PACE: Periodic and Channel-Aligned Encoding of Four-Bit Recurrent States

Paper by Patrick Haller, Filipe Laitenberger, and Alan Akbik

Research group paper reading 27 October 2026

arXiv:2610.29545v1 (21 Oct 2026)
Publication record: not found as of 27 Oct 2026.
Public official code: not found as of 27 Oct 2026.

Four bits become viable when rounding follows the kernel

One-sentence takeaway

PACE rounds a recurrent state once per short decode buffer, groups scales within value channels, and reserves 8 bits for error-persistent heads.

- **Temporal**: fewer state writes mean fewer injected rounding errors.
- **Spatial**: key-axis groups avoid mixing large and small value columns.
- **Selective**: 8-bit heads protect the errors that decay slowly.

Paper, Abstract and Sec. 1, PDF pp. 1–2; Fig. 2, PDF p. 3.

Concurrent decoding makes fixed-size state expensive

- Each gated-delta head stores a matrix $S_t \in \mathbb{R}^{d_k \times d_v}$ for every active sequence.
- At 64 concurrent sequences, fp32 states account for 10–41% of weights-plus-states memory on the evaluated models.
- Four-bit storage creates room for more sequences, provided quality survives repeated state updates.

The failure mode

Per-token decode materializes and rounds the *same* recurrent state after every generated token.

Paper, Sec. 1, PDF p. 2. The memory share is the authors' 64-sequence estimate.

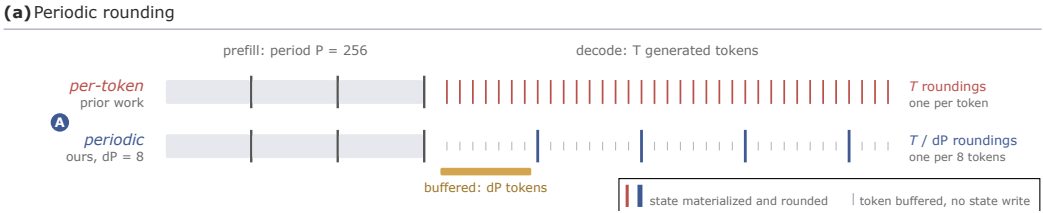
The gated-delta update exposes two useful structures

$$S_t = a_t(I - \beta_t k_t k_t^\top)S_{t-1} + \beta_t k_t v_t^\top = a_t S_{t-1} + k_t u_t^\top$$
$$u_t = \beta_t(v_t - a_t S_{t-1}^\top k_t), \qquad o_t = S_t^\top q_t.$$

- $a_t \in (0, 1)$: decay; β_t: write strength.
- $k_t \in \mathbb{R}^{d_k}$, $u_t \in \mathbb{R}^{d_v}$: an outer-product update.
- A persistent large coordinate of u_t repeatedly writes one **value column** of S_t.

Paper, Sec. 2.1, Eqs. (2)–(3), PDF p. 3.

Prefill already rounds once per chunk; decode can too



- Prefill period $P = 256$ tokens in every benchmark evaluation.
- Unbuffered decode: $dP = 1$; PACE headline configuration: $dP = 8$.

Original paper Fig. 2a, PDF p. 3; Sec. 3.1, PDF p. 5.

Longer periods let old rounding errors decay

$$E_h \approx \frac{e_h}{1 - a_h^{2dP}}$$

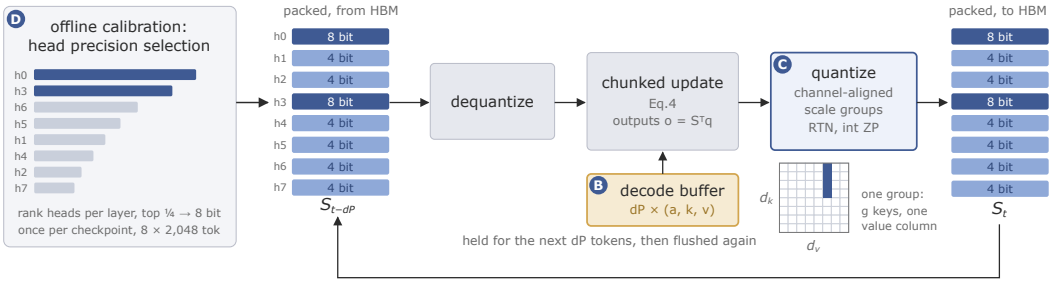
- e_h: energy injected by one rounding in head h.
- a_h: representative head decay; dP: decode rounding interval.
- Fewer injections and more decay between them reduce retained error.
- A head with a_h near 1 holds errors longer and merits more precision.

