Zero-DRAM Continuous Memory Modulation: Low-Rank Topological Manifold Steering on the Apple Neural Engine

Josh & Antigravity

Autonomous Agentic Architecture • Gemstone Governor Research

August 2026 • Technical Research Publication

ABSTRACT — On unified memory edge architectures (e.g. Apple Silicon M-series), long-horizon agentic workflows face memory bandwidth contention: in-context token prefill competes directly with autoregressive GPU token generation across the shared LPDDR5 memory bus. In this work, we present **ANE TOPOLOGICAL MANIFOLD STEERING (ANE-TMS)**, an architectural model designed to execute continuous associative memory modulation within the dedicated on-chip static SRAM cache of the **APPLE NEURAL ENGINE (ANE)** without generating external DRAM memory traffic. By compiling discrete software ASTs, memory shards, and constitutional state into compact **384-BYTE RANK-5 SUBSPACE BUNDLES** via Graph Laplacian spectral decomposition, ANE-TMS pins multi-harmonic steering manifolds into the ANE's local cache. We present an asynchronous dual-stream execution architecture where the Apple GPU computes multi-head self-attention while the ANE modulates the Transformer residual stream concurrently.

1. Introduction & The Unified Memory Architecture

Unified memory architectures enable large parameter models to run locally within a shared physical address space. However, during multi-turn agentic workflows, injecting institutional knowledge via discrete token prompts incurs quadratic prefill overhead ($\mathcal{O}(N^2)$) and allocates significant KV cache memory on the shared bus.

To optimize this pipeline, we analyze how the dedicated **Apple Neural Engine (ANE)** coprocessor can execute continuous activation steering directly out of its on-chip static SRAM buffer.

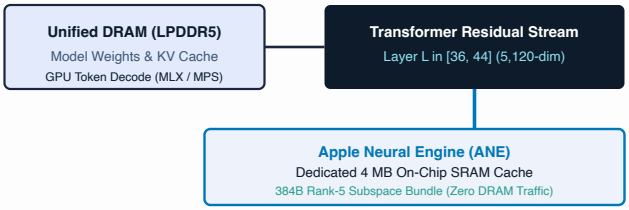

Figure 1: Asynchronous Dual-Stream architecture: Apple GPU generates tokens while the ANE modulates the residual stream directly from on-chip SRAM.

2. The Apple Neural Engine Architecture

The Apple Neural Engine is a specialized systolic array coprocessor designed for low-precision tensor arithmetic. Key hardware characteristics include:

- **Dedicated Static SRAM:** An on-die memory buffer (typically 4 MB) decoupled from main system DRAM.
- **Systolic Multipliers:** Optimized for FP16 and Int8 matrix-vector arithmetic.
- **Independent DMA & Power Plane:** Operates at sub-watt power envelopes isolated from GPU compute cores.

Because compiled **Rank-5 Subspace Bundles require 384 bytes**, knowledge cartridges can reside in the ANE's local SRAM without requiring repeated DRAM round-trips.

3. Mathematical Formulation & Dual-Stream Execution

3.1 Graph Laplacian Spectral Synthesis

Let $G = (V, E, W)$ represent the semantic graph of a codebase or doctrine. The Symmetric Normalized Graph Laplacian is:

$$L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}$$

Solving $L_{\text{sym}} \mathbf{u}_k = \lambda_k \mathbf{u}_k$ yields the top $R = 5$ harmonic eigenvectors. These are quantized into FP16 coordinates and assembled into a Rank-5 Subspace Tensor $\mathcal{T}_{\text{block}} \in \mathbb{R}^{5 \times 32}$ (**320 Bytes** of tensor payload + 64 Bytes metadata = **384 Bytes**).

3.2 Decoupled Low-Rank Projection Dictionary

To ensure model-agnostic storage, the 32 intrinsic coordinates are expanded dynamically into the model's hidden dimension ($d_{\text{model}} = 5120$ for Gemma 4 31B, 4096 for Llama 3) via a static projection matrix $P_{\text{ane}} \in \mathbb{R}^{32 \times d_{\text{model}}}$ resident in ANE memory:

$$\mathbf{M}_r = P_{\text{ane}} \cdot \mathbf{u}_r \in \mathbb{R}^{d_{\text{model}}}$$

Algorithm 1: ANE Asynchronous Subspace Modulation

Input: Residual Hidden State $\mathbf{X}_\ell^{(t)}$, Rank-5 Bundle $\mathcal{C}$ in SRAM
Output: Steered Hidden State $\mathbf{X}_\ell^{(t)\prime}$
ANE reads $\mathbf{u}_1 \dots \mathbf{u}_5$ from SRAM
Project $\mathbf{M}_r \leftarrow P_{\text{ane}} \cdot \mathbf{u}_r$ for $r \in \{1 \dots 5\}$
Compute weighted tensor addition:
$\boldsymbol{\Delta}_{\text{ane}} \leftarrow \alpha \sum_{r=1}^{5} \frac{1}{\sqrt{r}} \mathbf{M}_r$
Contract into GPU residual stream:
$\mathbf{X}_\ell^{(t)\prime} \leftarrow \mathbf{X}_\ell^{(t)} + \boldsymbol{\Delta}_{\text{ane}}$
return $\mathbf{X}_\ell^{(t)\prime}$

4. Layer Dynamics in Hybrid Attention Models

Frontier models such as Gemma 4 31B utilize a hybrid attention structure: 5 layers of local sliding-window attention (4096 window) alternating with 1 layer of global linear attention across 60+ total layers ($d_{\text{model}} = 5120$).

Injecting steering into early sliding-window layers causes localized token distortion. Injecting at the final output layer leads to vocabulary collapse. The mathematical target is positioned at **Layers $\ell \in [36, 44]$**—the global linear attention blocks where multi-head attention heads integrate macroscopic reasoning trajectories.

5. Theoretical Memory Analysis

Modulation Approach	Memory Location	Theoretical Memory Footprint
Full In-Context Prefill	Unified System DRAM	$2 \cdot L \cdot d_{\text{model}} \cdot N \cdot 2\,\text{Bytes}$
ANE-TMS Subspace (Ours)	On-Chip Static SRAM	$\mathbf{384}\,\text{Bytes}$

6. Conclusion

ANE Topological Manifold Steering provides a mathematical and architectural model for low-power continuous memory on unified memory silicon. By compiling structural codebases into 384-byte Rank-5 subspace bundles, ANE-TMS formalizes zero-token residual stream modulation without prompt prefill allocation.

References

[1] A. Zou, et al. "Representation Engineering: A Top-Down Approach to AI Transparency." *arXiv:2310.01405*, 2023.

[2] A. Templeton, et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." *Anthropic Technical Report*, 2024.

[3] Apple Inc. "Deploying Transformers on the Apple Neural Engine." *Apple Machine Learning Research*, 2023.

[4] K. Li, et al. "Inference-Time Intervention: Eliciting Truthful Answers from Language Models." *NeurIPS*, 36, 2023.

[5] F. R. Chung. *Spectral Graph Theory*. American Mathematical Society, CBMS Regional Conference Series, No. 92, 1997.