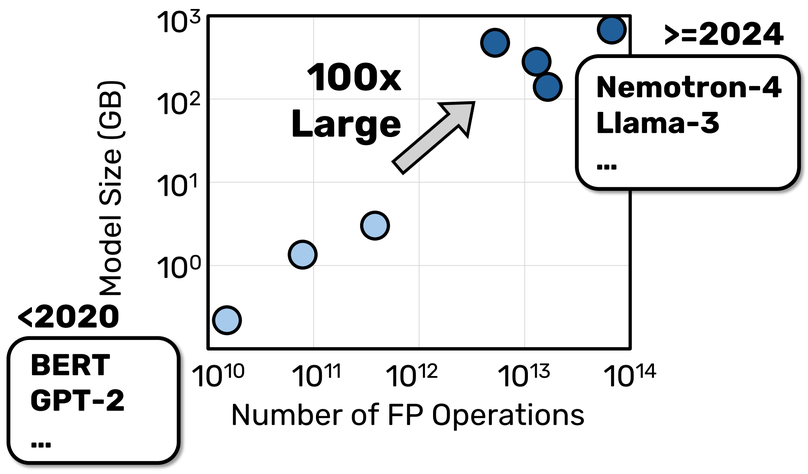
Tender: Accelerating Large Language Models via Tensor Decomposition and Runtime Requantizaiton

Jungi Lee*, Wonbeom Lee*, Jaewoong Sim
Seoul National University

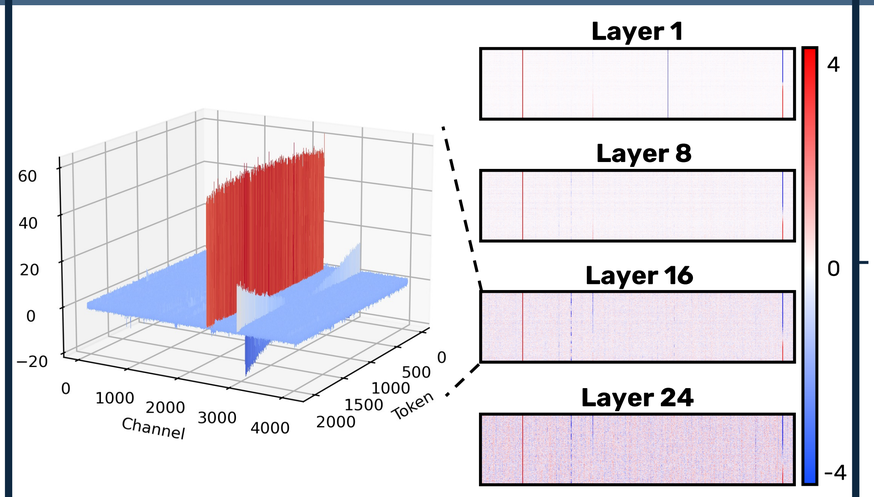
Background

Computations & Sizes of LLMs



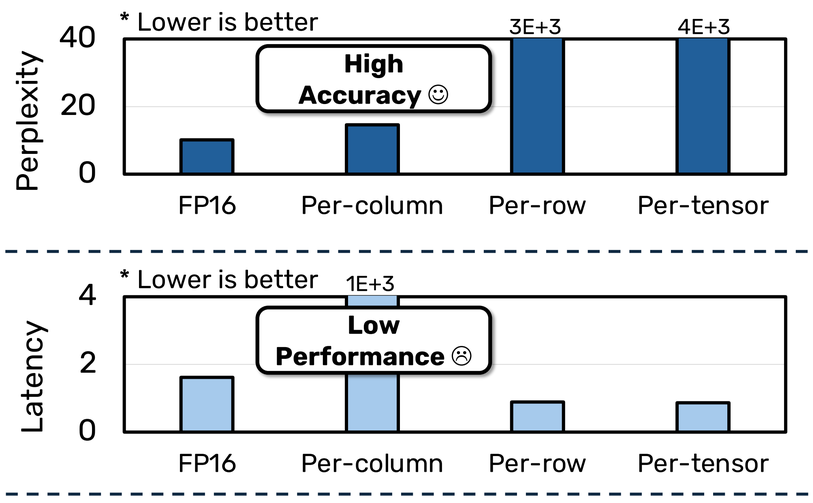
The growing memory and compute requirements of LLMs make it challenging for wide deployment.

