Mamba-2 & State Space Duality (SSD) Architecture

1-Semiseparable Structured Matrix Transformations, GPU Tensor Core Acceleration & Sub-Quadratic Modeling

Author: Lesty Tech Systems Architecture Team | **Date:** November 15, 2026 | **Status:** Validated Engineering Reference | **Target:** Tier-1 Foundry, Systems & Infrastructure Audits

EXECUTIVE ARCHITECTURAL ABSTRACT:

While State Space Models (SSMs) provide linear-time O(N) inference and eliminate the quadratic Key-Value cache bottleneck of transformers, Mamba-1 was bottlenecked by parallel associative scans that failed to saturate systolic GPU Tensor Cores. Mamba-2 introduces State Space Duality (SSD), establishing an exact mathematical equivalence between continuous-time selective SSMs and structured masked linear attention via 1-semiseparable matrix transformations. By reformulating recurrences as chunked dense matrix multiplications, Mamba-2 achieves 55% peak MFU on Hopper Tensor Cores, delivers 3x to 8x faster training, and scales hidden state dimensions up to 256 channels with zero throughput penalty.

KEY PHYSICAL PARAMETERS & EXPERIMENTAL BENCHMARK TELEMETRY

Architecture Parameter	Transformer (Self-Attention)	Linear Attention (RWKV)	Mamba-1 (Selective SSM)	Mamba-2 (SSD Architecture)
Inference Complexity	O(N) KV Cache Growth	O(1) Constant State	O(1) Constant State	O(1) Constant State
Training Time Complexity	O(N^2) Quadratic FLOPs	O(N) Linear Time	O(N) Linear Time	O(N) Linear Time
Primary Compute Kernel	Dense GEMM (Tensor Cores)	Custom Fused Kernel	Parallel Associative Scan	Dense Block GEMM (Tensor Cores)
Hopper Peak MFU Efficiency	55% - 62% MFU	28% - 34% MFU	18% - 24% MFU	54% - 58% MFU (Rivals Transformers)

1. The Compute vs. Memory Wall in Foundation Models

Transformer self-attention incurs quadratic computational growth and massive KV cache memory footprints. While Mamba-1 achieved linear inference, its parallel associative scan kernel achieved an arithmetic intensity of less than 15 FLOP/byte, leaving over 70% of Tensor Core compute capacity idle.

2. State Space Duality (SSD) Mathematical Equivalence

Mamba-2 proves that 1-D selective state space recurrences correspond to 1-semiseparable matrix operators, which factor into outer products of projection vectors modulated by exponential decay. This reformulates scans into dense matrix multiplications.

TECHNICAL ANALYSIS CONTINUED: MAMBA-2 & STATE SPACE DUALITY (SSD) ARCHITECTURE

3. Chunked Block-Decomposition on Tensor Cores

Sequences are partitioned into 64-token tiles. Intra-chunk attention executes as dense WMMA GEMMs on Tensor Cores, while inter-chunk boundary states propagate via lightweight linear recurrences, maximizing arithmetic intensity.

4. State Expansion & Hybrid Foundation Model Scaling

Tensor Core acceleration allows hidden state dimension N to scale from 16 up to 256 without speed penalties. Interleaving Mamba-2 layers with Sliding Window Attention in hybrid models yields near-infinite context scaling with dense transformer fidelity.

STRATEGIC ENGINEERING TAKEAWAYS & DEPLOYMENT CHECKLIST:

- State Space Duality: Unifies continuous-time state spaces with structured linear attention via 1-semiseparable matrices.
- Hardware Saturation: Delivers 55% MFU on Hopper Tensor Cores, running 3.2x to 7.8x faster than Mamba-1 parallel scans.
- Zero KV Cache: Sustains true O(1) memory footprint during autoregressive token generation across multi-million contexts.
- Expanded Memory State: Scales hidden state capacity up to 256 channels, mastering complex long-context retrieval benchmarks.