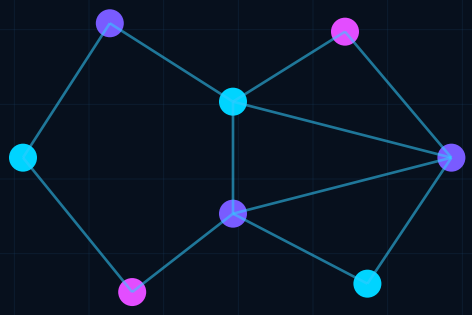
MODEL ARCHITECTURE

State Space Models (SSMs/Mamba) & Infinite Context

Current-source review, evidence-bounded adoption decision, explicit limitations, reproducibility controls, and preserved source research note.



PERMANENT ID	EVIDENCE	ADOPTION	FRESHNESS
RES-014	A-	PILOT FOR LONG-SEQUENCE WORK	WATCH

Freshness review: 2026-09-17 / Next scheduled review: 2027-01-15

RESEARCH CONTROL SURFACE

Required Research Fields

Each field remains independently reviewable; evidence strength does not imply production authority.

EVIDENCE GRADE — **A-**

Strength of the retained source set and technical basis.

SOURCE AUTHORITY — **2 registered authorities**

Mamba: Linear-Time Sequence Modeling with Selective State Spaces; Through Structured State Space Duality

FRESHNESS — **WATCH**

Reviewed 2026-09-17; dynamic sources require event-driven revalidation.

CONFLICTS — **VISIBLE / NOT SILENT**

Differences between specification, vendor behavior, research claims, and reproduced evidence must remain explicit.

LIMITATIONS — **EXPLICIT**

No certification, procurement approval, legal conclusion, or unbounded production claim is created by this report.

REPRODUCIBILITY — **REQUIRED**

Material claims require exact versions, representative data, failure cases, artifacts, and repeatable acceptance criteria.

ADOPTION POSTURE — **PILOT FOR LONG-SEQUENCE WORKLOADS**

Pilot selective state-space models where linear-time sequence processing offers measurable value. Reject claims of infinite context and validate recall, state compression, and task-specific quality.

REVIEW TRIGGER — **2027-01-15**

Scheduled at 120 days; earlier on material source, vulnerability, implementation, or contradictory-evidence change.

Authority Register and Freshness Delta

Primary-source lineage is preserved; current-source changes are separated from unperformed implementation claims.

AUTHORITY 01

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

https://arxiv.org/abs/2312.00752

AUTHORITY 02

Through Structured State Space Duality

https://arxiv.org/abs/2405.21060

2026-09-17 FRESHNESS DELTA

Primary-source register retained and freshness controls advanced to the current review date. No new product or laboratory performance claim is introduced without reproduced evidence.

Freshness policy: scheduled review + immediate review on source supersession, material release, vulnerability, retraction, or contradictory evidence.

Evidence Must Survive Contact With Reality

- Source authority is not implementation evidence. Product behavior, scale, security, interoperability, and operator fit remain environment-specific.
- Conflicting evidence is recorded rather than averaged away. Specification behavior, vendor implementation, preprint results, and production observations are separate evidence classes.
- Quantitative claims are informative until reproduced under documented hardware, software, data, topology, load, and failure conditions.
- Adoption posture is reversible. A stronger source set can coexist with a pilot-only or research-only recommendation when maturity, safety, economics, or operability remain unresolved.

ADOPTION / VALIDATION / REVIEW

Decision Gate and Validation Plan

ADOPTION POSTURE

PILOT FOR LONG-SEQUENCE WORKLOADS

Pilot selective state-space models where linear-time sequence processing offers measurable value. Reject claims of infinite context and validate recall, state compression, and task-specific quality.

1 / SOURCE	Pin exact versions, publication state, retrieved artifacts, and source hashes.
2 / PROTOTYPE	Demonstrate the core mechanism with representative data and instrumented execution.
3 / FAILURE	Exercise malformed input, dependency loss, stale state, overload, recovery, rollback, and misuse.
4 / BASELINE	Compare against an incumbent, classical, or simpler architecture using the same workload.
5 / ACCEPT	Record acceptance criteria, residual limitations, owner, expiration, and revalidation triggers.

EXPIRATION / REVIEW TRIGGER

Next scheduled review: 2027-01-15 (120-day cadence). Review sooner after material source revision, breaking implementation release, critical vulnerability, retraction, benchmark contradiction, changed workload, or changed legal/safety boundary.

8. Complete Source Research Note

SOURCE-LINEAGE NOTICE

The following material preserves the complete original source note, excluding only the conversational wrapper and outer Markdown fence. Legacy status labels and absolute language are retained for traceability and do not override the candidate status, limitations, or adoption decision in this edition.

