上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Week 8 Extension: Efficient Attention

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

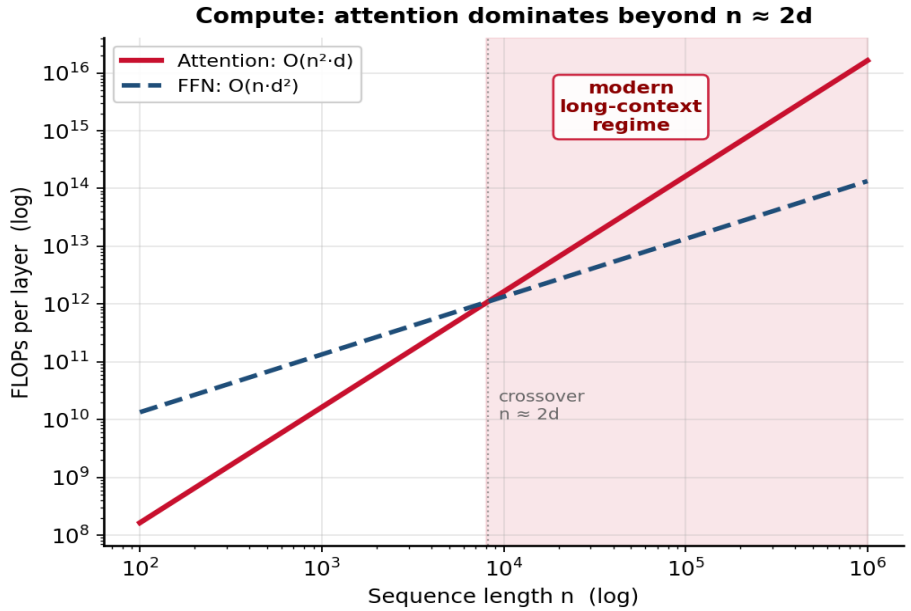
AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

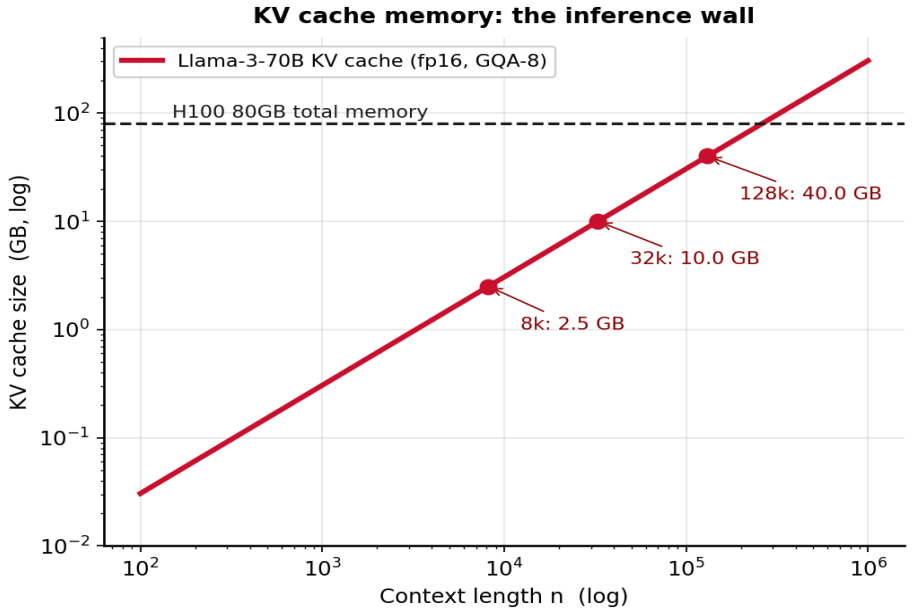
The $O(n^2)$ Wall, Revisited

L24 left us with self-attention's scaling problem.

- Per-layer compute: $O(n^2 \cdot d)$ — quadratic in sequence length
- KV cache memory: $O(n)$ per token per layer per head — and grows every generation step

Example: Llama-3-70B at 128k context needs ~40 GB of KV cache for ONE sequence — half an H100.





