Atom: Low-bit Quantization for Efficient and Accurate LLM Serving

Yilong Zhao[1,2], Chien-Yu Lin[1], Kan Zhu[1], Zihao Ye[1], Lequn Chen[1], Size Zheng[1,3], Luis Ceze[1,4], Arvind Krishnamurthy[1], Tianqi Chen[5], Baris Kasikci[1]

UW[1], SJTU[2], PKU[3], OctoAI[4], CMU[5]

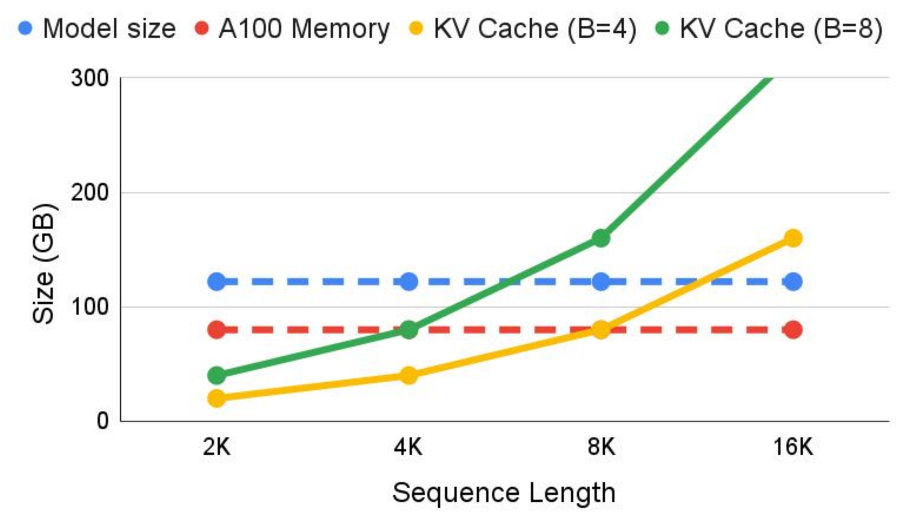
Serving LLMs is Challenging

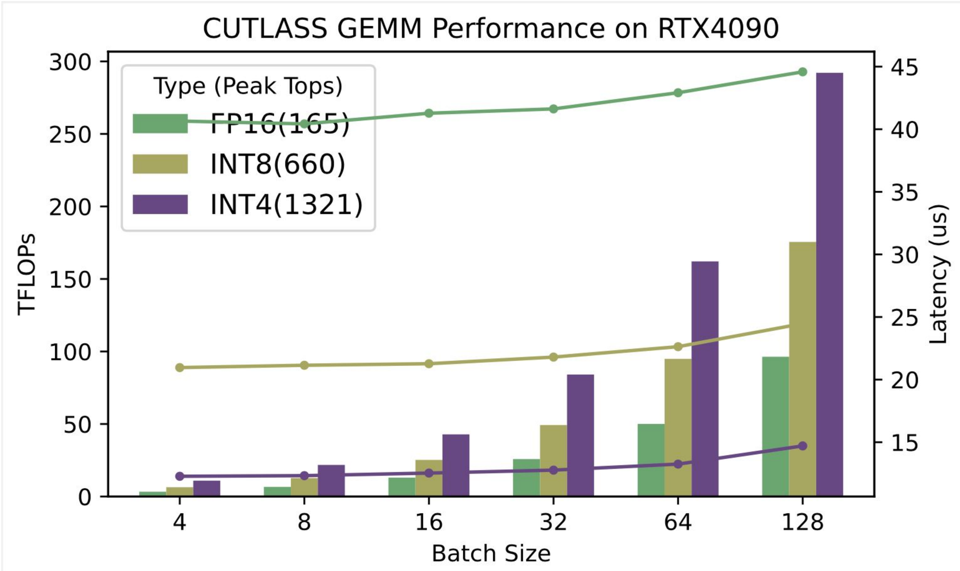
1. Large Memory Consumption

- Model **weights and KV-Cache** consumes significant memory.
- High memory demand **limits #requests** can be served concurrently.

2. Low Computate Utilization

- GPU's compute is **under utilization** when **batch size is small**
- Batch size can be increased if model and KV cache are compressed





Memory Consumption (Llama-65B)

GEMM performance of Llama2-7B

Why Quantization?

- **Save memory** by reducing effective bits per element.
- **Boost compute** by increasing batch and using low-bit hardware, tensor cores.

