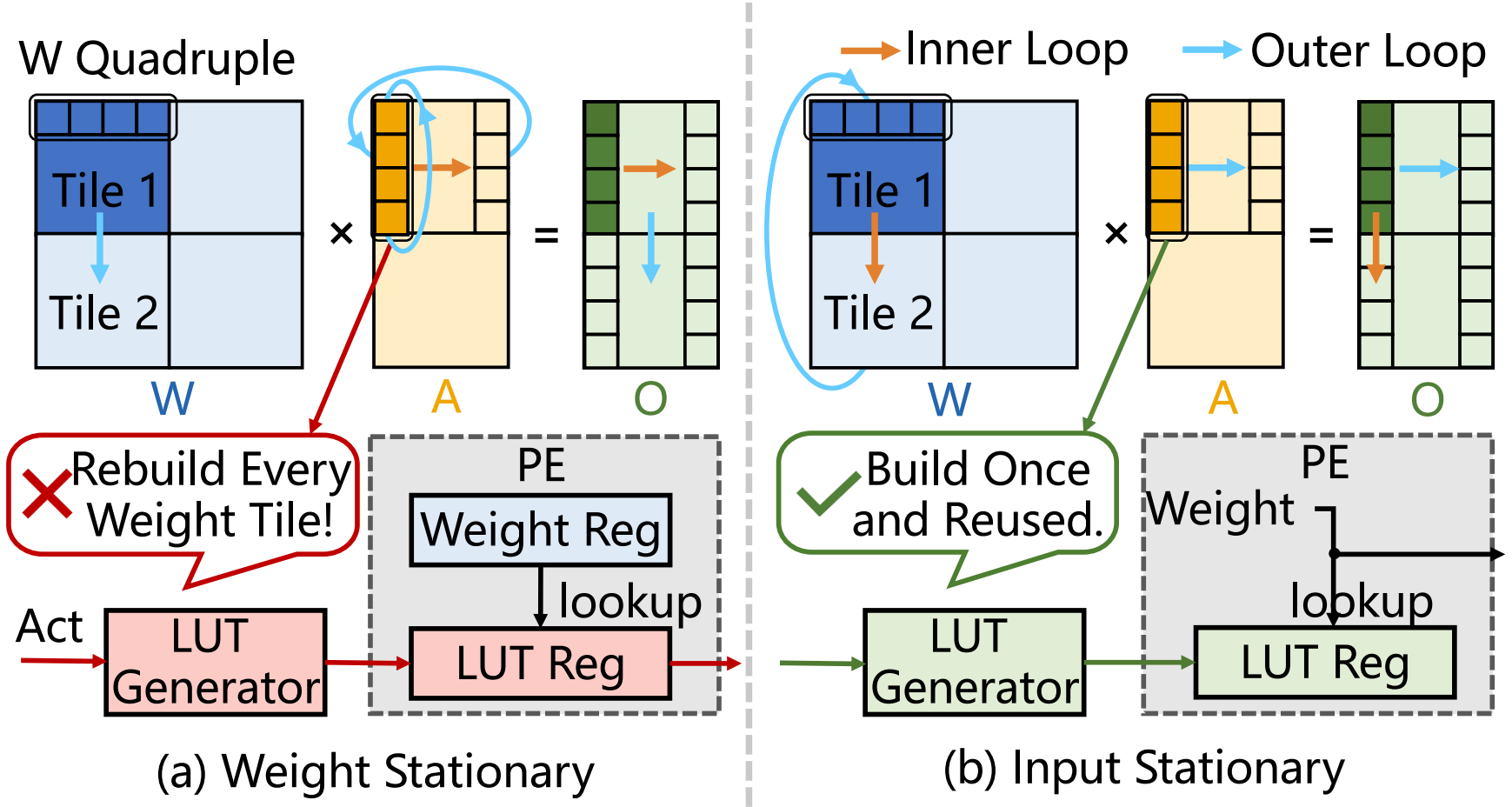
The FLUTE Accelerator Architecture

Fused Preprocessing: Directly Generate Aligned LUT Inputs



The FLUTE Accelerator Architecture

Weight-Stationary vs. Input-Stationary Dataflow



Input-stationary execution improves reuse of activation-derived LUTs.

