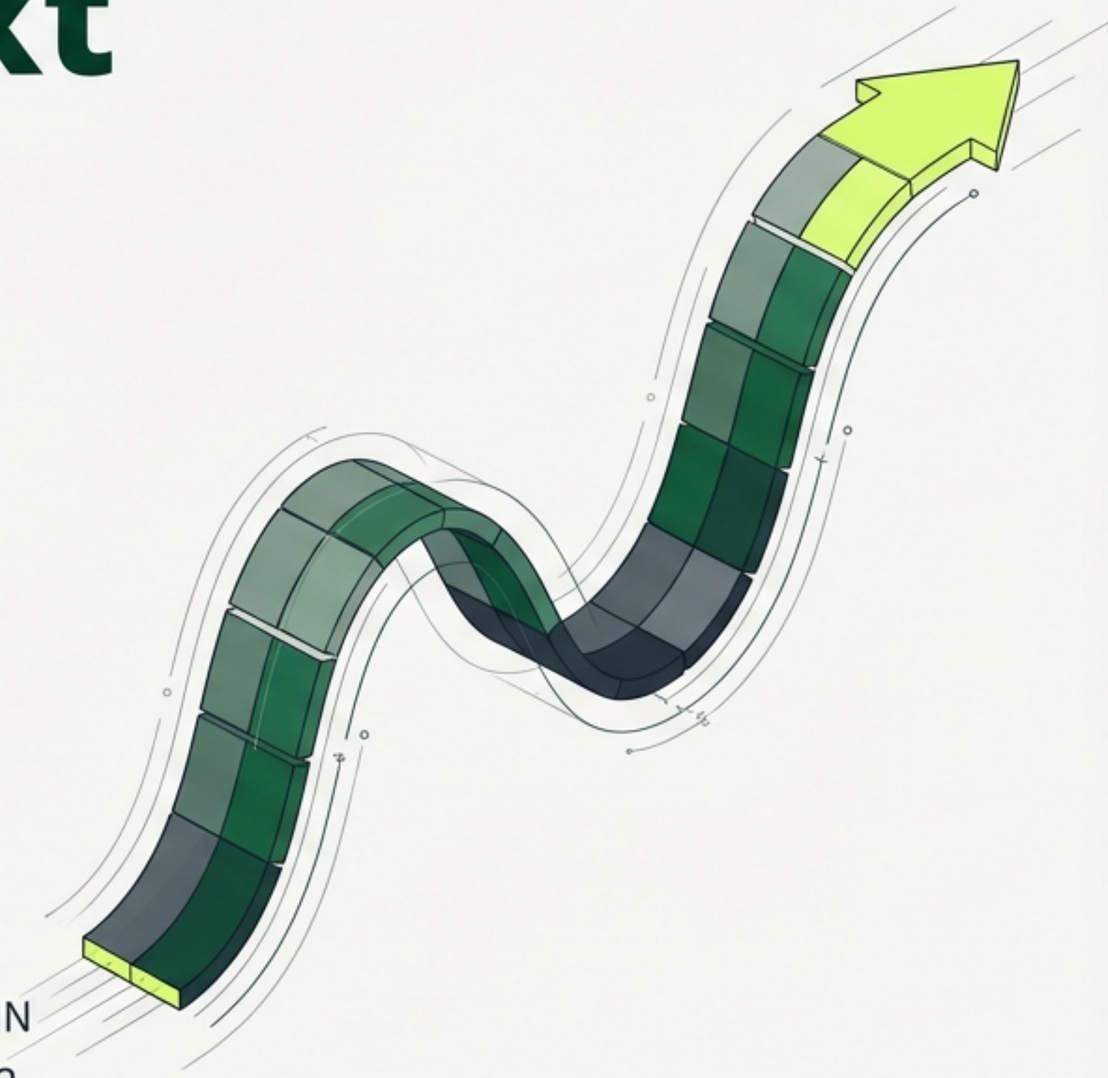
Mamba: The Next Evolution of Sequence Modeling

