Sub-Quadratic Recurrent Sequence Models: RWKV-v6 vs. Mamba-3

Eagle Dynamic Token-Shift, Input-Dependent Selective State Spaces, and Tensor-Core Associative Scans

Document ID: LT-WP-2026-1703 | **Classification:** Public Technical Audit | **Release:** October 31, 2026 | **Author:** Lesty Tech Research Lab

EXECUTIVE ABSTRACT: *As context lengths scale beyond 1 million tokens, quadratic self-attention O(N^2) faces prohibitive memory and FLOP scaling limits. RWKV-v6 (Finch/Eagle) and Mamba-3 (Selective State Spaces) represent two leading sub-quadratic architectures. RWKV leverages dynamic multi-head token-shift mechanics and linear recurrent state tracking, whereas Mamba-3 generalizes structured state spaces into 2D tensor-core associative scans. This dossier evaluates their mathematical foundations, GPU kernel operational intensity, and inference memory economics.*

Architectural Dimension	RWKV-v6 (Eagle / Finch)	Mamba-3 (SSM)	Transformer (FlashAttention-3)
Time Complexity (Sequence)	O(N) Linear Time	O(N) Linear Time	O(N^2) Quadratic Time
Inference Memory Complexity	O(1) Constant State	O(1) Constant State	O(N) Growing KV Cache
Token-Mixer Mechanism	Dynamic Token Shift + WKV	Input-Dependent B(t), C(t), Delta	Softmax Causal Attention
1M-Token Context VRAM	3.2 GB Fixed Memory	3.8 GB Fixed Memory	128+ GB KV Cache Footprint

1. RWKV-v6 Eagle: Dynamic Token Shift Formulation

RWKV-v6 reformulates recurrent attention through dynamic token shifting: each token's representation is linearly interpolated with past tokens using input-dependent weights: x_t = x_t + w * (x_{t-1} - x_t). This continuous interpolation operates across multiple heads with vector-valued decay parameters, granting the network high-order memory retention while maintaining purely linear O(N) training parallelization through chunkwise matrix decompositions.

2. Mamba-3: Input-Dependent Selective State Spaces

Mamba-3 extends continuous-time state-space models h'(t) = A * h(t) + B * x(t) by making discretization parameters Delta, B, and C explicit dynamic functions of the current input token x_t. This selective mechanism allows the model to selectively propagate salient tokens into its hidden recurrent state while actively flushing noise, matching dense Transformer recall on associative retrieval tasks without quadratic compute.

3. GPU Kernel Acceleration: Associative Scan vs. GEMM

While traditional Transformers rely heavily on standard GEMM operations on Tensor Cores, recurrent models historically suffered from memory bandwidth constraints during sequential state updating. Modern Flash-Linear and chunked parallel associative scans tile the sequence across GPU SRAM, converting sequential scans into hierarchical intra-chunk GEMMs and inter-chunk state transfers, achieving >70% tensor core utilization.

4. Inference Economics & Needle-in-a-Haystack Benchmarks

At 128k to 1M token context lengths, both RWKV-v6 and Mamba-3 deliver constant O(1) inference memory footprints. While standard Transformers require massive distributed multi-GPU clusters solely to house the expanding KV cache, Mamba-3 and RWKV-v6 execute single-chip 1M token autoregressive decoding, reducing enterprise deployment inference serving costs by over 75%.

Key Strategic Architecture Takeaways

- Both RWKV-v6 and Mamba-3 achieve strictly linear O(N) training compute and constant O(1) inference memory.
- RWKV-v6 Eagle's dynamic token shift enables nuanced local context interpolation without complex discretization steps.
- Mamba-3's input-dependent selectivity solves the classic SSM forgetting problem on needle-in-a-haystack associative recall.
- Chunked parallel associative scans allow linear recurrent models to achieve high GPU Tensor Core operational intensity.