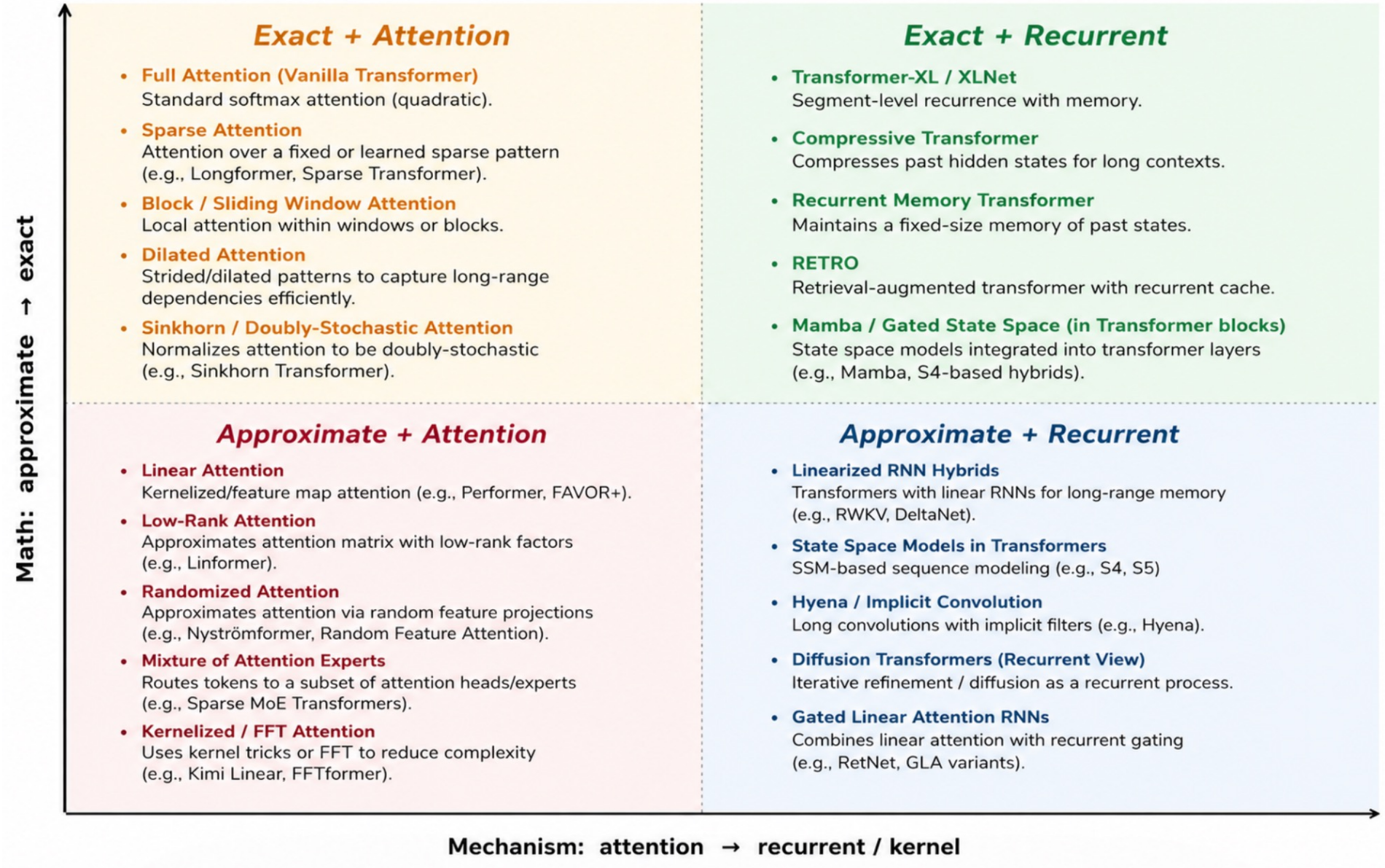
Three Strategies for Climbing the Wall

Each family attacks a different assumption of vanilla attention:

- Sparse attention — keep the math, drop most of the $n \times n$ entries (Longformer, BigBird)
- Linear attention — change the math: $O(n \cdot d^2)$ instead of $O(n^2 \cdot d)$ (Performer, Linformer)
- I/O-aware exact — same math, run it correctly on GPU memory hierarchy (FlashAttention)
- Drop attention entirely — return to recurrence, done right (Mamba / SSMs)

We will tour all four, then look at how 2025–26 frontier labs combine them.