The FLUTE Accelerator Architecture

LUT Sharing Granularity

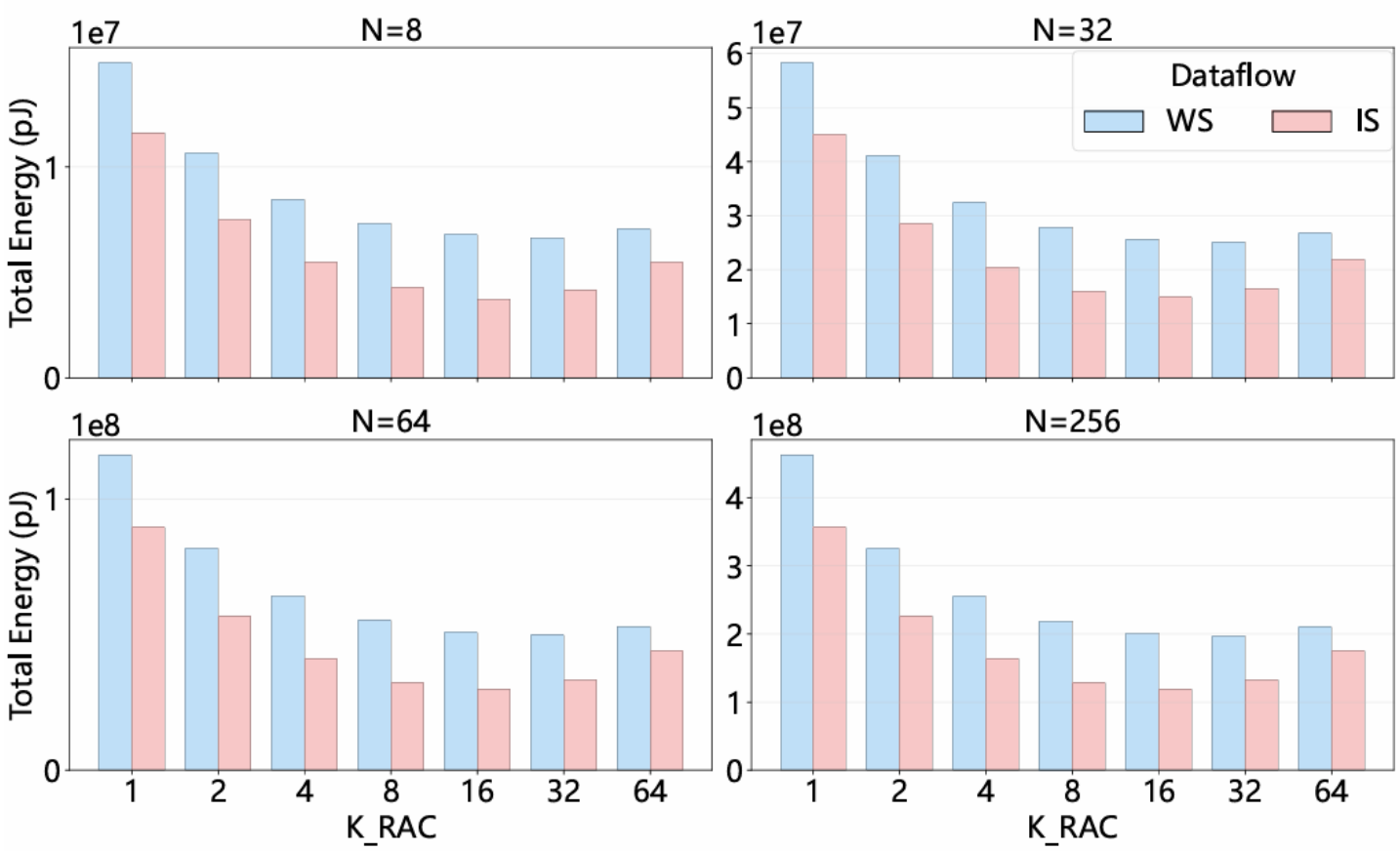


Fig. 9. Energy comparison across dataflows and K_{RAC}.