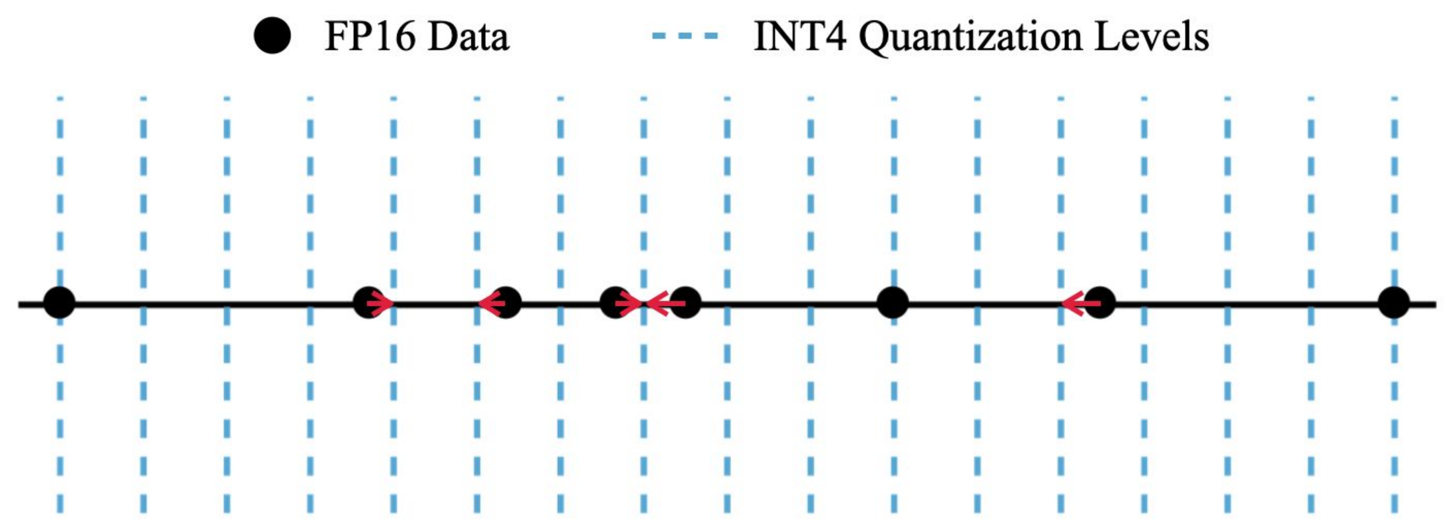
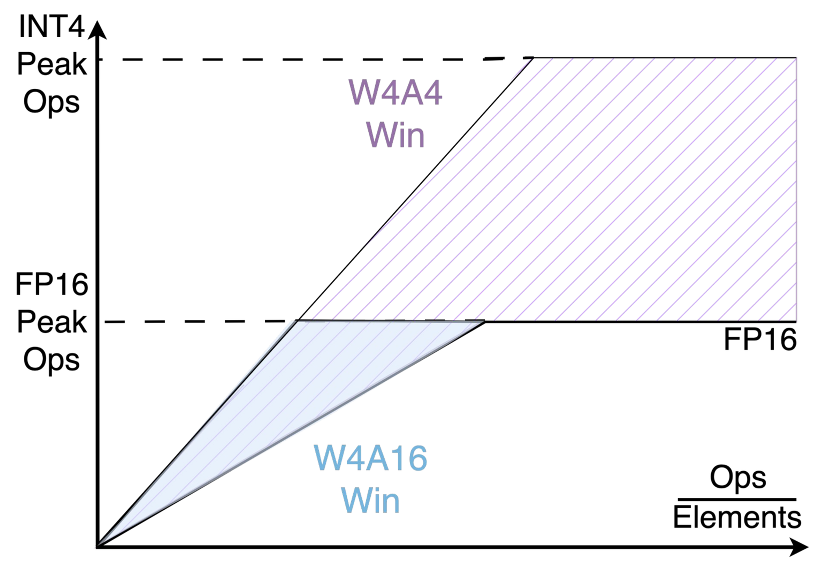
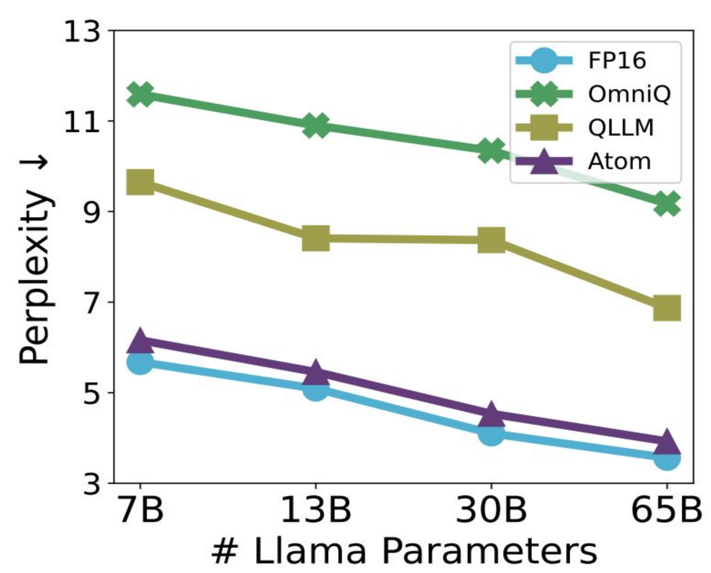