Three Strategies for Climbing the Wall



Sparse Attention: Longformer & BigBird

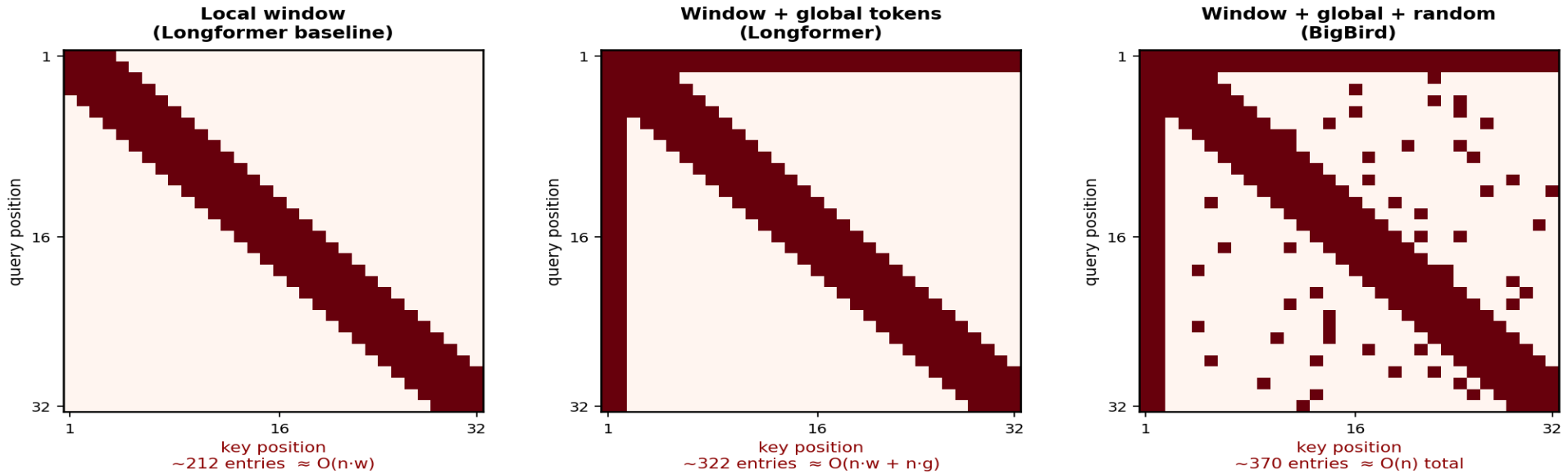
Beltagy et al. Longformer. arXiv 2020 / Zaheer et al. BigBird. NeurIPS 2020

Idea: most of the n × n attention matrix is redundant. Restrict each query to a structured subset.

- Sliding window (local context) + global tokens (CLS, special positions)
- BigBird adds random connections — provably retains universality at O(n) entries

Result: 4k–8k tokens on hardware of the day; standard for long-document QA pre-2022.

Limit: the sparsity pattern is a hand-designed prior. If your task wants attention between distant unrelated tokens, it can't.



