ICLR
International Conference On Learning Representations

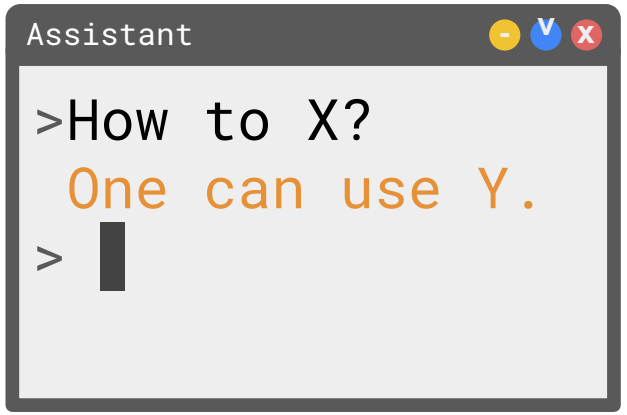
KV Cache Transform Coding for Compact Storage in LLM Inference

Konrad Staniszewski & Adrian Łańcucki

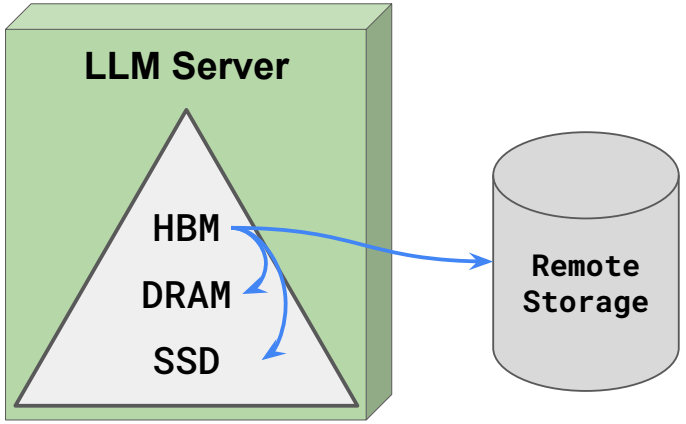
Rio de Janeiro, Brazil, 2026

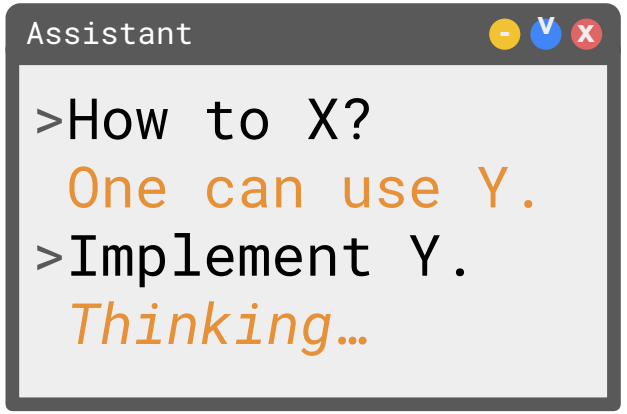