Illustration of uniform quantization

Prior Works Fall Short

- **Weight-only** quantizations (GPTQ, AWQ, QUIP…) **falls short to boost efficiency** when **op intensity increases** (larger batch)
- Prior 4-bit weight-activation quantizations fails to **maintain accuracy**





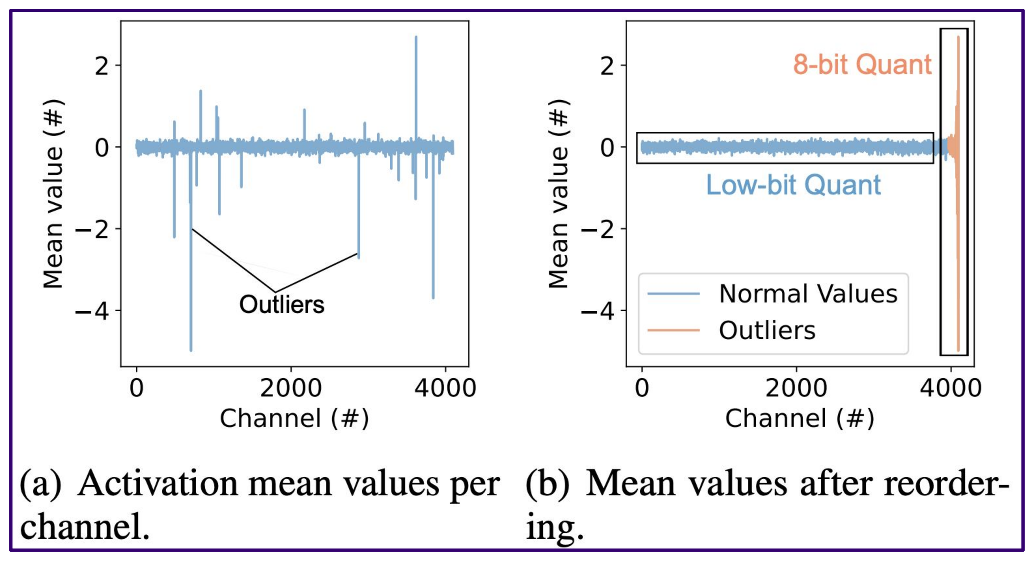
Roofline for 4-bit W-only and W-Act Quant