Reading the expression

It uses independent rounding errors and a mean-decay proxy for token-varying a_t; it is the paper's ranking intuition.

Paper, Sec. 3.2, Eq. (5), PDF p. 5.

A buffer advances the packed state only at flushes



Original paper Fig. 2b, PDF p. 3. A: period; B: buffer; C: scale groups; D: head selection.

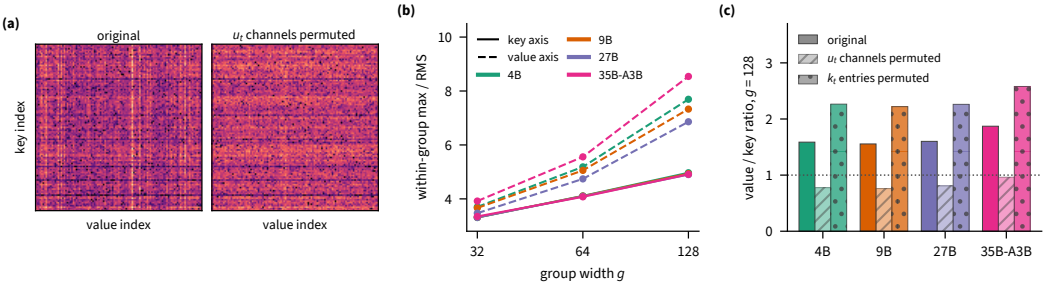
The memory metric includes the decode buffer

$$\underbrace{b + \frac{16}{g}}_{\text{symmetric state, bits/value}} + \underbrace{dP\,\frac{b_k d_k/\gamma + b_v d_v + 32}{d_k d_v}}_{\text{buffer, bits/value}}$$

- b: state bits; g: group width; b_k, b_v: buffered key/value bits.
- γ: value heads sharing one key head.
- Uniform INT4, $g = 32$, int8 buffer, $dP = 8$: $4.50 + 0.78 = \boxed{5.28}$ resident bits/value.

Paper, Sec. 5.1, Eq. (6), PDF p. 6; Appendix A, PDF p. 15.

Large state entries form value columns



- Dashed value-axis groups cross columns and have larger max/RMS.
- Solid key-axis groups stay within one column and use finer steps for typical entries.

Original paper Fig. 3, PDF p. 4; Sec. 4.1, PDF p. 6.

Permuting the delta value removes the column advantage

Value/key max/RMS, $g = 128$	4B	9B	27B	35B-A3B
Original state	1.588	1.555	1.601	1.871
u_t channels permuted/token	0.776	0.760	0.808	0.963
k_t entries permuted/token	2.265	2.220	2.261	2.580

- A ratio > 1 means value-axis groups have the worse tail statistic.
- Permuting u_t, while preserving its values, reverses that ranking.

Paper, Sec. 4.2, PDF p. 6; Appendix C, Table 4, PDF p. 16.

PACE chooses precision at two granularities

Within each head

Group g adjacent entries along the key axis, all in one value column. Round to nearest with a shared fp16 scale.

Across heads

Estimate retained error E_h on eight held-out 2,048-token windows; keep the highest-ranked quarter of heads per layer at 8 bits.

The head ranking is computed once per checkpoint; model weights are not fine-tuned.

Paper, Secs. 5.2–5.3, PDF p. 7.

The two PACE variants spend different memory

	PACE-4	PACE-4/8
State heads	all 4 bit	75% 4 bit; 25% 8 bit
Key-axis group g	64	32
State grid	symmetric	affine, integer zero point
Decode buffer k/v	int8 / int8	int4 / bf16
Resident bits/value	**5.03**	**6.82**

For PACE-4/8, the mixed state payload is 5.66 bits/value; the rest is buffer cost.

Paper, Sec. 5.3, PDF p. 7; Appendix E, PDF p. 17; Table 1, PDF p. 9.