Best configuration: K_RAC = 16 and input-stationary dataflow.

Evaluation

Accuracy Evaluation: Perplexity

TABLE II

GEOMETRIC AVERAGE PERPLEXITY (PPL) ON GENERATIVE TASKS. THE EVALUATION SPANS 3-BIT, 4-BIT, AND 6-BIT TARGET PRECISIONS.

Models	FP16	3-bit						4-bit						6-bit					
		FP4	INT3	BFP3	BBAL	BiE	**Ours**	FP6	INT4	BFP4	BBAL	BiE	**Ours**	FP8	INT6	BFP6	BBAL	BiE	**Ours**
OPT-6.7B	11.88	12.76	1146	2068	8102	8815	**12.85**	11.94	12.71	13.60	14.98	174.6	**12.09**	11.89	11.90	11.91	11.92	11.93	**11.88**
OPT-13B	11.16	12.00	531.1	706.2	8528	3458	**12.34**	11.20	11.72	12.77	13.68	20.92	**11.29**	11.17	11.19	11.20	11.20	11.21	**11.16**
OPT-30B	10.57	11.15	760.8	1071	6988	3587	**11.28**	10.60	11.02	12.05	12.66	35.39	**10.67**	10.57	**10.57**	10.63	10.59	10.61	**10.57**
LLaMA-2-7B	9.24	11.14	22.59	25.21	nan	9542	**14.02**	9.39	10.32	10.21	10.78	17.08	**9.54**	9.23	9.31	9.39	9.31	9.29	**9.27**
LLaMA-2-13B	9.68	10.63	13.35	14.79	2398	7840	**11.23**	9.80	10.00	10.11	11.04	10.93	**9.70**	9.69	9.69	9.73	9.71	9.69	**9.67**
LLaMA-2-70B	6.58	6.87	8.09	10.96	215.1	2281	**7.54**	6.59	6.79	6.99	7.65	7.75	**6.67**	6.58	6.58	6.62	6.60	6.61	**6.57**
Qwen3-8B	12.93	14.44	31.00	96.38	nan	4211	**16.71**	13.18	13.74	14.86	15.59	20.57	**13.55**	12.91	12.92	12.92	12.97	13.03	**12.82**
Qwen3-30B-A3B	12.02	13.70	141.7	592.6	nan	5516	**17.59**	12.15	12.98	18.51	28.74	176.2	**12.55**	12.10	12.06	12.07	12.11	12.08	**11.93**

6-bit FLUTE: near-FP16 perplexity across the evaluated models.

4-bit LLaMA-2-7B: PPL 9.54 vs. 9.24 for FP16.

Evaluation

Accuracy Evaluation: Zero-shot Reasoning

TABLE III
ZERO-SHOT ACCURACY ON DOWNSTREAM REASONING TASKS ACROSS DIVERSE LLMS.