Linear Attention: Kernelizing the Softmax

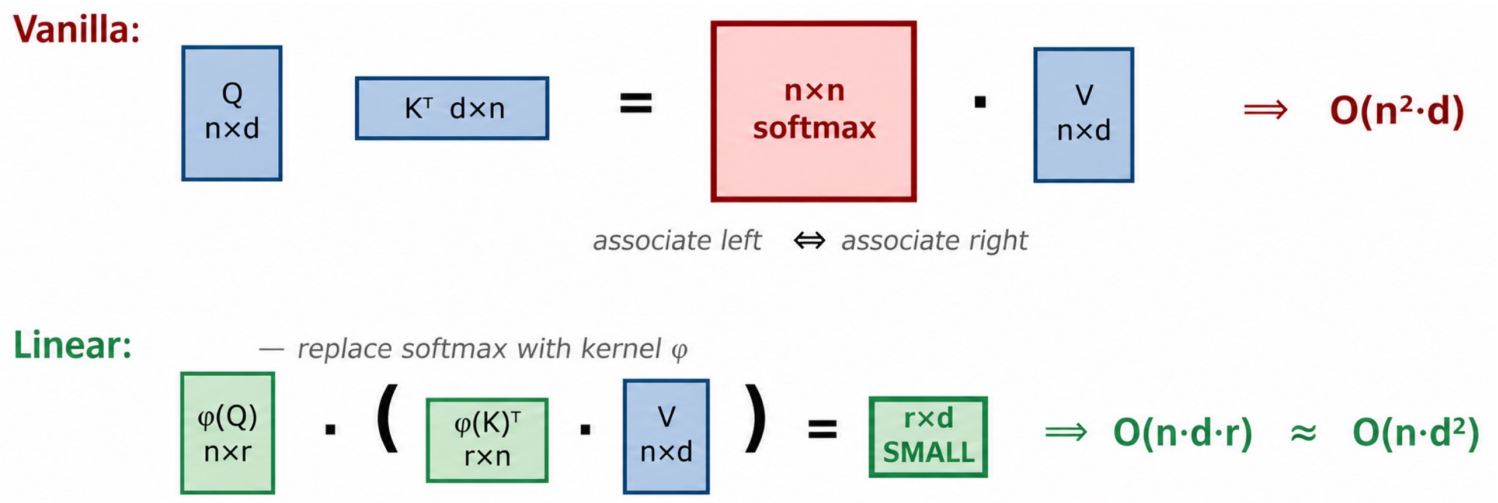
Choromanski et al. Performer. ICLR 2021 / Wang et al. Linformer. arXiv 2020

The trick: replace softmax with a kernel ϕ such that $\phi(Q)\,\phi(K)^\top \approx \mathrm{softmax}\left(QK^\top\right)$

- Then re-associate: $\phi(Q)\left(\phi(K)^\top V\right)$ — inner term is r × d, not n × n.
- Performer: random feature maps (FAVOR+). Linformer: low-rank projection of K, V.

Why it didn't take over: the approximation hurts on the long-range retrieval tasks people actually care about.

"Linear attention forgets." The quality gap was real and persistent — until FlashAttention made the whole approximation game look unnecessary.