Outliers in LLMs



Large magnitude **outliers** show **column-wise** patterns and exist across the layers of LLMs.

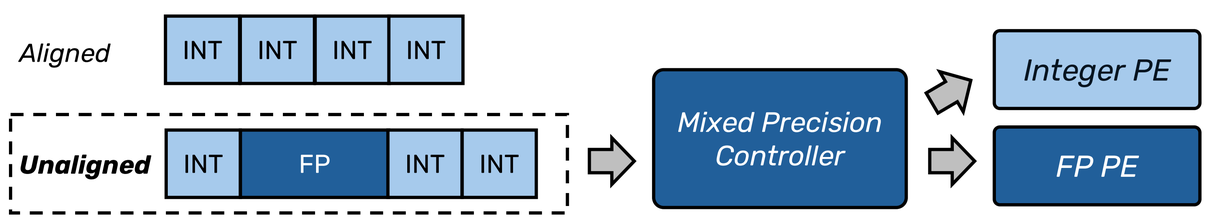
Accuracy & Performance Tradeoff



Per-column grouping **isolates** column-wise outliers, but it incurs **large performance overhead**.

Limitations of Prior Outlier-Aware Works

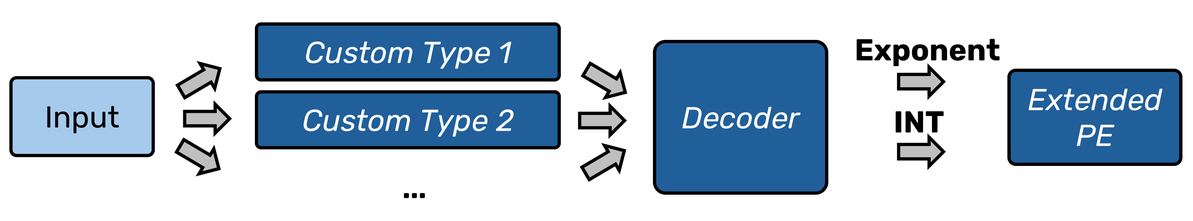
Mixed Precision



Mixed precision results in **complex hardware** design to handle different precisions.