Bit	Tasks	Qwen3-8B							OPT-13B							OPT-30B						
		FP16	FP	INT	BFP	BBAL	BiE	**Ours**	FP16	FP	INT	BFP	BBAL	BiE	**Ours**	FP16	FP	INT	BFP	BBAL	BiE	**Ours**
3-bit	Avg	0.723	0.702	0.501	0.371	0.317	0.322	**0.623**	0.629	0.606	0.349	0.340	0.307	0.307	**0.621**	0.658	0.640	0.377	0.368	0.313	0.305	**0.641**
4-bit	ARC-C	0.565	0.573	0.538	0.509	0.514	0.433	**0.562**	0.354	0.349	0.354	0.340	0.335	0.307	**0.355**	0.385	0.387	0.370	0.346	0.333	0.321	**0.373**
	ARC-E	0.811	0.813	0.768	0.719	0.717	0.679	**0.784**	0.612	0.609	0.610	0.600	0.586	0.529	**0.618**	0.653	0.651	0.636	0.619	0.604	0.578	**0.648**
	BoolQ	0.867	0.865	0.846	0.830	0.831	0.741	**0.858**	0.666	0.671	0.662	0.666	0.664	0.660	**0.691**	0.716	0.709	0.673	0.681	0.691	0.650	**0.729**
	Hella	0.750	0.747	0.730	0.701	0.691	0.636	**0.738**	0.699	0.698	0.683	0.664	0.637	0.588	**0.695**	0.722	0.720	0.710	0.700	0.680	0.573	**0.710**
	Lambada	0.611	0.612	0.556	0.530	0.531	0.386	**0.587**	0.635	0.630	0.610	0.587	0.576	0.462	**0.635**	0.650	0.645	0.614	0.613	0.609	0.408	**0.614**
	PIQA	0.777	0.780	0.760	0.762	0.755	0.733	**0.767**	0.767	0.769	0.761	0.757	0.748	0.733	**0.761**	0.787	0.790	0.782	0.772	0.767	0.748	**0.782**
	Wino	0.680	0.683	0.664	0.639	0.650	0.591	**0.665**	0.668	0.665	0.665	**0.669**	**0.671**	0.638	0.657	0.697	0.693	0.676	0.639	0.652	0.634	**0.676**
	Avg	0.723	0.725	0.694	0.670	0.670	0.600	**0.709**	0.629	0.627	0.621	0.612	0.603	0.560	**0.630**	0.658	0.656	0.637	0.624	0.619	0.559	**0.637**
6-bit	Avg	0.723	0.722	0.718	0.721	0.719	0.715	**0.722**	0.629	0.629	0.628	0.627	0.628	0.629	**0.630**	0.658	0.658	0.658	0.656	0.657	0.655	**0.659**

Bit	Tasks	LLaMA-2-7B							LLaMA-2-13B							LLaMA-2-70B						
		FP16	FP	INT	BFP	BBAL	BiE	**Ours**	FP16	FP	INT	BFP	BBAL	BiE	**Ours**	FP16	FP	INT	BFP	BBAL	BiE	**Ours**
3-bit	Avg	0.701	0.693	0.613	0.567	0.327	0.316	**0.642**	0.729	0.720	0.666	0.624	0.320	0.315	**0.694**	0.777	0.764	0.740	0.680	0.379	0.309	**0.758**
4-bit	ARC-C	0.450	0.451	0.448	**0.454**	0.442	0.422	0.448	0.489	0.485	0.482	**0.486**	0.469	0.449	0.477	0.573	0.566	0.550	**0.561**	0.550	0.521	0.549
	ARC-E	0.739	0.742	0.725	0.712	0.714	0.677	**0.733**	0.765	0.763	0.758	0.761	0.751	0.719	**0.766**	0.796	0.796	0.785	0.795	0.792	0.780	**0.801**
	BoolQ	0.794	0.788	0.776	0.756	0.745	0.742	**0.783**	0.821	0.823	0.818	0.814	0.819	0.793	**0.821**	0.852	0.848	0.840	0.823	0.817	0.811	**0.851**
	Hella	0.762	0.758	0.755	0.749	0.736	0.722	**0.760**	0.796	0.795	**0.789**	0.784	0.766	0.765	0.783	0.844	0.841	**0.835**	0.824	0.809	0.795	0.831
	Lambada	0.688	0.687	0.672	0.647	0.633	0.588	**0.678**	0.706	0.713	0.692	0.691	0.687	0.663	**0.693**	0.748	0.750	0.738	0.738	0.737	0.701	**0.761**
	PIQA	0.788	0.786	0.782	0.780	0.784	0.770	**0.789**	0.802	0.808	0.803	0.796	0.797	0.785	**0.804**	0.826	0.823	0.820	0.819	0.817	0.815	**0.824**
	Wino	0.694	0.695	0.692	0.690	0.679	0.680	**0.684**	0.723	0.718	0.716	0.715	0.708	0.703	**0.734**	0.804	0.811	0.792	0.779	0.779	0.769	**0.800**
	Avg	0.702	0.701	0.693	0.684	0.676	0.657	**0.697**	0.729	0.729	0.723	0.721	0.714	0.697	**0.725**	0.778	0.776	0.765	0.763	0.757	0.742	**0.774**
6-bit	Avg	0.702	0.700	0.700	0.701	0.700	0.698	**0.701**	0.729	0.729	0.729	0.728	0.728	0.729	**0.731**	0.778	0.777	0.776	0.774	0.775	0.776	**0.777**

4-bit LLaMA-2-7B: 69.7% vs. 70.2% — a 0.5 percentage-point gap.