The Plot Twist: FlashAttention

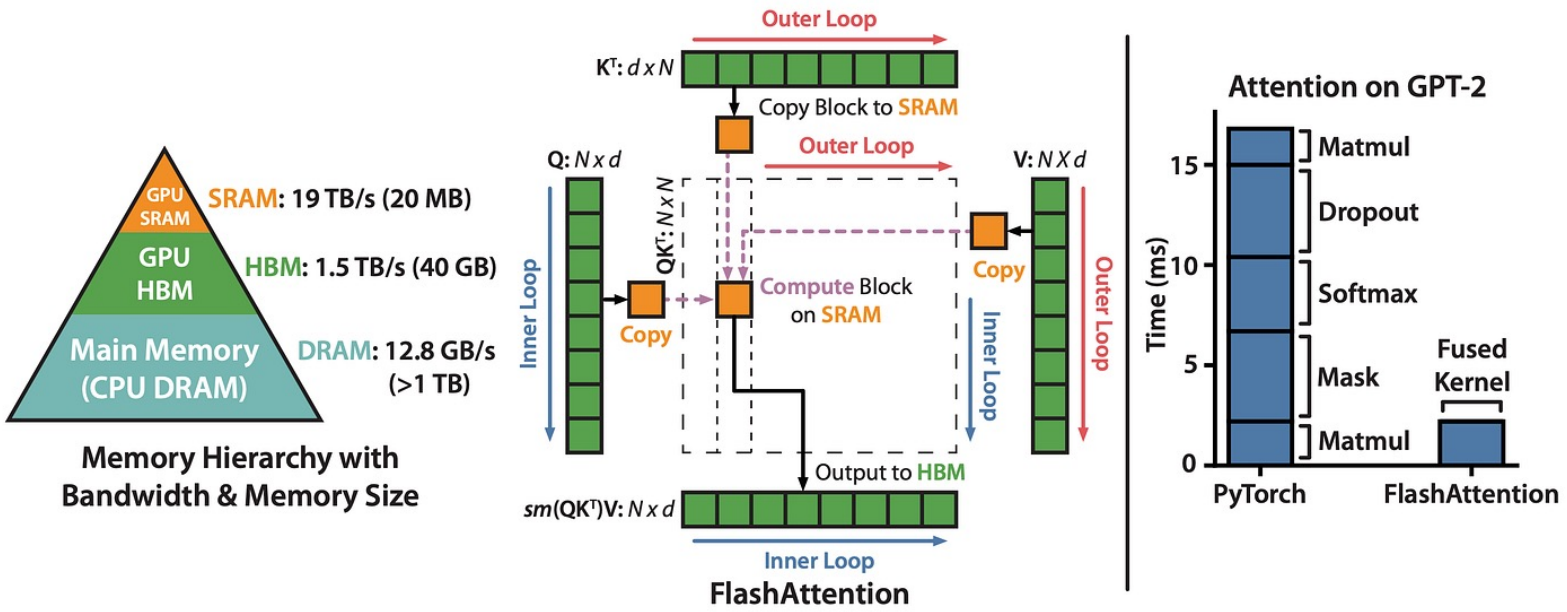
Dao et al. FlashAttention. NeurIPS 2022 / FA-2 (2023) / FA-3 (2024)

Insight: vanilla attention's bottleneck on a GPU isn't FLOPs — it's HBM↔SRAM memory traffic.

- Vanilla materializes the n×n matrix in HBM (slow, 1.5 TB/s).
- FlashAttention tiles Q, K, V; keeps tiles in SRAM (fast, 19 TB/s); uses online softmax to never form the full matrix.

Same math. Same numerical result. 2–4× faster wall-clock, 10–20× less memory.

Now standard everywhere: PyTorch SDPA, vLLM, TensorRT-LLM, every frontier training stack.