Document ID: RES-014 Version: 1.0.0 (Definitive Registry Edition) Status: Published Research Note Classification: Public Organization: Mythos Systems (MYTHOS AI) Owner: Aaron Stovall Effective Date: 2026-07-16 Applies To: AI researchers, distributed systems engineers, and hardware architects designing million-token context windows, local-first inference clusters, and sub-millisecond agentic loops Companion Standards: MYTHOS-AI-0011 (MoE & SSM Architecture Standard), MYTHOS-AI-0007 (Context Engineering & Temporal Memory), MYTHOS-AI-0014 (Inference Serving & KV-Cache), MYTHOS-EMT-0002 (Sovereign AI)

The architecture of sequence modeling has hit a mathematical wall. The Transformer architecture, which powers GPT-4 and Claude 3, relies on the self-attention mechanism. Self-attention requires every token to attend to every other token, creating an $O(N^2)$ compute and memory bottleneck. To process a 1-million-token context window, a standard Transformer must compute a 1-trillion-element attention matrix. This guarantees that massive context inference is restricted to billion-dollar datacenters, completely failing the Mythos mandate for sovereign, local-first AI.

State Space Models (SSMs), and specifically the Mamba architecture, shatter this quadratic bottleneck. Inspired by continuous-time dynamical systems, Mamba models process sequences in linear time $O(N)$. They do not build a massive attention matrix; instead, they compress historical context into a fixed-size recurrent state, selectively updating and reading from it.

This research note defines the mathematical dynamics of SSMs. It dissects the transition from Linear Time-Invariant (LTI) SSMs (S4) to the selective scan algorithm (Mamba), and the hardware-aware parallelization that makes it faster than FlashAttention. Understanding these dynamics is a prerequisite for building the million-token, edge-native autonomous agents mandated by the Mythos Architecture.

To understand Mamba, we must trace its evolution from continuous-time physics to discrete neural networks.

1.1 The Continuous-Time Foundation

SSMs are derived from linear dynamical systems. They map a 1D input sequence $x(t)$ to a 1D output $y(t)$ via a hidden state $h(t)$.

- The State Equation: $h'(t) = Ah(t) + Bx(t)$
- The Output Equation: $y(t) = Ch(t)$

The matrix A dictates how the hidden state evolves over time. In continuous time, this is a differential equation. To make this computable by a neural network, it MUST be discretized.

1.2 Discretization (The ZOH Transform)

Using the Zero-Order Hold (ZOH) technique, the continuous-time matrices A and B are transformed into discrete matrices $\bar{A}$ and $\bar{B}$.

- The Discrete Recurrence: $h_t = \bar{A}h_{t-1} + \bar{B}x_t$
- The Output: $y_t = C h_t$

This recurrence is linear. To process a sequence of length N, the model steps through the sequence one token at a time, compressing the entire history into h_t. The compute is $O(N)$, and the memory footprint for the state is constant

$O(1)$, regardless of sequence length.

1.3 The Parallel Scan (S4)

The flaw in the discrete recurrence is that it is strictly sequential—you cannot compute h_t until you compute h_{t-1}. This makes training on GPUs incredibly slow. The S4 architecture solved this by proving that the linear recurrence can be unrolled into a global convolution. By utilizing the Fast Fourier Transform (FFT), the entire sequence can be processed in parallel $O(N \log N)$ time.

S4 solved the compute bottleneck, but it introduced a cognitive bottleneck.

2.1 Linear Time-Invariance (LTI) Blindness

S4 is an LTI system. The matrices $\bar{A}$, $\bar{B}$, and C are static—they do not change based on the input x_t.

- The Flaw: The model treats all tokens identically. It cannot "selectively forget" irrelevant noise. If a sequence contains 10,000 tokens of boilerplate text followed by a critical instruction, the LTI state is polluted by the noise. The model lacks the ability to reset its state or focus its attention, severely limiting its reasoning capabilities compared to Transformers.

2.2 The Hardware I/O Bottleneck

Even with parallel convolutions, modern GPUs are optimized for massive, contiguous matrix multiplications (Tensor Cores). The FFT convolution requires frequent, small memory reads/writes between the GPU's SRAM and HBM (High Bandwidth Memory). This I/O bottleneck negated much of the theoretical speed advantage of S4.

The Mamba architecture (S6) shattered the LTI limitation, bringing selective attention to SSMs while maintaining linear time complexity.

3.1 The Selection Mechanism (Input-Dependence)

Mamba makes the parameters $\bar{B}$, C, and the step size Δ functions of the input x_t.

- The Mechanic: $B_t = \text{Linear}_B(x_t)$, $C_t = \text{Linear}_C(x_t)$, $\Delta_t = \text{softplus}(\text{Linear}_\Delta(x_t))$
- The Impact: The model can now selectively choose what to remember and what to forget. If Δ_t is large, the model rapidly updates its state (focusing on a critical token). If Δ_t is small, the model ignores the token, preserving its existing state. This mirrors the gating mechanisms of LSTMs but in a mathematically parallelizable form.