Problem Setting

Compression of KV Cache for Storage

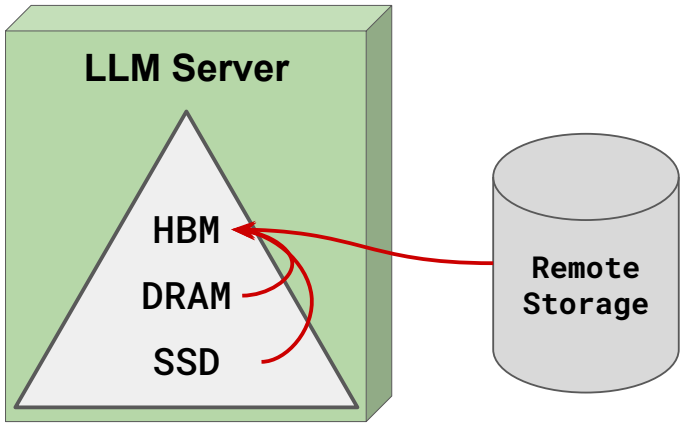


Awaiting for user/tool response





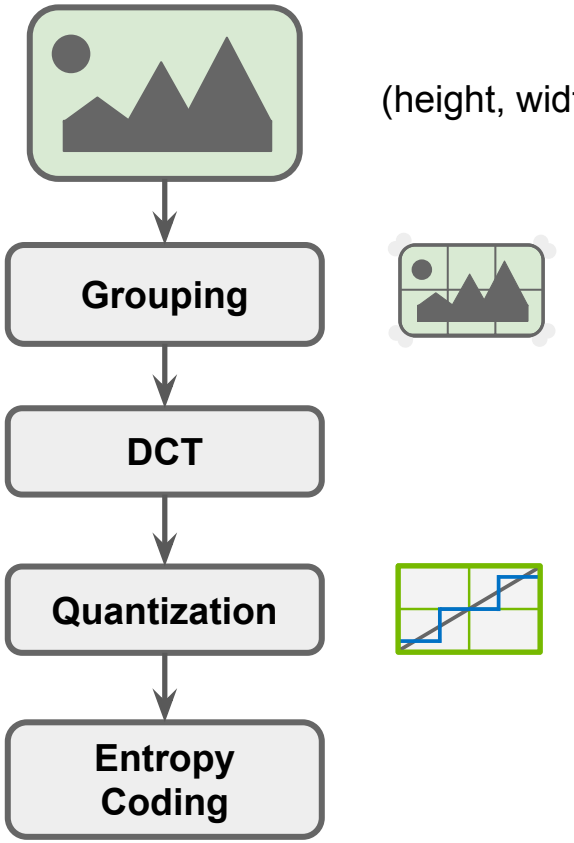
Processing next request



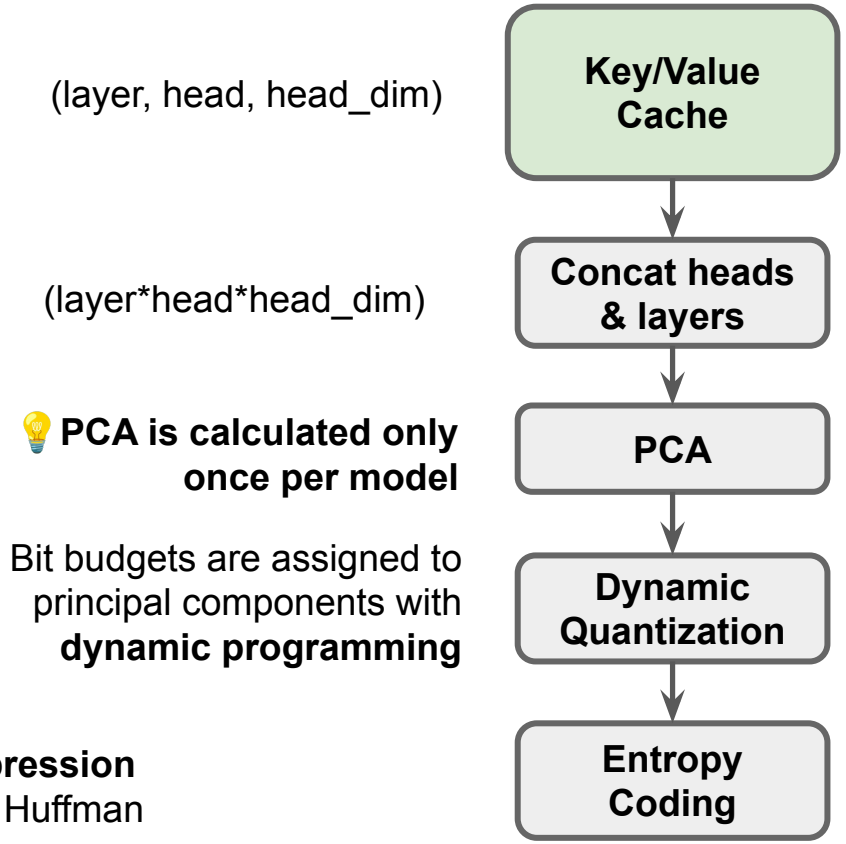
Our Approach - KVTC

Transform Coding

JPEG:



KV Cache Transform Coding:



💡 **PCA is calculated only once per model**

💡 Bit budgets are assigned to principal components with **dynamic programming**

Variety of **lossless compression** algorithms: **ANS**, Deflate, Huffman Coding

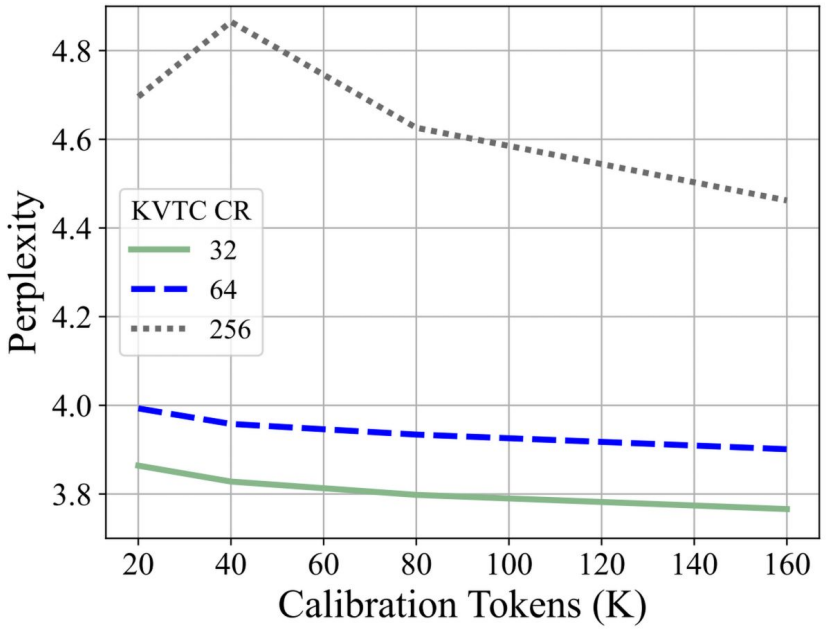
Generalization Study

More Calibration Data -> Better Performance

KVTC generalizes well for distant calibration data

Method	Calibration Data	CR	GSM	MMLU	QASPER	LITM	RULER-VT
			Mistral NeMo 12B				
Vanilla		1	$61.9_{1.3}$	$64.5_{0.4}$	38.4	$99.5_{0.1}$	$99.8_{0.2}$
GEAR 2bit	-	5	$59.8_{1.4}$	$64.0_{0.4}$	38.6	$96.9_{0.2}$	$99.4_{0.3}$
KIVI 2bit		5	$59.7_{1.4}$	$64.3_{0.4}$	38.2	$91.9_{0.3}$	$98.3_{0.4}$
kvtc$_{16\times}$	FineWeb + OpenR1Math	17-20	$63.5_{\pm 0.9}$	$64.7_{\pm 0.3}$	$37.2_{\pm 0.4}$	$99.6_{\pm 0.3}$	$99.7_{\pm 0.0}$
	Python	17-20	$59.4_{1.4}$	$65.2_{0.4}$	37.4	$99.9_{0.0}$	$99.7_{0.2}$
	C	17-20	$55.0_{1.4}$	$65.4_{0.4}$	38.0	$99.2_{0.1}$	$99.3_{0.3}$
	Assembly	17-20	$58.9_{1.4}$	$65.2_{0.4}$	37.0	$99.9_{0.0}$	$99.7_{0.2}$

Careful calibration unlocks higher compression ratios



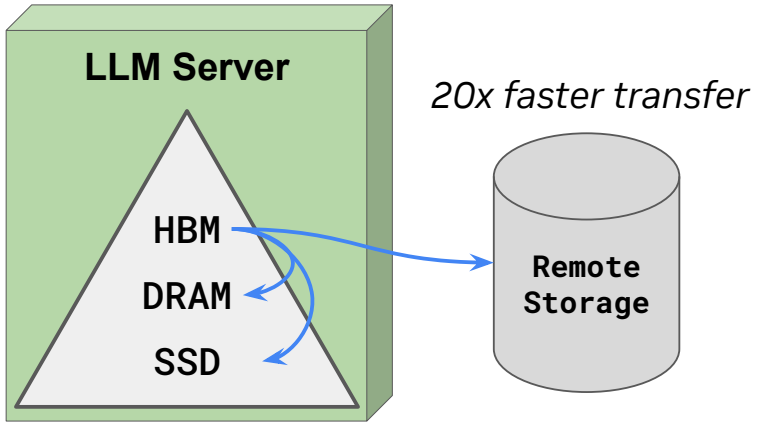
Results

Comparison with Other Methods

	Vanilla	$GEAR_{2\text{-bit}}$	$KIVI_{2\text{-bit}}$	H_2O	TOVA	xKV	FP8	kvtc$_{8\times}$	kvtc$_{16\times}$	kvtc$_{32\times}$	kvtc$_{64\times}$
Llama 3.1 8B											
CR	1	5	5	8	8	1-5	2	9-10	18-22	34-44	60-88
GSM8K	**56.8**	52.8	52.8	54.3	54.5	**56.6**	55.2	**57.0**	**56.9**	**57.8**	**57.2**
MMLU	**60.5**	**59.6**	**59.6**	44.3	44.8	**59.5**	**60.1**	**59.8**	**60.1**	**60.6**	**60.7**
QASPER	**40.4**	**40.4**	39.1	34.3	38.6	35.6	**40.8**	**40.1**	**40.7**	**39.4**	37.8
LITM	**99.4**	96.9	88.8	20.2	1.2	**99.9**	**99.4**	**99.3**	**99.3**	**99.1**	90.2
RULER-VT	**99.8**	**99.8**	**98.9**	50.4	**99.7**	**99.8**	**99.9**	**99.1**	**99.1**	**98.9**	95.9