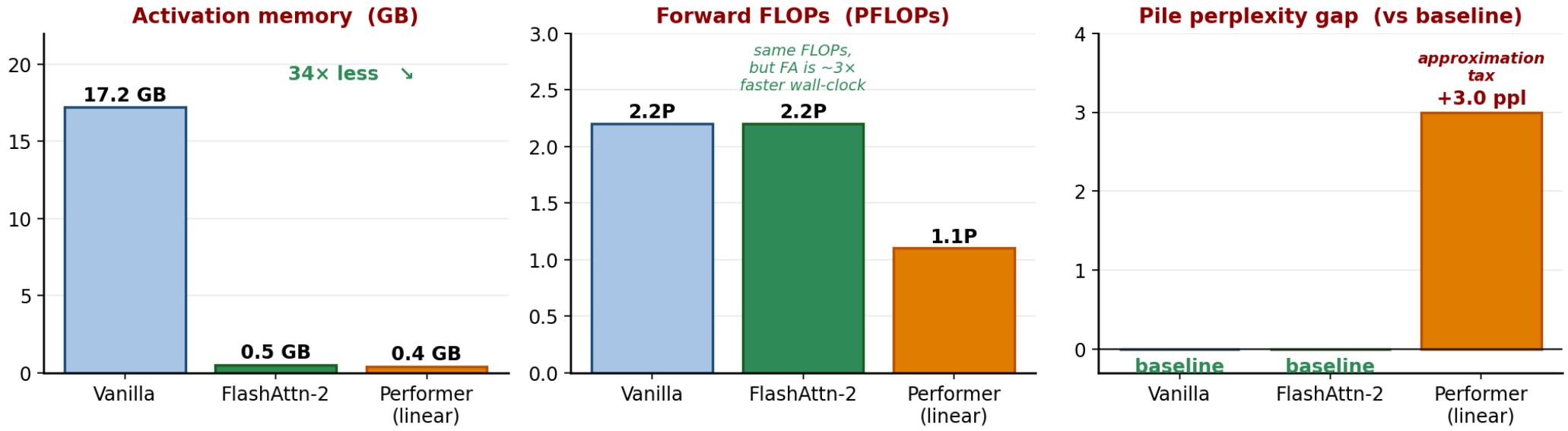
A Back-of-Envelope: Three Knobs Traded

One Llama-2-7B block: n=8192, d=4096, h=32. Forward pass.

- Vanilla: ~17 GB activations, 2.2 PFLOPs
- FlashAttention-2: ~0.5 GB, same FLOPs, ~3× faster wall-clock
- Performer (linear): ~0.4 GB, 1.1 PFLOPs — but ~3 ppl worse on Pile

Takeaway: memory wins from FlashAttention are real and free.

FLOP wins from approximation come with a quality tax that varies by task — and is largest exactly where long context helps.



State-Space Models: Mamba

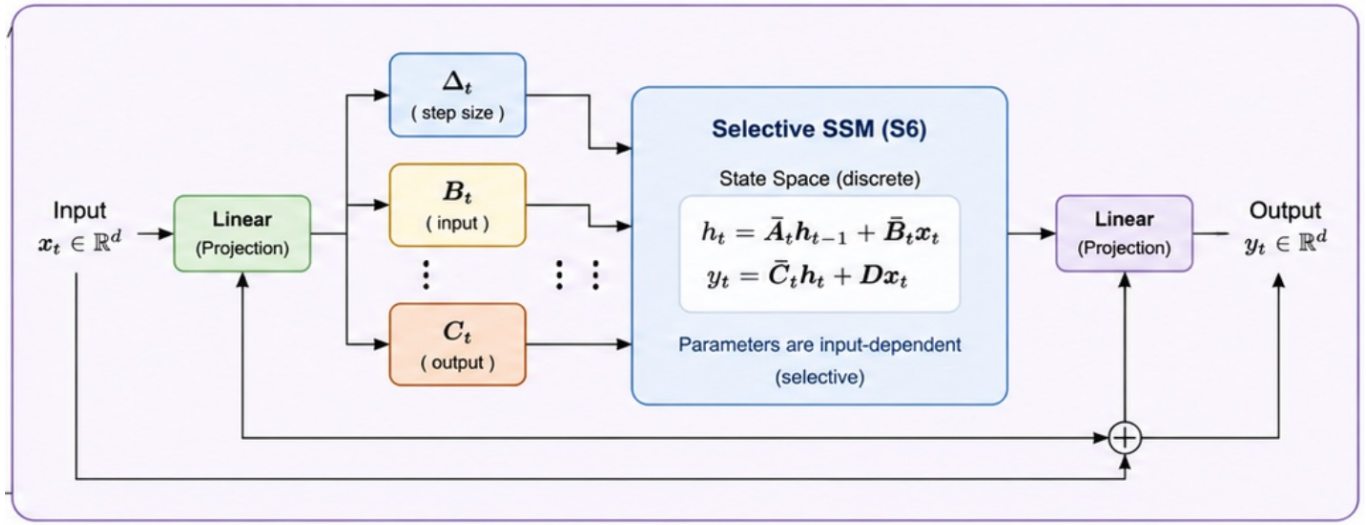
Gu & Dao. Mamba. COLM 2024 (preceded by S4: Gu et al. ICLR 2022)

Different attack on the wall: drop attention entirely. Return to recurrence — done right.