Evaluation

Performance Comparison: GEMM and End-to-End Decode

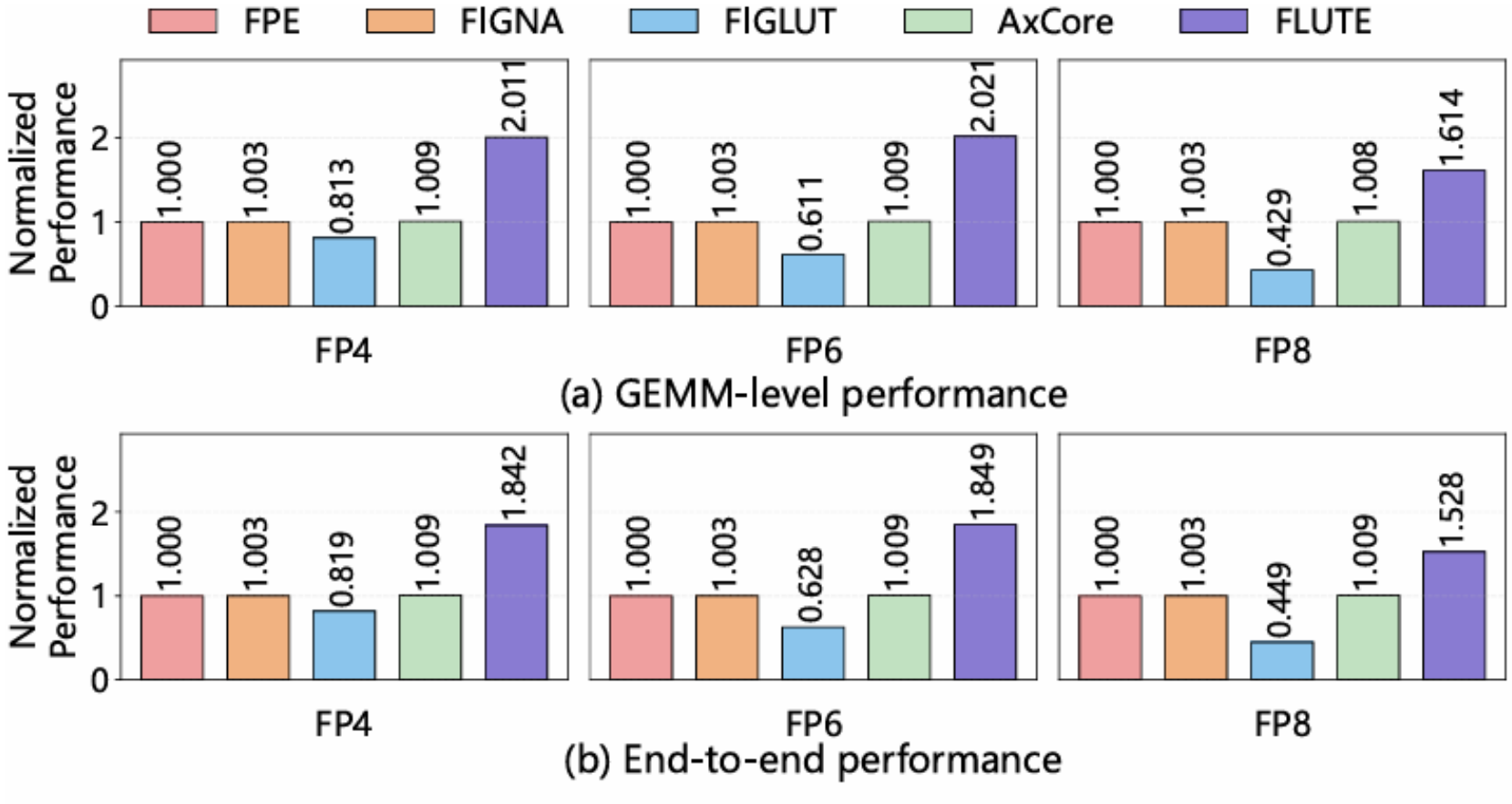


Fig. 11. Normalized performance of various architectures.

