On the Expressive Power and Limitations of Multi-Layer SSMs

NeurIPS 2026 / TMLR Journal-to-Conference

Nikola Zubić[1], Qian Li[2], Yuyi Wang[3] and Davide Scaramuzza[1]

[1] University of Zurich [2] Shenzhen Research Institute of Big Data [3] CRRC Zhuzhou Institute

University of Zurich[UZH]

ROBOTICS & PERCEPTION GROUP

Motivation: SSMs compress long sequences into a small state. What can depth, width and precision buy? Function composition exposes sharp limits, and shows why the timing of thought tokens matters.

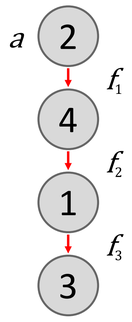
Task: compute $f_k(\cdots f_1(a))$ from ordered lookup tables

$N = 4,\ k = 3,\ a = 2$

j	1	2	3	4
$f_1(j)$	3	**4**	1	2
$f_2(j)$	2	4	3	**1**
$f_3(j)$	**3**	1	4	2

Read f_1, then f_2, then f_3

Explicit Lookup Tables

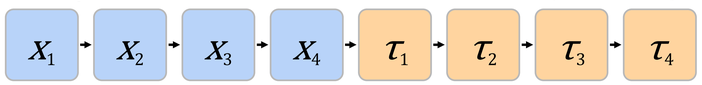


Sequential Composition

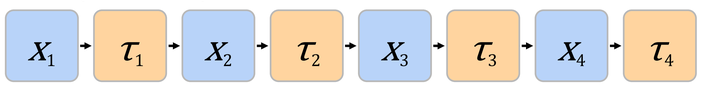
Key Insight: When the Model Thinks

- **Post-input CoT:** only processes the final state. The communication lower bound remains.
- **Online CoT:** feeds the state back through thought tokens before the next input arrives.
- **One thought per input:** exact composition with one layer and O(log *N*) state bits.

Post-input CoT — x: input τ: thought



Online (interleaved) CoT



SSMs as Communication

- **Model:** *L* layers, width *d*, *p* bits per scalar. Each state update is affine.
- **Block summary:** $h \mapsto Ah + b$ needs $d^2 + d$ scalars.
- **Communication:** *L* layers yield *L* forward rounds with $O(d^2p)$-bit messages.
- **Lower bound:** serialize into *L* + 1 messages, then apply pointer chasing.

Each input block is one affine map

Block 1 → Block 2 → Block 3

(A_1, b_1) (A_2, b_2) (A_3, b_3)

$$h_{\text{out}} = A^{\text{blk}}\, h_{\text{in}} + b^{\text{blk}}$$

$d^2 + d$ scalars, p bits each

L-layer SSM: layer 1 ⋮ layer L → L-round protocol: round 1 ⋮ round L

Pointer chasing, with k = L + 3

$$d^2p = \Omega(N/L^3)$$

Error probability at most 1/3

Main Theoretical Results

A Four-Layer Gap for Function Composition

Same task: $k = L + 3$

L layers

$$d^2p = \Omega(N/L^3)$$

Lower bound

Error at most 1/3

L + 4 layers

$$d = 1$$
$$p = \Theta(\log N)$$

Exact construction

Logarithmic bits per layer

Reasoning	Layers	Resource result
No CoT	L	$d^2p = \Omega(N/L^3)$
Post-input CoT	L	$d^2p = \Omega(N/L^3)$
Online CoT	**1**	$dp = O(\log N)$

First two rows: $k = L + 3$. Online CoT: any k, exactly one thought token per input.

Width and Precision

Equal bit budgets

1 bit | 1 bit | 1 bit vs. 3 bits

width 3, p = 1 width 1, p = 3

No universal simulation

Exact, step-preserving simulation in the base affine-state ring model

Online CoT

Total persistent memory

$$O(Ldp + \log T_n)$$

Width–precision simulation via streaming memory

T_n: worst-case processed steps

Check out our Paper and Proofs!

Paper link | Full proofs | ML research at RPG | Nikola Zubić's homepage

Sponsors: Horizon2020 European Union Funding for Research & Innovation; erc

Online CoT Recovers Streaming Power

- **SSM to streaming:** simulate the state stack using O(Ldp + log T□) persistent bits.
- **Streaming to SSM:** one layer simulates any deterministic S-bit streaming algorithm when dp ≥ S.
- **One thought token per input:** exactly 2n recurrent updates for n input tokens.

Takeaways and Scope

- Depth and thought-token timing are distinct computational resources.
- Results concern generalized finite-precision SSMs on explicit-table composition.
- Upper constructions use vector-valued thoughts. Learnability and hardware efficiency remain open.