

Mamba & Selective State Space Models (SSM)

Overcoming Transformer Quadratic Attention via Hardware-Aware Associative SRAM Parallel Scans

EXECUTIVE ABSTRACT: Transformer self-attention incurs quadratic $O(N^2)$ compute and memory scaling, causing KV-cache explosion (335GB at 128K context for 70B LLMs) that starves HBM bandwidth. Mamba (S6 SSM) establishes input-dependent state parameters (B, C, Δ) to selectively filter context while maintaining constant $O(1)$ inference time per step. By fusing parallel associative scans inside fast on-chip GPU SRAM, Mamba eliminates KV-caches entirely and delivers 5x higher inference throughput.

Characteristic	Standard Transformer (Llama)	FlashAttention-2 Transformer	Mamba Selective SSM (S6)
Complexity	Quadratic $O(N^2 * d)$	Quadratic $O(N^2 * d)$	Strictly Linear $O(N * d)$
Latency Time	$O(N * d)$ (Linear in Context)	$O(N * d)$ (Linear in Context)	$O(1)$ Strictly Constant Time
KV-Cache	Hundreds of GBs (VRAM Bound)	Hundreds of GBs (VRAM Bound)	ZERO KV-Cache (Fixed Latent State)
Inference Throughput	~10 to 15 tokens / second	~30 to 45 tokens / second	180 to 250+ tokens / second (5x)
Context Length Ceiling	Limited by VRAM (~64K-128K)	Limited by VRAM (~128K-256K)	1,000,000+ Tokens on Single GPU

1. The Quadratic Attention Wall and the KV-Cache Crisis

The foundational bottleneck of multi-head self-attention lies in scaled dot-product computation: $\text{Attention}(Q, K, V) = \text{Softmax}(QK^T / \sqrt{d}) * V$. Materializing the $N \times N$ attention matrix demands quadratic compute and memory. During autoregressive generation, past Key and Value vectors must remain resident in GPU VRAM (the KV-cache).

For a 70-billion parameter model operating at 128,000 tokens, the KV-cache consumes over 335 gigabytes of VRAM. A single user session requires four 80GB Nvidia H100 GPUs solely to hold past context, choking HBM memory buses and running tensor cores at under 2% computational utilization.

2. Foundations of State Space Models and the S6 Selection Breakthrough

Continuous State Space Models map a 1D sequence $x(t)$ through a hidden state $h(t)$: $h'(t) = A \cdot h(t) + B \cdot x(t)$, $y(t) = C \cdot h(t)$. Using Zero-Order Hold (ZOH) discretization, Linear Time-Invariant (LTI) SSMS (such as S4) compute recurrence in $O(1)$ time during inference or 1D convolution via FFT during training.

However, LTI models treat all tokens identically, failing at content-based associative recall. Albert Gu and Tri Dao solved this with S6: making parameters B , C , and Δ dynamic functions of the current input token x_t . The step size Δ_t acts as a continuous reset/filter gate, selectively retaining critical tokens and purging noise.

3. The Hardware-Aware SRAM Fused Scan Kernel

Because input-dependent selection destroys time-invariance, convolution via FFT is mathematically impossible. Naive sequential recurrence would run hundreds of times slower on GPUs. Tri Dao engineered a hardware-aware parallel prefix scan kernel.

Parameters are loaded from slow HBM (3.35 TB/s) into on-chip SRAM (33 TB/s). State expansion ($B \times L \times D \times N$) and Blelloch parallel associative tree scans evaluate entirely within fast SRAM registers before being contracted back to (B, L, D) . The massive expanded state never touches DRAM, eliminating memory bus stalls.

CRITICAL MICROARCHITECTURAL TAKEAWAYS & PRODUCTION GUIDELINES

- Selective State Spaces (S6) eliminate the KV-cache completely, compressing context into a fixed-size latent state vector.
- Hardware-aware fused scan kernels execute parallel associative scans in on-chip SRAM, bypassing HBM bandwidth walls.
- Mamba maintains constant $O(1)$ compute and memory per token during autoregressive generation regardless of context length.
- Hybrid architectures (e.g., Jamba) interleaving 85% Mamba layers with 15% Transformer attention achieve optimal recall.