Evaluation

Area Efficiency

Average gain: 6.32× / 2.97× / 2.66× / 1.62× over FPE / FIGNA / FIGLUT / AxCore.

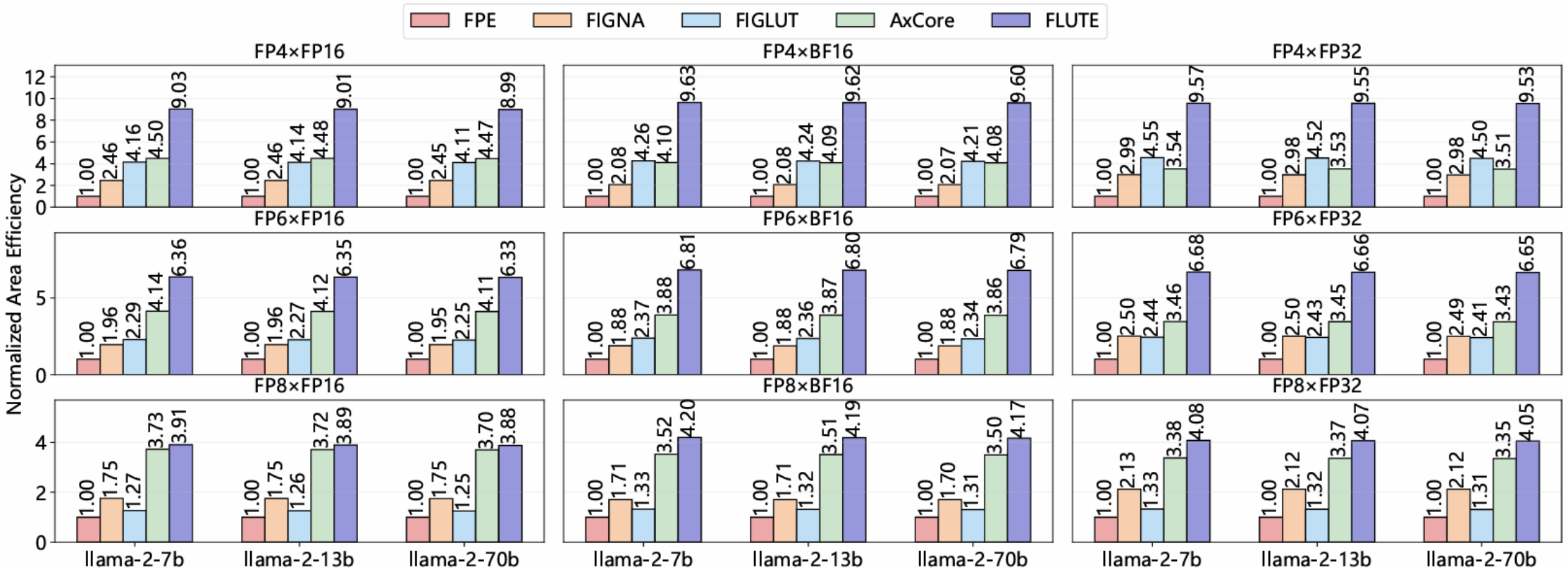


Fig. 10. Normalized area efficiency (Performance/Area) of the evaluated architectures under diverse weight precisions.

Compact bit-planes and PE optimizations improve performance per unit area.

Evaluation

Compute Energy Efficiency

Average gain: 6.01× / 4.02× / 3.35× / 2.38× over FPE / FIGNA / FIGLUT / AxCore.