คลื่นลูกใหม่แห่งวงการ AI



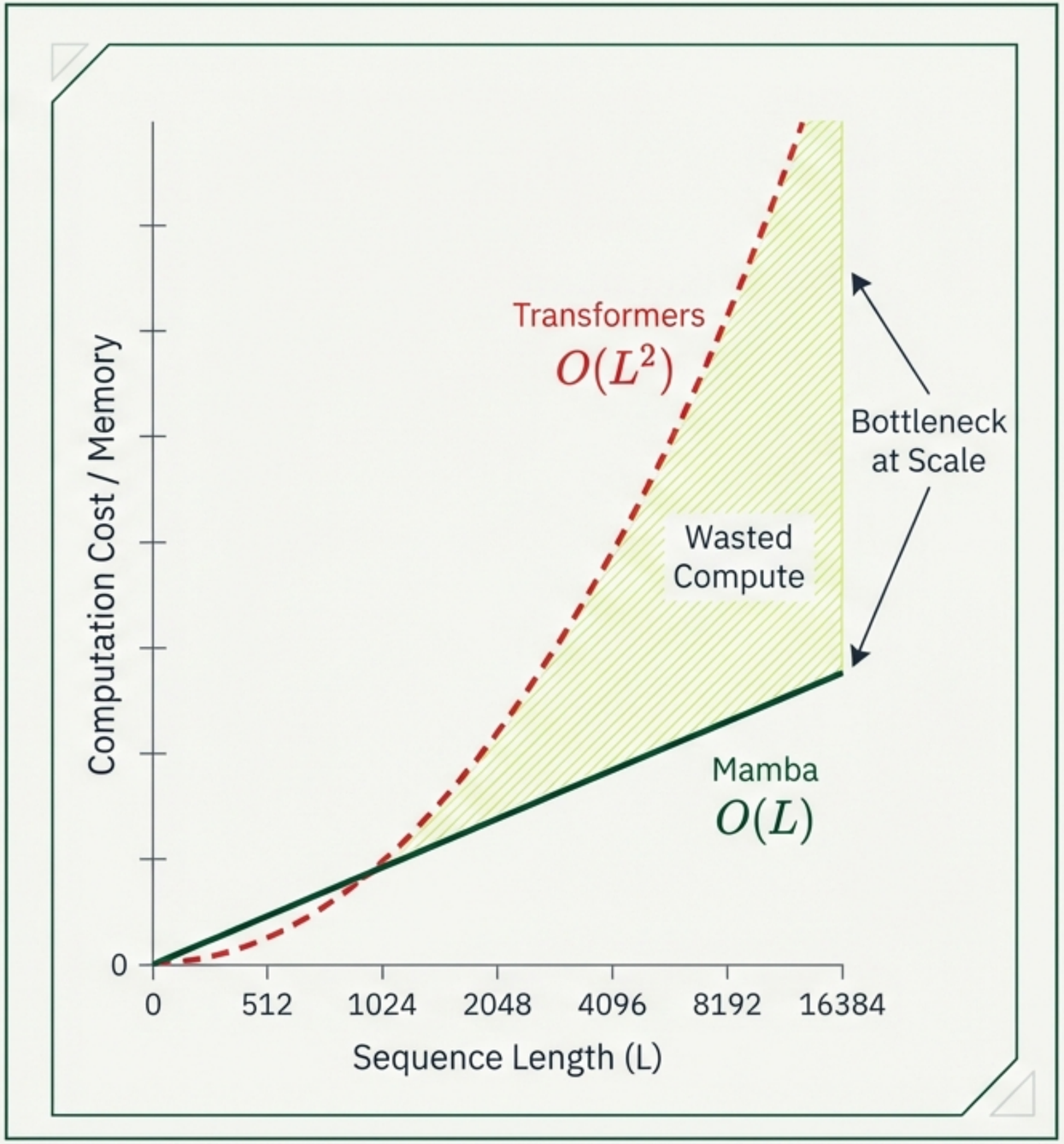
- Beyond Transformers: สถาปัตยกรรมใหม่ที่ไม่พึ่งพา Attention แบบเดิม
- The Best of Both Worlds: ผสานความเร็วของ RNN และความแม่นยำของ CNN
- The Promise: ประมวลผลข้อมูลระดับ Infinite Context ด้วยประสิทธิภาพสูงสุด

