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

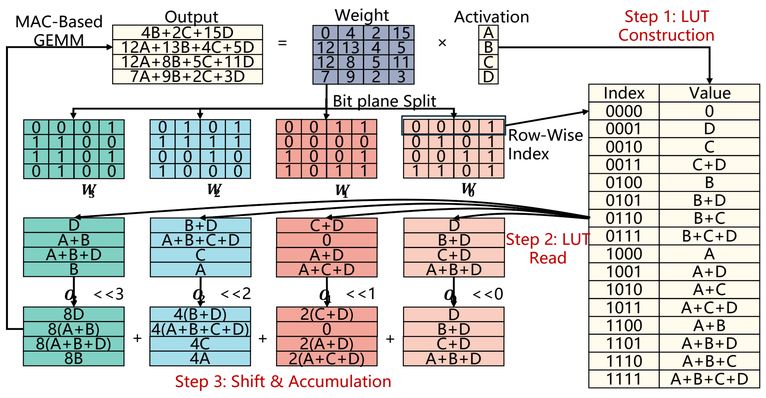
Tielong Liu[1,2,*], Zeyu Zhu[1,2,*], Gang Li[1,†], Zitao Mo[1], Chen Bai[3], Junwen Si[1]
Peihuan Ni[1], Minnan Pei[1], Peisong Wang[1], Zhuoran Song[4], Xiaoyao Liang[4], Jian Cheng[1,†]
[1] C²DL, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Future Technology, University of Chinese Academy of Sciences, Beijing, China
[3] Fudan University, Shanghai, China | [4] Shanghai Jiao Tong University, Shanghai, China
* Equal contribution. † Corresponding authors.

Introduction

Low-bit floating-point (FP) quantization preserves LLM accuracy, but naïve exponent alignment expands bit-planes and undermines efficient LUT execution.
FLUTE exploits exponent clustering through **Core-Tail Hierarchical Exponent Alignment (CT-HEA)**, producing compact weights with shared and local exponents.
An **8-entry LUT**, optimized PEs and input-stationary dataflow turn this representation into faster decoding and higher area and compute energy efficiency.

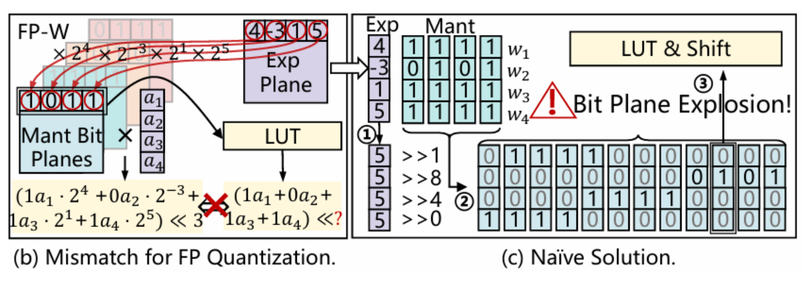
Background and Motivation

LUTs reuse activation sums



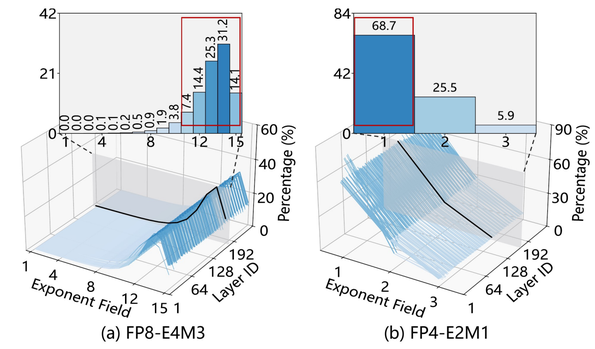
FP alignment expands bit-planes

FP exponents break common bit significance.
Naïve FP8 alignment: up to 19 bit-planes.



(b) Mismatch for FP Quantization. (c) Naïve Solution.