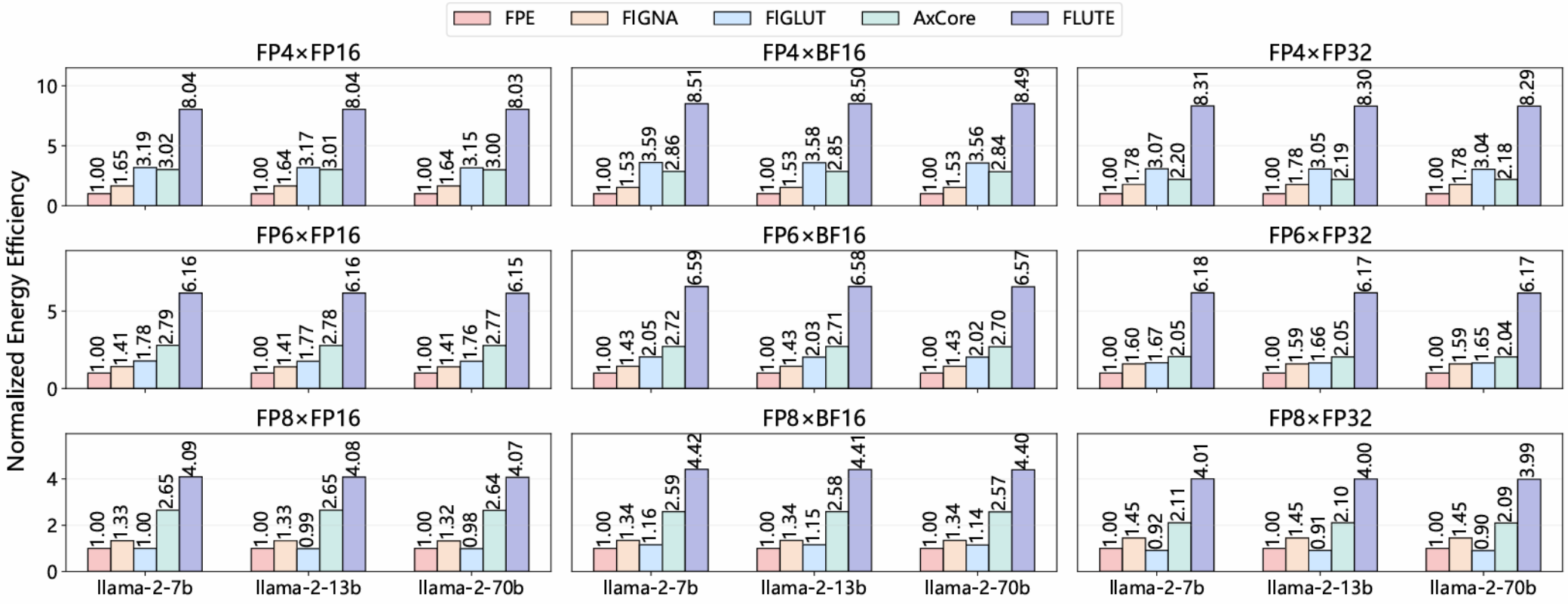


Fig. 12. Normalized compute energy efficiency across diverse models and precision configurations.

Compact bit-planes and lightweight BEU logic reduce dynamic switching.