➔ Design Goal: Use a Single Datatype

Custom & Multiple Types

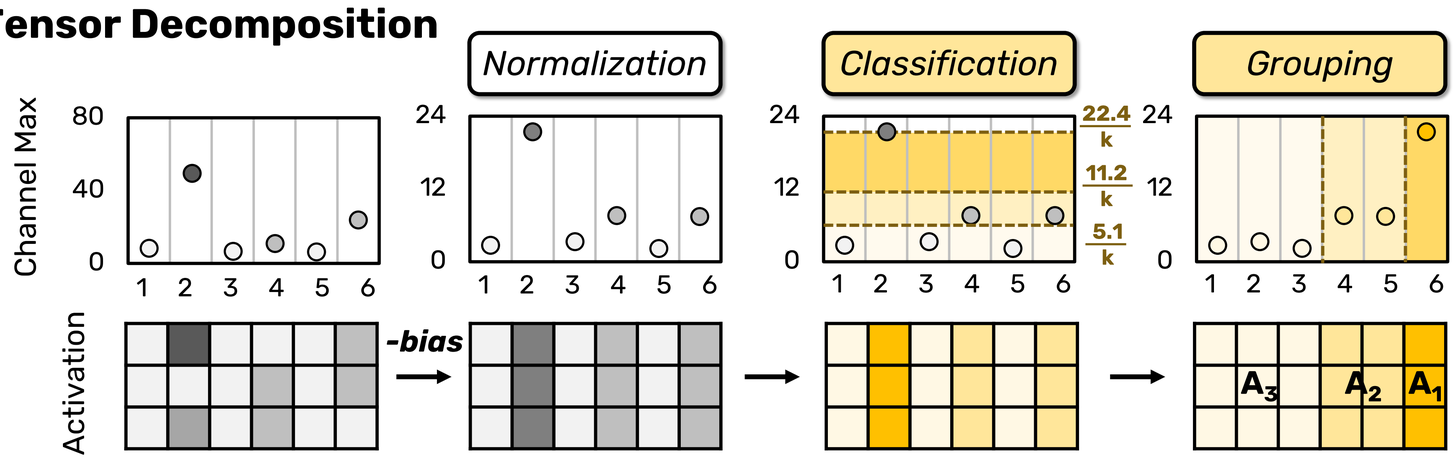


Custom data types result in **intrusive design** and require a decoder and extension of PE.

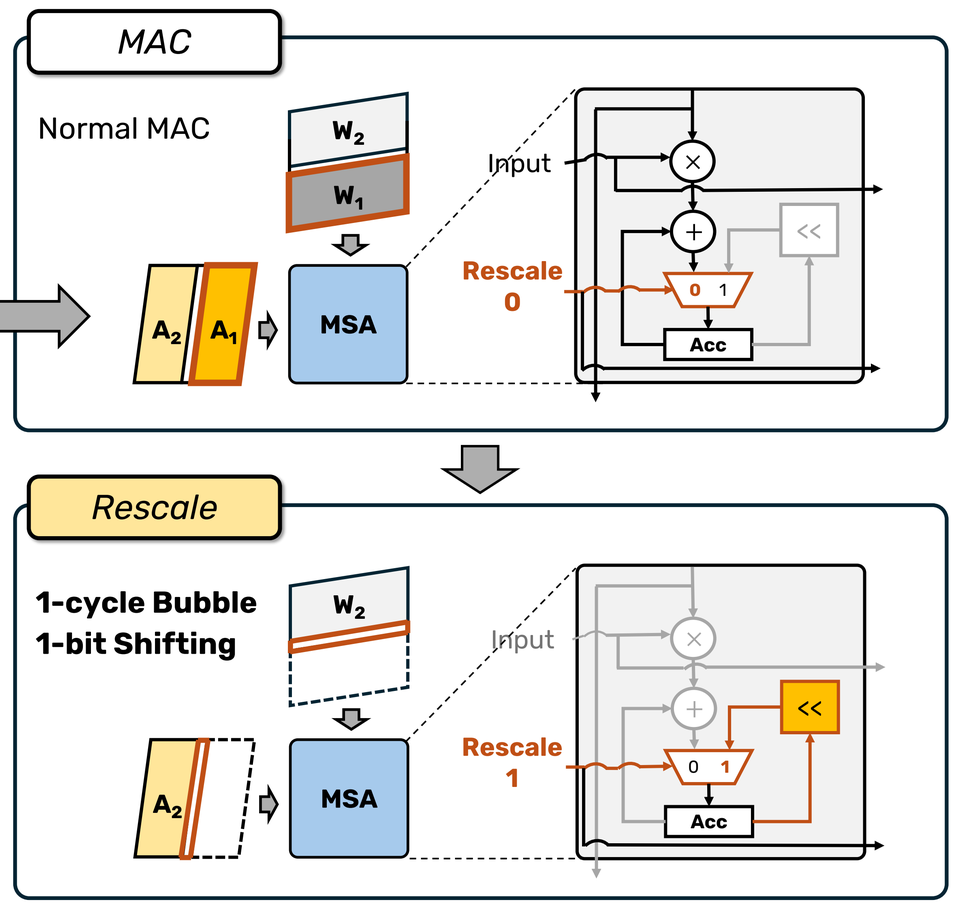
➔ Design Goal: Use Existing Integer Pipeline