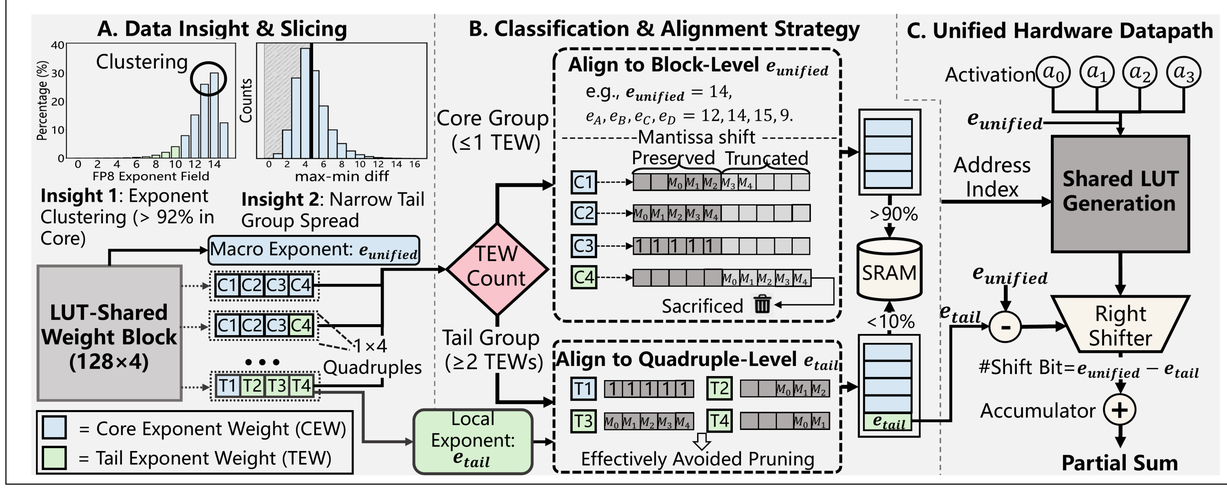
A narrow core covers >92% (FP8)



(a) FP8-E4M3 (b) FP4-E2M1

Most FP8 weights occupy a **narrow exponent core**.
Local alignment retains information in **sparse tail groups**.

Core-Tail Hierarchical Exponent Alignment



Offline alignment

1. Find the densest core window in each 128 × 4 block.
2. 0 / 1 tail: use the block exponent; truncate a lone tail.
3. ≥2 tails: use a local exponent; compensate the LUT output by shifting.

FP8 → 6-bit payload: 1 sign + 5 mantissa bits, plus exponent/flag metadata.

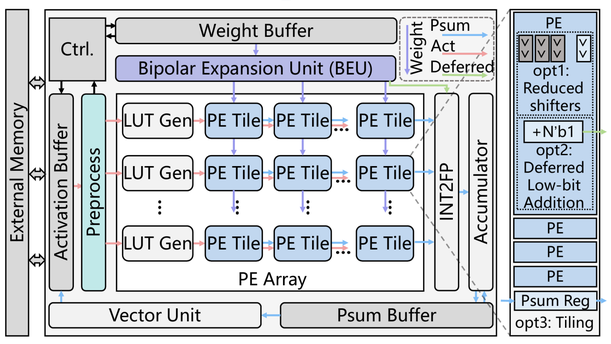
Algorithm 1 CT-HEA Encoding for One LUT-Shared Weight Block

Input: FP-quantized weights $\mathcal{W}$; core window width r.
Output: Aligned payloads $\mathcal{P}$; block exponent e_{unified}; tail flags $\mathcal{F}_{\text{tail}}$; tail metadata $\mathcal{M}_{\text{tail}}$.

Extract (s_i, e_i, m_i) for each $w_i \in \mathcal{W}$.
Find the width-r exponent window $[e_{\text{low}}, e_{\text{high}}]$ covering the most weights.
Select e_{unified} from $[e_{\text{low}}, e_{\text{high}}]$ and split $\mathcal{W}$ into quadruples G_j.
for each quadruple G_j **do**
  $\mathcal{T}_j \leftarrow w_i \in G_j \mid e_i \notin [e_{\text{low}}, e_{\text{high}}]$.
  if $|\mathcal{T}_j| = 0$ **then**
    Align G_j to e_{unified}; $\mathcal{F}_{\text{tail}}[j] \leftarrow 0$.
  else if $|\mathcal{T}_j| = 1$ **then**
    Align core weights to e_{unified} and truncate the single TEW; $\mathcal{F}_{\text{tail}}[j] \leftarrow 0$.
  else
    $e_{\text{tail}}[j] \leftarrow \max_{w_i \in G_j} e_i$; align G_j to $e_{\text{tail}}[j]$.
    Store $e_{\text{tail}}[j]$ in $\mathcal{M}_{\text{tail}}$; $\mathcal{F}_{\text{tail}}[j] \leftarrow 1$.
  end if
end for
return $\mathcal{P}$, e_{unified}, $\mathcal{F}_{\text{tail}}$, and $\mathcal{M}_{\text{tail}}$.