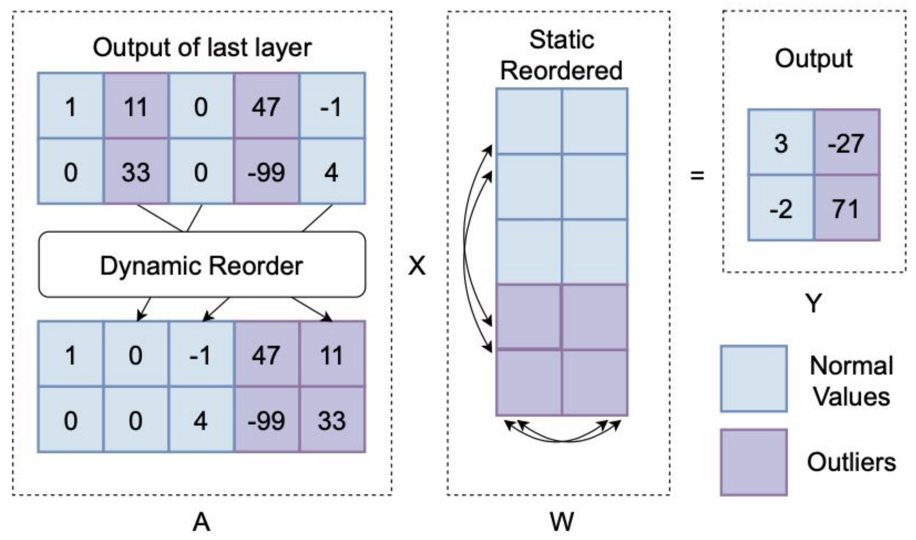
Perplexity: Prior works vs Atom

Overview of Atom's design

Reorder-based Mixed Precision

- **Outliers** severely degrades quant accuracy, calling **higher precision.**
- **Reorder-based** method avoids irregular memory access, with 30% speedup.



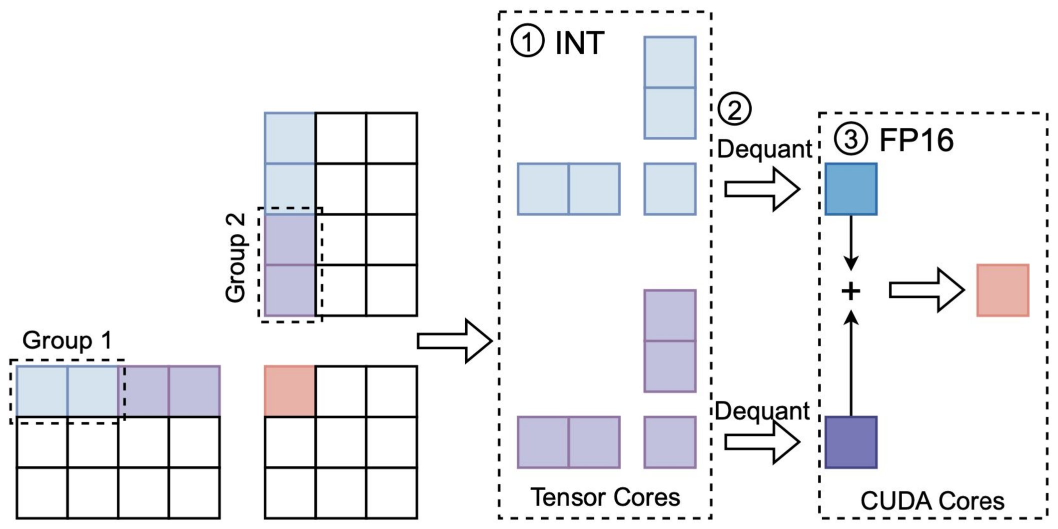


Activation outliers in LLMs

Atom's GEMM with reordering

Fined-grained Group Quantization

- Atom performs quant at a **finer granularity**, with small group sharing parameter
- Atom manages the dequant overhead by a **specialized GPU kernel**



Quantization method	WikiText2 PPL↓
FP16 baseline	5.68
W4A4 RTN	2315.52
+ Keeping 128 outliers in FP16	11.34 (2304.2↓)
+ Quantizing outliers to INT8	11.39 (0.05↑)
+ Group size 128	6.22 (5.17↓)
+ Clipping	6.13 (0.09↓)
+ GPTQ	6.04 (0.09↓)
+ Quantizing KV-cache to INT4	6.16 (0.12↑)

Atom's group quant

Ablation study of quant. techniques

Results

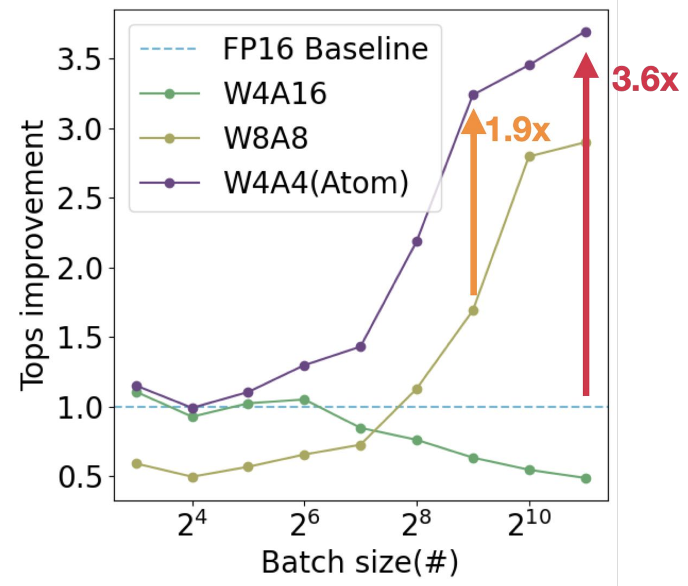
- Atom can **maintain accuracy** while increasing serving throughput for up to **7.7x**
- Performance is measured on a RTX 4090 GPU and based on Llama-7B