Tender

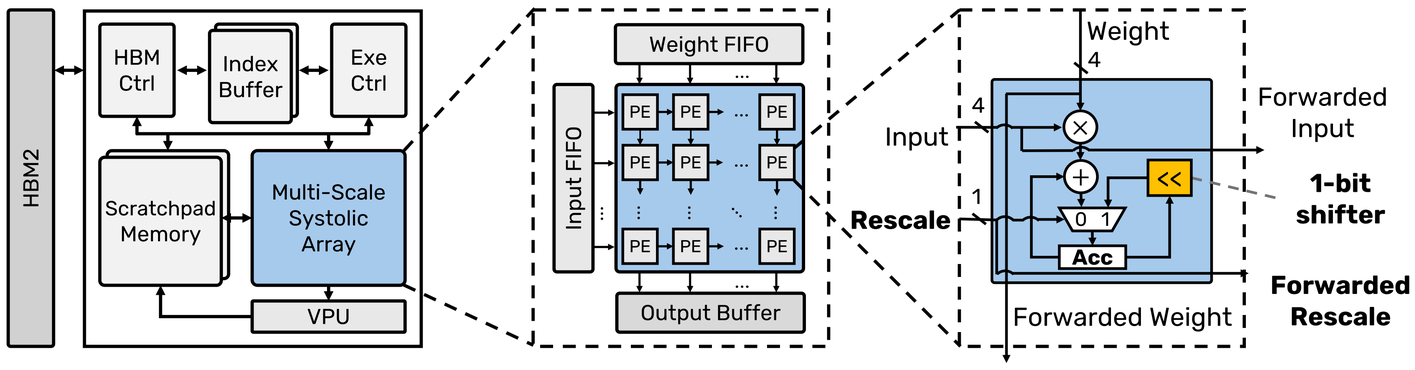
Tensor Decomposition



Runtime Requantization



Hardware Architecture



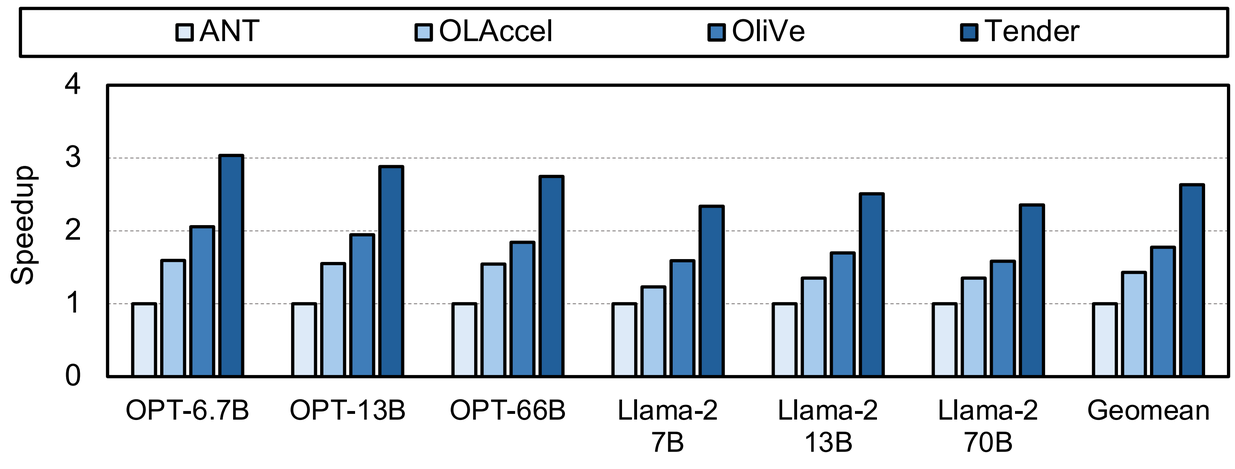
Results

Quantization Accuracy

Precision	Scheme	OPT-66B	Llama-2-70B	LLaMA-13B
FP16	Base	9.34	3.32	5.09
INT8	SmoothQuant	9.87	17.30	16.02
	OliVe	**9.43**	50.94	7.62
	Tender	**9.43**	**3.48**	**5.28**
INT4	SmoothQuant	6E+4	7E+4	2E+5
	OliVe	6E+3	99.91	94.32
	Tender	**12.38**	**13.43**	**13.68**

➔ Tender Shows **Better Quantization Accuracy**

Speedup over Other Accelerators



➔ Tender Provides **Speedup** Over the Other Accelerators