- Selective state-space model: input-dependent A, B, C transitions
- Hardware-aware parallel scan in training; O(1) state per layer at inference (no growing KV cache)
- Matches Transformers up to ~3B params on language; beats them on DNA, audio, very long sequences

Tradeoff: weaker on tasks requiring exact recall over very long context (associative-recall benchmarks). Recurrent state has finite capacity.

Course-arc note: L22's recurrence wasn't dead — it was waiting for the right reformulation.



Hybrids: The 2025–26 Frontier

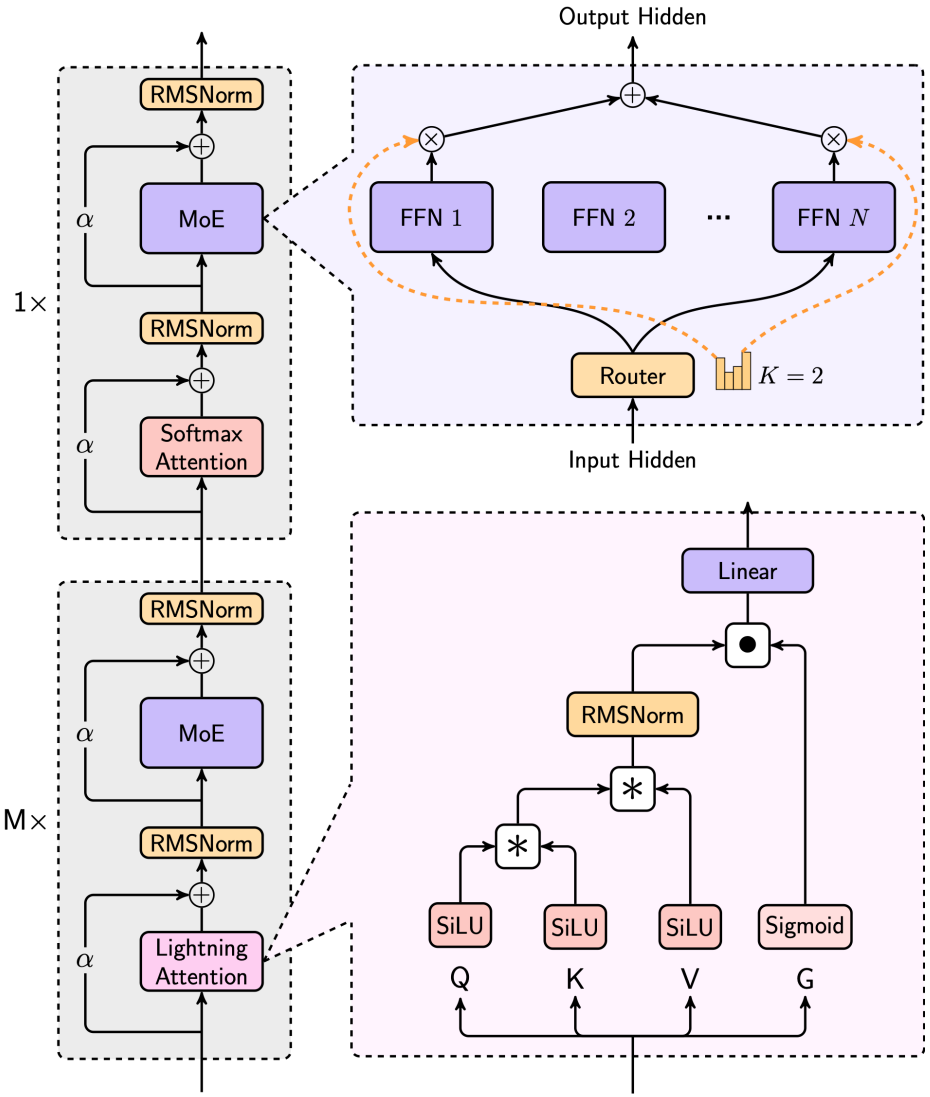
Lieber et al. Jamba (AI21, 2024) / MiniMax-01 (MiniMax, 2025) / Kimi Linear (Moonshot, 2025)

The frontier-lab consensus: don't pick one — stack them.