NotebookLM

The Bottleneck: เมื่อความฉลาดแลกมาด้วยต้นทุน

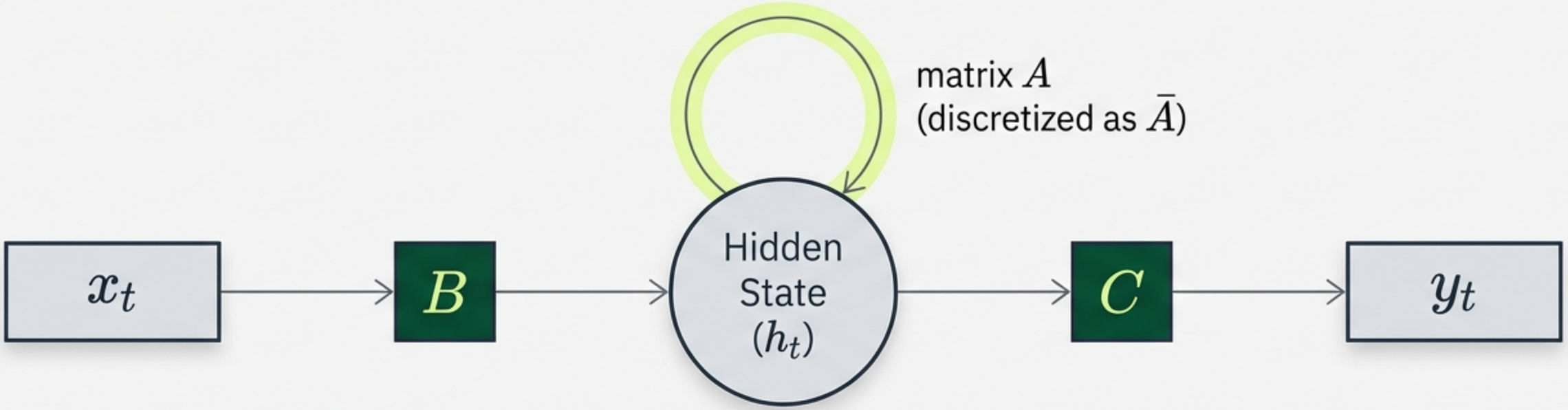
Quadratic Cost vs. Linear Efficiency

- **The Attention Trap:** Transformers ใช้การคำนวณแบบ Quadratic $O(L^2)$ ยิ่งข้อความยาวภาระการคำนวณยิ่งทวีคูณ
- **Memory Wall:** KV Cache กินหน่วยความจำมหาศาลเมื่อต้องประมวลผล Long Context
- **The Challenge:** เราต้องการความฉลาดระดับ GPT แต่เร็วและเบาเหมือน RNN ($O(L)$)



NotebookLM

Enter the State Space: ย้อนกลับสู่รากฐานเพื่อก้าวกระโดด



Continuous Systems

มองข้อมูลเป็นระบบต่อเนื่องทางคณิตศาสตร์ (Continuous-time) แล้วแปลงเป็น Digital (Discretization)