Evaluation

中国科学院 自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

PE Optimization Ablation

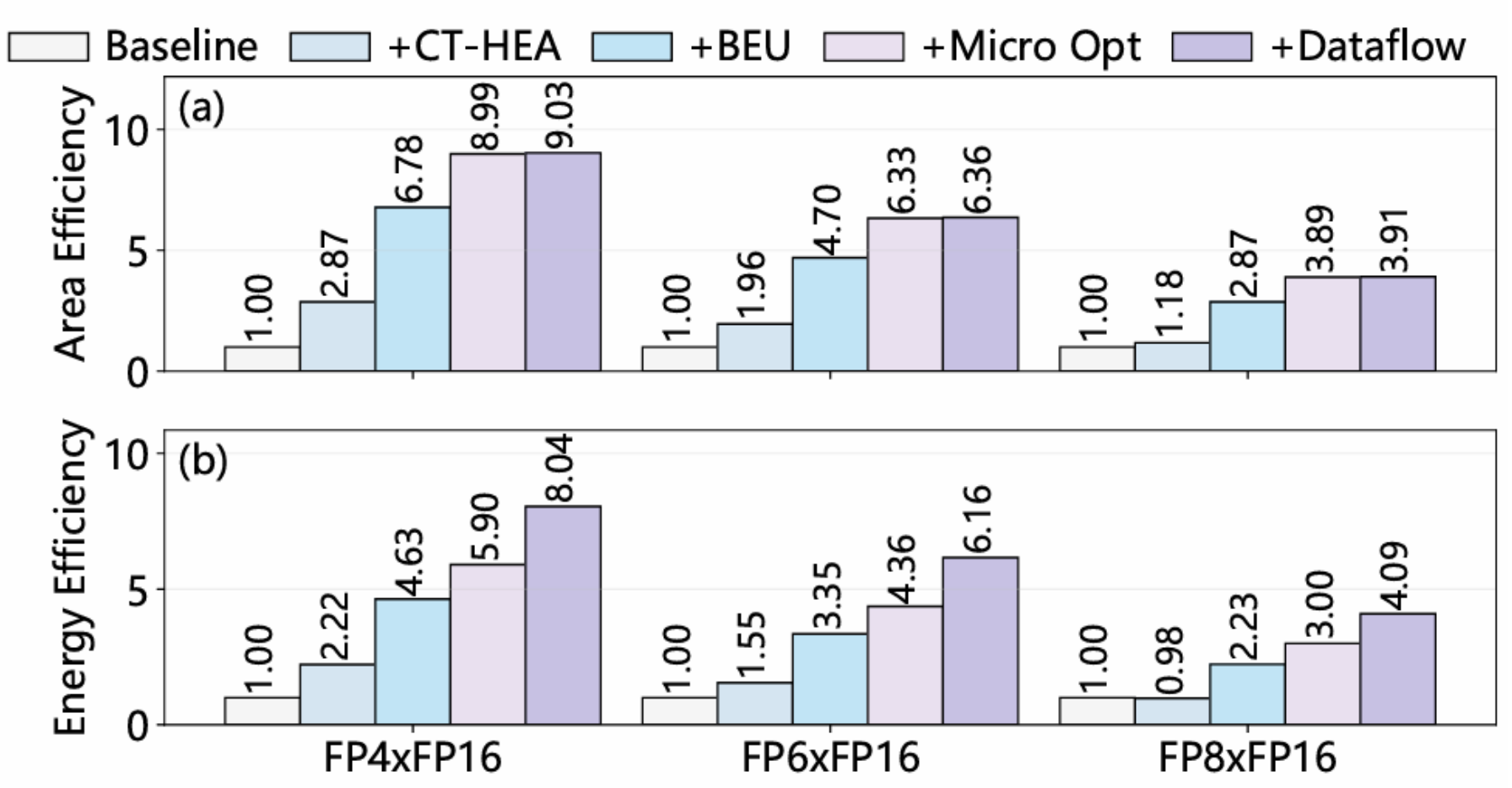
PE OPTIMIZATION ABLATION: AREA (μm^2) AND POWER (MW), NORMALIZED TO THE 81-ENTRY PE.

Design	Area	Area Eff.	Power	Power Eff.
81-entry PE	16338	1.00x	4.085	1.00x
8-entry PE	6455	2.53x	1.706	2.39x
+ Shared shifters	5440	3.00x	1.606	2.54x
+ Deferred "+1"	5203	3.14x	1.578	2.59x
+ Tiling	4646	3.52x	1.087	3.76x

The optimized PE uses less than 30% of the baseline area and power.

Evaluation

System-level Benefit Breakdown



Algorithm, BEU, microarchitecture, and dataflow contribute complementary benefits.

Q&A

中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

Thanks! Any Questions?