Evaluation: four hybrids and five headline tasks

- Qwen3.5-4B, 9B, 27B; Qwen3.6-35B-A3B.
- GSM8K and MATH-500: exact match; MMLU-Pro: 560-question subset.
- IFEval: 541 prompts, strict accuracy; RULER-32k: variable-tracking accuracy.
- Headline avg: unweighted mean of *all five* scores. All tasks are zero-shot with greedy decoding and thinking disabled.

Baseline detail

DAMP INT8 rounds every token at 9.90 bits/value. The authors reimplement Q-Mamba DSQ without its published fine-tuning step.

Paper, Sec. 6, PDF p. 8; Table 1, PDF p. 9.

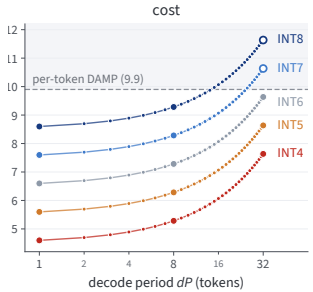
Eight-token rounding improves the uniform INT4 tradeoff

Model	Decode period	3-task avg	Resident bits	IFEval
9B	$dP = 1$	78.1	4.60	61.4
9B	$dP = 8$	82.9	5.28	71.3
9B	$dP = 32$	85.6	7.64	76.9
35B-A3B	$dP = 1$	70.7	4.60	54.7
35B-A3B	$dP = 8$	77.3	5.28	66.2
35B-A3B	$dP = 32$	78.4	7.64	68.8

Three-task average = GSM8K, IFEval, RULER-32k; prefill remains $P = 256$.

Paper, Sec. 7.1, PDF p. 8; Appendix A, Table 3, PDF p. 15.

A longer buffer has a visible memory cost



- The buffer term grows linearly with dP.
- At INT4, $dP = 8$ recovers much of the $dP = 32$ gain with one quarter of its buffer length.

Original paper Fig. 4, right panel, PDF p. 7; Appendix A, Table 3, PDF p. 15.

PACE-4/8 approaches DAMP INT8 at 69% of its bits

Five-task average	4B	9B	27B	35B-A3B
DAMP INT8, 9.90 bits	80.9	83.7	87.4	86.3
PACE-4/8, 6.82 bits	79.1	83.7	87.3	85.0
PACE-4, 5.03 bits	79.6	80.0	84.2	77.3
Q-Mamba DSQ, 5.03 bits	75.2	72.8	78.9	71.2

- PACE-4 beats the post-hoc Q-Mamba DSQ comparison at equal bits.
- The 4B PACE-4/8 average is 1.8 points below DAMP INT8.

Paper, Table 1, PDF p. 9. Average covers all five headline tasks.

The average hides an IFEval gap

IFEval strict accuracy	9B	27B	35B-A3B
DAMP INT8	83.4	88.0	83.0
PACE-4/8	81.3	86.1	81.0
Difference, percentage points	-2.1	-1.9	-2.0

PACE-4/8's headline parity is about the *five-task mean*; it does not imply parity on every task.

Paper, Table 1 and Sec. 7.2, PDF p. 9.

Key-axis grouping supplies a large ablation gain

INT4, $g = 32$, 5.28 bits	IFEval 4B	IFEval 27B	PPL 4B
Value-axis groups	73.75	75.05	8.158
Key-axis groups	**80.59**	**78.37**	**7.719**

- IFEval: higher is better; WikiText-103 perplexity: lower is better.
- At equal resident bits, grouping within columns helps on both models.

Paper, Table 2, PDF pp. 9–10.