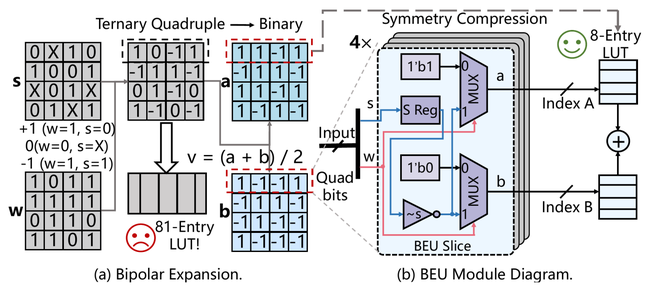
The FLUTE Accelerator Architecture

Unified LUT datapath



Activations → LUTs → shared PE array.
One datapath for core and tail weights.

BEU: 81 → 8 LUT entries

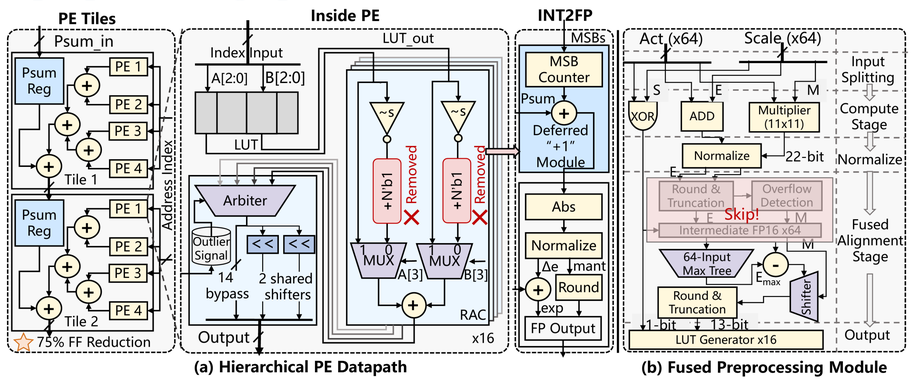


(a) Bipolar Expansion. (b) BEU Module Diagram.

v = (a + b) / 2, a, b ∈ {−1, +1}

Two parallel bipolar lookups recover each ternary query. Sign symmetry reduces storage to 8 entries.

PE and preprocessing optimizations



(a) Hierarchical PE Datapath (b) Fused Preprocessing Module

4-PE tiles: 75% fewer psum registers; 16 RACs share 2 shifters.

Evaluation

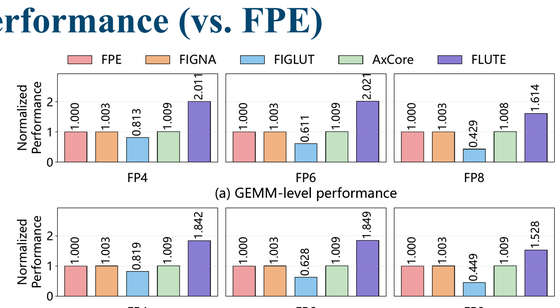
Accuracy: zero-shot reasoning on LLaMA-2 (TABLE III)

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**

Experimental setup

Models: OPT, LLaMA-2 and Qwen3.
PPL + 7 zero-shot tasks; group size 128.
FP4/FP6/FP8 → 3/4/6-bit payloads.
28 nm, 1 GHz; HBM2 256 GB/s; matched peak throughput.
Synthesis + cycle-accurate simulation.
Decode: LLaMA-2-7B; batch 64; prefill/decode 512/128.
Baselines: FPE, FIGNA, FIGLUT, AxCore.
Area/energy averages follow this baseline order.
4-bit LLaMA-2-7B: PPL 9.54 vs. 9.24 (FP16);
zero-shot **69.7% vs. 70.2%**. 6-bit: near FP16.