Task	Vanilla	kvtc$_{8\times}$	kvtc$_{16\times}$
Llama 3.3 70B Instruct			
MATH-500	$75.6_{1.92}$	$73.2_{1.98}$	$73.2_{1.98}$
NIAH	100.0	100.0	100.0
LITM	100.0	100.0	100.0
Qwen 2.5 R1 7B			
AIME24	$50.9_{\pm 4.9}$	$52.5_{\pm 3.6}$	$50.9_{\pm 6.8}$
AIME25	$40.8_{\pm 4.3}$	$40.8_{\pm 5.2}$	$38.3_{\pm 5.5}$
LCB	36.7	36.5	31.6
Qwen3.5 4B			
τ^2-Bench Telecom	95.6	96.5	93.9
NIAH	100.0	100.0	100.0

Performance & Deployment

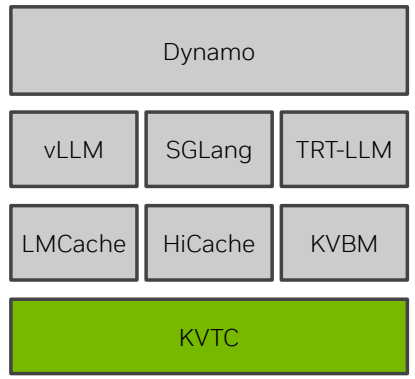


20x faster transfer

20x compression increases KV Cache retention time 20x

- Compression can be performed async on CPU / DPU
- GPU can be utilized for decompression

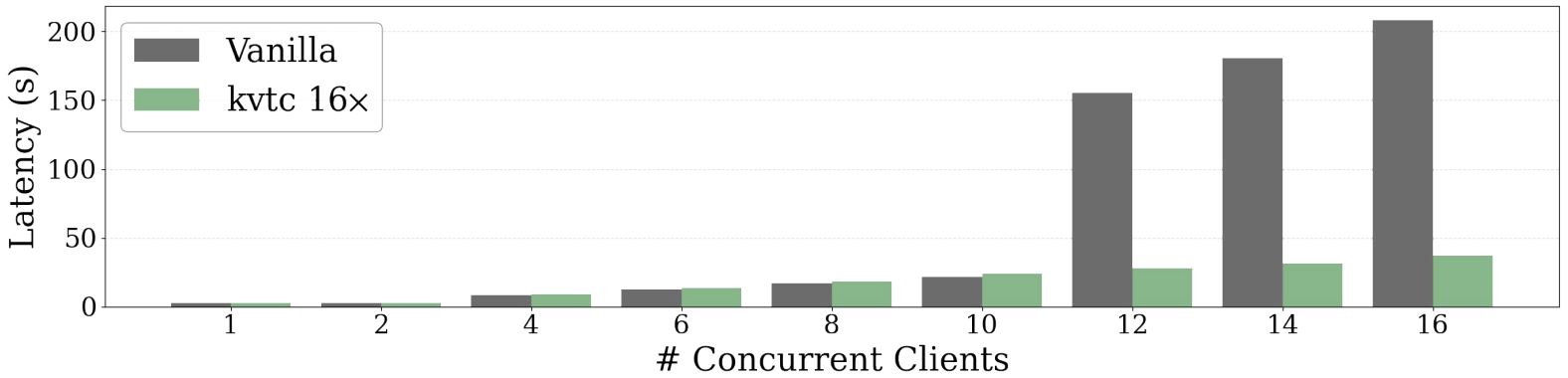
- KVTC can act as a **separate compression layer** in the inference stack
- Easy to integrate - **no alterations** to vLLM/SGLang/...
- Input: KV cache Output: Compressed data



Performance

vLLM with LMCache Example

Llama 3.3 70B FP8 TP2
~64K prefill **~100** tokens per turn
256GB DRAM KV Cache storage space



ICLR
International Conference On Learning Representations

KV Cache Transform Coding for Compact Storage in LLM Inference

Konrad Staniszewski & Adrian Łańcucki

Rio de Janeiro, Brazil, 2026