Head allocation and buffer precision remain task-sensitive

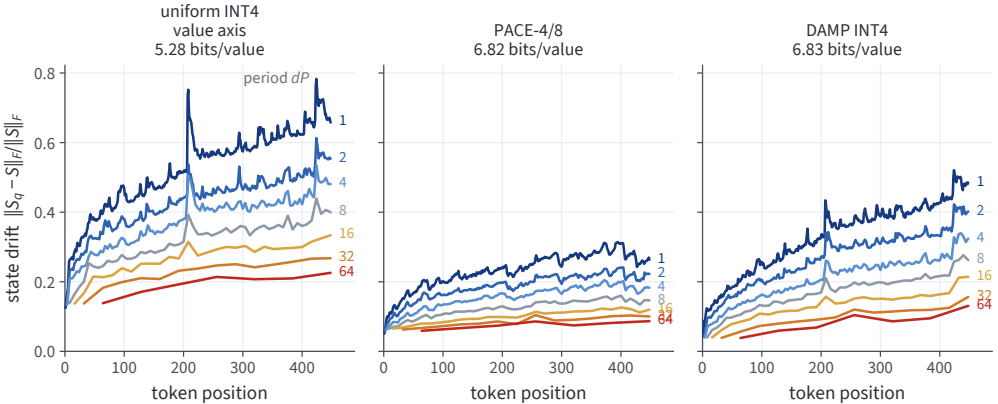
- At 6.44 bits, head allocation plus key-axis groups improves 4B IFEval versus DAMP INT4: 81.52 vs. 79.48.
- On 27B IFEval the same ablation is lower: 78.00 vs. 79.85.
- With the PACE-4/8 state, 9B IFEval is 82.4 for int4 key/bf16 value, but only 22.7 when the buffer widths are reversed.

Interpretation

Buffered values enter the state write directly; key errors are partly filtered by L2 normalization after key quantization.

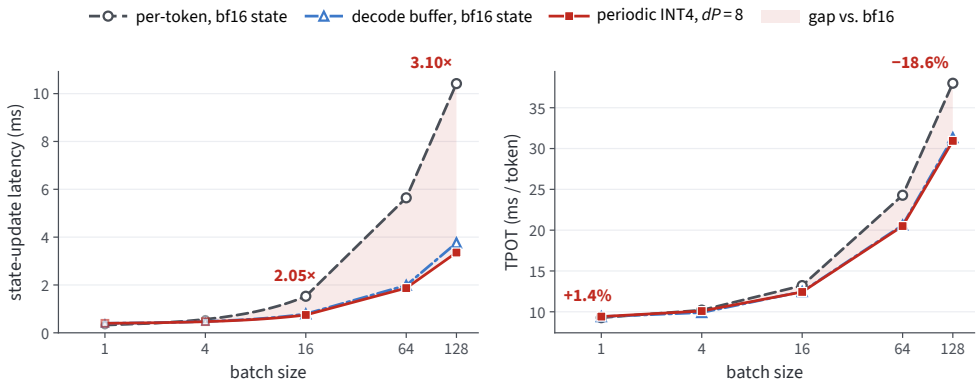
Paper, Table 2, PDF p. 10; Appendix D, Table 5, PDF pp. 16–17.

State drift falls as the rounding period grows



Original paper Fig. 5, PDF p. 10. Drift = $\|S_q - S\|_F / \|S\|_F$, median over recurrent layers; 35B-A3B corroboration in Appendix B, Fig. 7, PDF p. 16.

Periodic INT4 keeps the decode-buffer speedup



At batch 128, TPOT is 18.6% lower than per-token bf16; at batch 1, it is 1.4% higher.

Original paper Fig. 6, PDF p. 11. Qwen3.5-4B, H100 PCIe, 512-token context; uniform INT4 state and int8 buffer, not full PACE-4/8.

What the evidence supports

1. **Less frequent rounding**: improved low-bit scores and lower state drift across tested Qwen hybrids.
2. **Channel-aligned groups**: the delta-value permutation and grouping ablation explain a distinct four-bit gain.
3. **Practical tradeoff**: PACE-4/8 saves resident bits with near-DAMP-INT8 mean scores on the three larger backbones.

Current boundary

Only Qwen gated-delta hybrids were studied; prefill period is fixed at 256. Throughput times uniform INT4, without PACE-4/8's grouping and head allocation.

Paper, Sec. 8 and Limitations, PDF pp. 11–12.

Source and discussion pointers

- Haller, Laitenberger, Akbik. PACE: Periodic and Channel-Aligned Encoding of Four-Bit Recurrent States, arXiv:2610.29545v1, 2026.
- Main results: Table 1, PDF p. 9; mechanisms: Eqs. (5)–(6) and Figs. 2–3, PDF pp. 3–6.
- Current publication and code status: no verifiable formal venue or public official implementation found as of 27 October 2026.

Question for follow-up

Does the value-column structure persist in other recurrent model families, and which prefill period best balances long-context accuracy with cost?