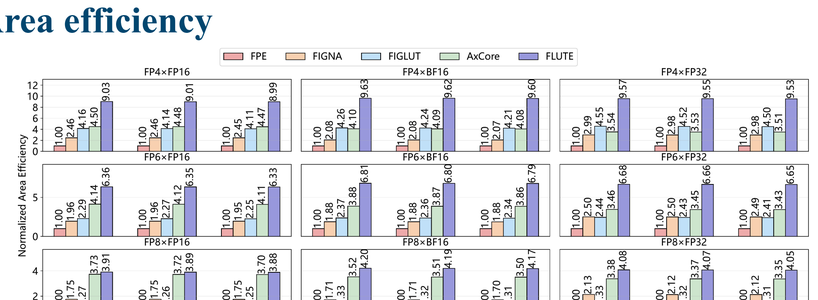
Performance (vs. FPE)



(a) GEMM-level performance (b) End-to-end performance

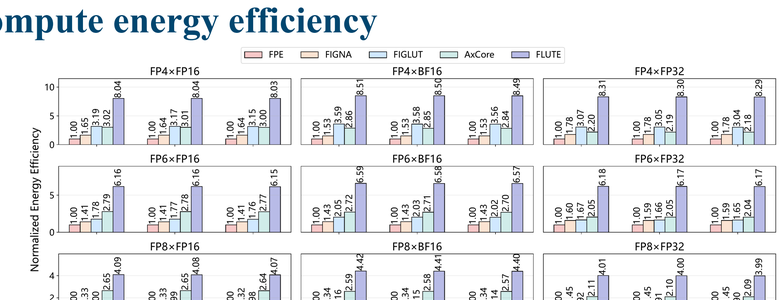
GEMM: **1.61–2.02×**; decode: **1.53–1.85×**.

Area efficiency



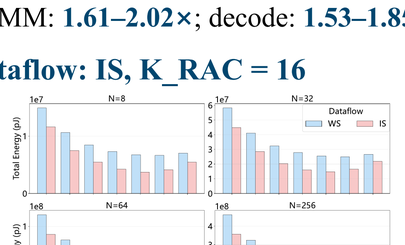
Average: **6.32× / 2.97× / 2.66× / 1.62×**

Compute energy efficiency

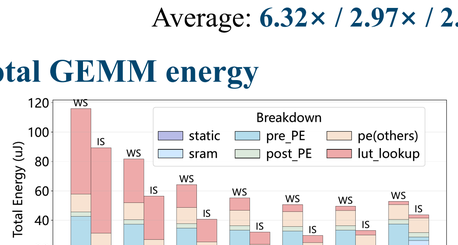


Average: **6.01× / 4.02× / 3.35× / 2.38×**

Dataflow: IS, K_RAC = 16

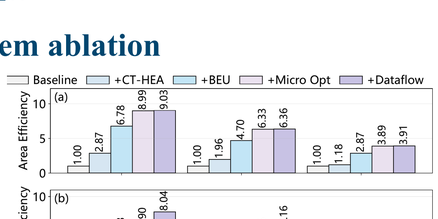


Total GEMM energy



LUT reuse lowers total energy.

System ablation



Accuracy and PE ablations

Accuracy decomposition (LLaMA-2-7B)

Setting	FP16	FP6	FP4	Data Sep.	CT-HEA	FLUTE Exec.
Avg. Acc.	0.702	0.699	0.694	0.698	0.697	0.697
Δ vs. prev.	–	-0.003	-0.005	+0.004	-0.001	0.000

Datapath area (mm²)

Act	Pre	LUT-G	BEU	PEs	INT2FP	Acc	Total
FP16	0.025	0.008	0.007	1.203	0.053	0.104	1.400
BF16	0.023	0.006	0.007	1.015	0.051	0.104	1.207
FP32	0.064	0.015	0.007	1.941	0.075	0.104	2.207

PE area (μm²) / power (mW)

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

Final PE: 28.4% of baseline area.

Conclusion

FLUTE combines **CT-HEA, compact LUTs and hardware co-design** to improve LLM decoding speed and efficiency while retaining accuracy at practical precisions.