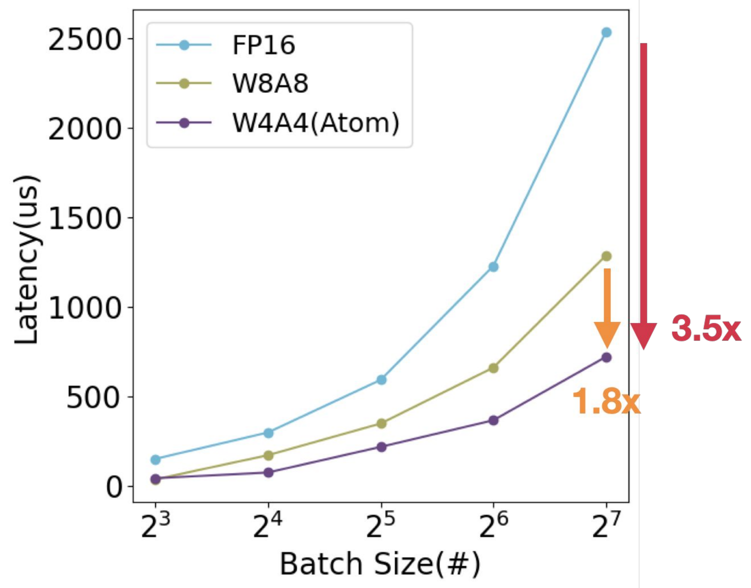
Size	#Bits	Method	Zero-shot Accuracy ↑ PQ	Arc-e	Arc-c	BQ	HS	WG	Avg.
65B	FP16	-	80.79	58.71	46.33	82.26	80.71	77.03	70.97
	W4A4	SmoothQuant	60.72	38.80	30.29	57.61	36.81	53.43	46.28
		OmniQuant	71.81	48.02	35.92	73.27	66.81	59.51	59.22
		QLLM	73.56	52.06	39.68	-	70.94	62.90	59.83
		Atom	**80.41**	**58.12**	**45.22**	**82.02**	**79.10**	**72.53**	**69.57**

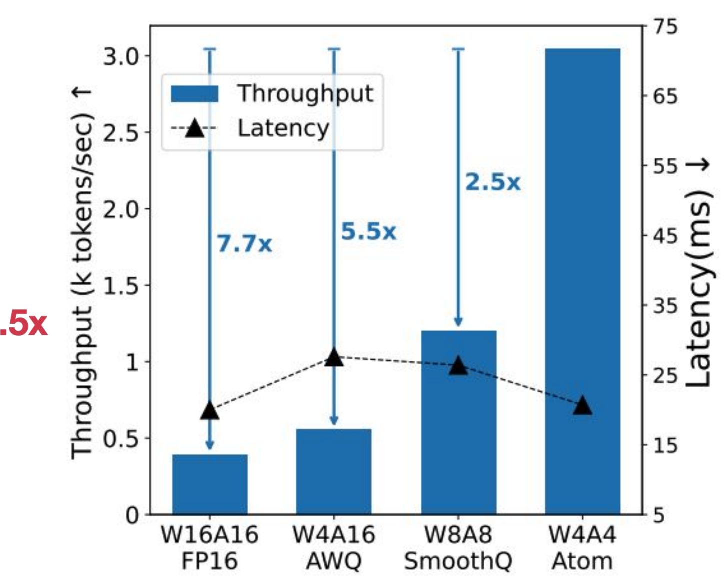
Llama-65B zero shot accuracy

# Bits	Method	Llama2 7B	Llama2 13B	Llama2 70B	Mixtral 8x7B
FP16	-	5.47	4.88	3.32	3.84
W4A4	SmoothQuant	83.12	35.88	-	-
	OmniQuant	14.61	12.3	-	-
	Atom (INT)	**6.03**	**5.27**	**3.68**	**4.41**
	Atom (FP)	**6.14**	**5.35**	**3.78**	**4.50**

Llama2 & Mixtral perplexity







Throughput of dense layer

Latency of self-attn layer

End-to-end performance