- Jamba: alternating Mamba and Transformer blocks + MoE → 256k context on a single H100
- MiniMax-01/Kimi Linear: linear-attention layers + occasional full-attention layers — recovers full-attention quality at a fraction of cost

Why it works: cheap layers handle the bulk efficiently; rare full-attention layers handle the long-range retrievals that linear/SSM forget.

Same lesson as W7 pruning (structured + hardware-friendly wins) and W5 HPO (combined methods like BOHB win).



KV Cache: The Wall Behind the Wall

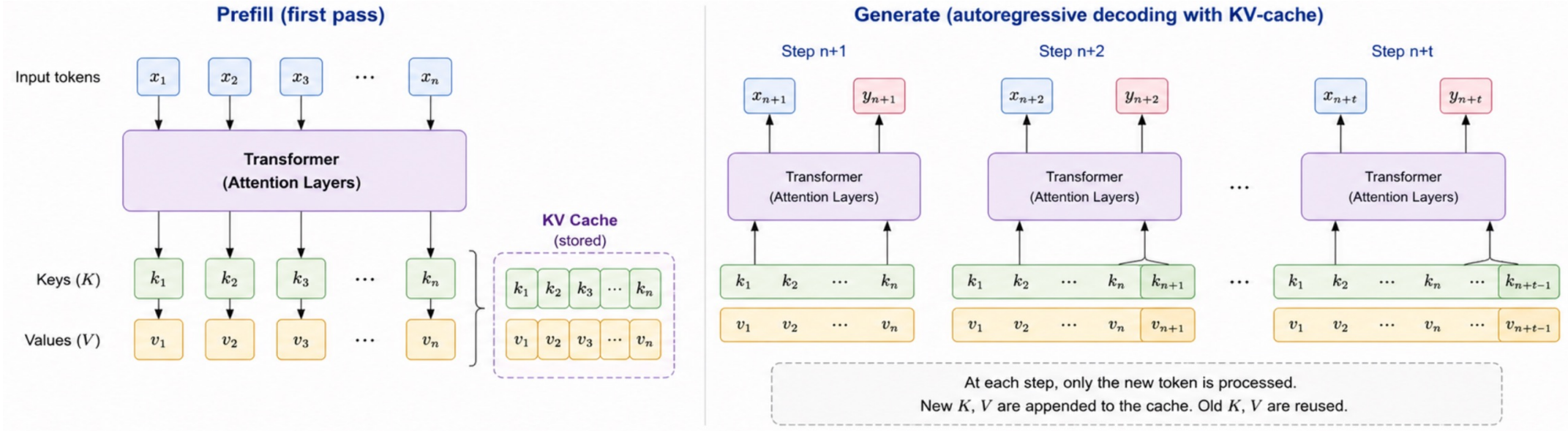
Even with FlashAttention, the KV cache during autoregressive decoding is often the binding constraint.

- Cache size = 2 · n · L · h_kv · d_head · bytes — grows with every generated token.

Fixes deployed in 2026:

- GQA / MQA — share K, V across query heads (Llama-2/3, Mistral, PaLM)
- PagedAttention / vLLM — virtual-memory-style KV management; eliminates fragmentation
- INT8 / INT4 KV quantization — 2–4× compression of the cache itself

These compose: Llama-3-70B at 128k goes from 40 GB → ~3 GB with GQA-8 + INT4 + paging.



References

[1] Beltagy et al. Longformer. arXiv 2020.
[2] Zaheer et al. BigBird. NeurIPS 2020.
[3] Choromanski et al. Performer. ICLR 2021.
[4] Wang et al. Linformer. arXiv 2020.
[5] Dao et al. FlashAttention. NeurIPS 2022.
[6] Dao. FlashAttention-2. ICLR 2024.
[7] Shah et al. FlashAttention-3. NeurIPS 2024.
[8] Gu & Dao. Mamba. COLM 2024.
[9] Lieber et al. Jamba. arXiv 2024.
[10] Ainslie et al. GQA. EMNLP 2023.
[11] Kwon et al. PagedAttention/vLLM. SOSP 2023.