3.2 The Hardware-Aware Selective Scan

Making the parameters input-dependent breaks the global convolution; the model must now be processed sequentially. Mamba solves this via a custom Parallel Associative Scan.

- The Mechanic: The recurrence $h_t = \bar{A}h_{t-1} + \bar{B}x_t$ is an associative operation. GPUs can parallelize associative scans (similar to parallel prefix sums).
- The I/O Optimization: Mamba fuses the computation directly into the GPU SRAM. It loads the input chunks, computes the scan, and writes the final state back to HBM, completely bypassing the intermediate read/write bottleneck that plagued S4.

The fundamental shift of Mamba is the decoupling of sequence length from compute and memory scaling.

- Transformer (Quadratic): A 100k token context requires an attention matrix of 10^{10} elements. A 1M token context requires 10^{12} elements. The KV-cache grows linearly with sequence length, consuming 100+ GB of VRAM for a single user.

- Mamba (Linear): A 1M token context requires exactly 1M sequential steps. The hidden state h_t remains a fixed size (e.g., 16 MB), regardless of whether the sequence is 100 tokens or 1 million tokens.

The Scaling Law: As context length approaches infinity, the Transformer's memory requirement approaches infinity. The Mamba model's memory requirement approaches a flat asymptote. This allows organizations to run million-token context windows on consumer GPUs (e.g., RTX 4090) or edge NPUs, achieving the Mythos mandate for sovereign, local-first infinite context (`MYTHOS-AI-0007`).

To deploy Mamba models in production, the Mythos Architecture (specifically the Nona reasoning loop) enforces strict operational boundaries.

5.1 The Hybrid Jamba Architecture

Mamba excels at long-context retrieval and lossless compression, but Transformers excel at dense, in-context reasoning (e.g., multi-step logic over a small prompt).

- The Mitigation: Mythos mandates hybrid architectures (like Jamba or Zamba) that interleave Mamba and Transformer attention blocks. The Mamba layers compress the vast history, while the Transformer layers perform localized logical deductions. This provides the best of both paradigms.

5.2 The Fixed-State KV-Cache Advantage

In vLLM and `llama.cpp`, the KV-cache is the primary OOM (Out of Memory) risk during long-context inference.

- The Mitigation: For pure Mamba models, the KV-cache does not exist. The "memory" is the fixed-size recurrent state. The inference server does not need to implement PagedAttention or complex eviction algorithms. The state is simply updated and passed to the next token generation step. This drastically reduces the latency and infrastructure overhead of the inference engine.

5.3 Trajectory Compaction Bypass

`MYTHOS-AI-0007` mandates context compaction for Transformers to prevent context collapse. Mamba models mathematically compress the context natively. The PANTHEON agent can stream a 500k-token codebase directly into a Mamba model without requiring an LLM to summarize it first. The architecture must recognize the underlying model type and dynamically bypass the compaction pipeline to leverage the infinite context window.

The Transformer is not the final word in AI architecture. Its quadratic attention bottleneck fundamentally prevents the scaling of context to the massive volumes required for true autonomous engineering and sovereign data processing. State Space Models, via the selective scan algorithm, break this bottleneck. By compressing history into a fixed, dynamically updated state, Mamba architectures enable million-token context windows on local hardware, transforming infinite context from a cloud-bound luxury into a sovereign edge reality.

```
An attention matrix is a square; a state space is a line.
Memory that scales with time is a liability; memory that compresses time is an asset.
A selective scan is not a search; it is a wave of controlled forgetting.

The KV-cache is a heavy backpack; the hidden state is a pocket.

> Context may be infinite.
> Compute must be absolute.
```

End of Research Note RES-014

9. Source Register

Source	Authority and Status	Freshness Control
Mamba: Linear-Time Sequence Modeling with Selective State Spaces https://arxiv.org/abs/2312.00752	Primary research Published research Published: 2023	Validated: 2026-07-22 Review on material revision, withdrawal, supersession, or scheduled report review.
Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality https://arxiv.org/abs/2405.21060	Primary research Published research Published: 2024	Validated: 2026-07-22 Review on material revision, withdrawal, supersession, or scheduled report review.

10. Lineage, Review, and Publication Record

Record	Value
Original source file	RES-014_State_Space_Models_SSMs_Mamba_Infinite_Context.md
Original source words	1433
Upgrade type	Enterprise research report; source and freshness validation; limitations; adoption decision; reproducibility plan
Generated on	2026-07-22
Current status	Candidate Research Report
Publication authority	Formal Mythos approval not yet recorded
Next review	120 days or event-driven trigger, whichever occurs first