The Core Equation

$$h'(t) = Ah(t) + Bx(t)$$

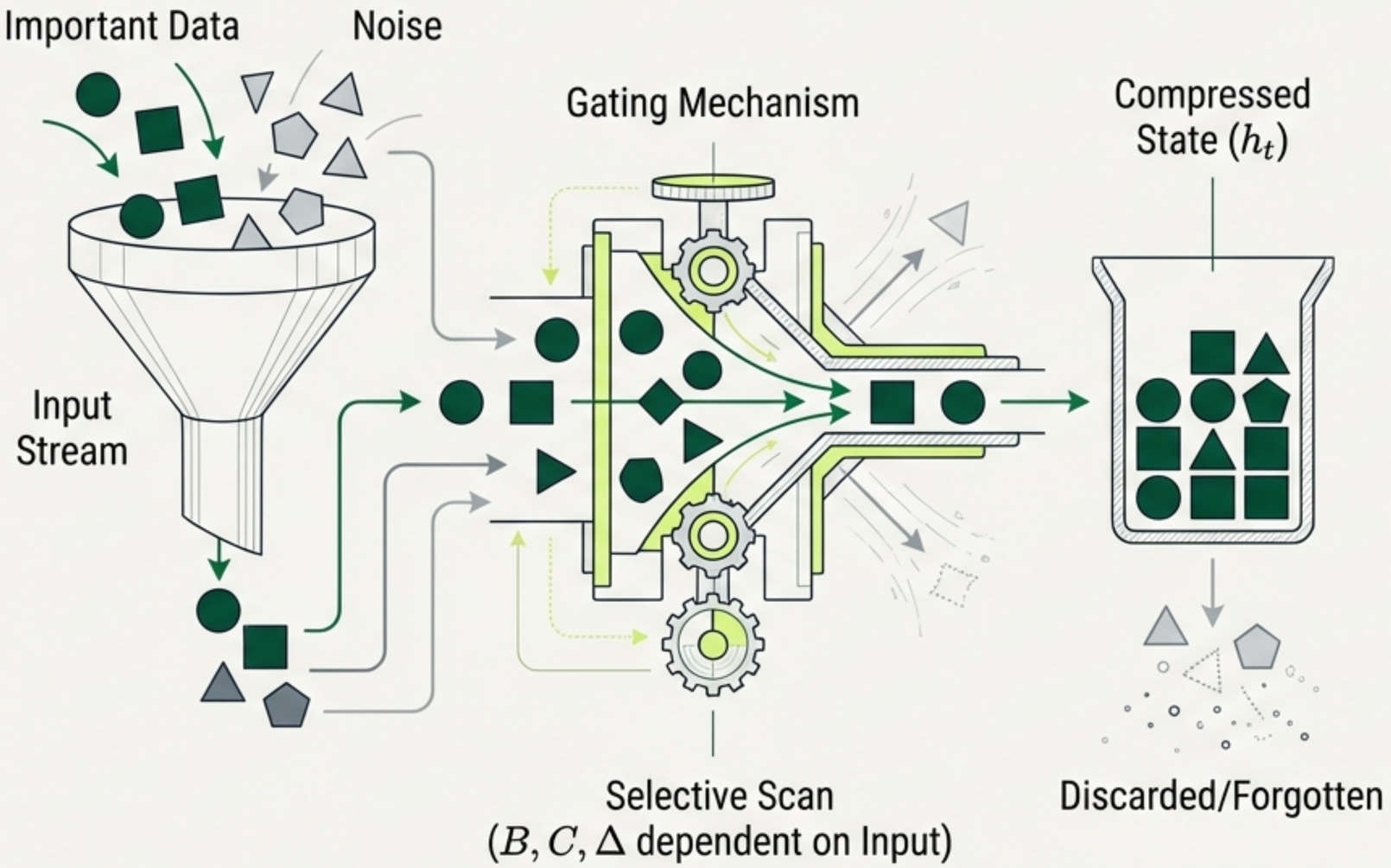
สมการหัวใจหลักในการส่งต่อสถานะ (State Transition)

Compressed Memory

ย่อประวัติข้อมูลทั้งหมดเก็บไว้ใน Hidden State ขนาดคงที่ ไม่ต้องอ่านข้อมูลเก่าซ้ำ

NotebookLM

The ‘Selection’ Magic: หัวใจสำคัญของความอัจฉริยะ



- **Problem with Old SSMs:** โมเดลรุ่นเก่าจำทุกอย่างเท่ากันหมด (Linear Time Invariant) ทำให้ข้อมูลขยะปนกับข้อมูลสำคัญ
- **Selective Mechanism:** Mamba เพิ่มกลไกให้โมเดล *เลือก* ได้ว่าจะ ‘จำ’ หรือ ‘ลืม’ ข้อมูลไหนตามบริบท
- **Dynamic Dynamics:** ค่าพารามิเตอร์เปลี่ยนแปลงตาม Input x_t ทำให้โฟกัสเฉพาะสิ่งที่ **จำเป็น** คล้ายกลไกของ LSTM แต่เร็วกว่า

NotebookLM

Mamba-2 & SSD: ทลายกำแพงด้วย Matrix Multiplication

Structured State Space Duality (SSD)

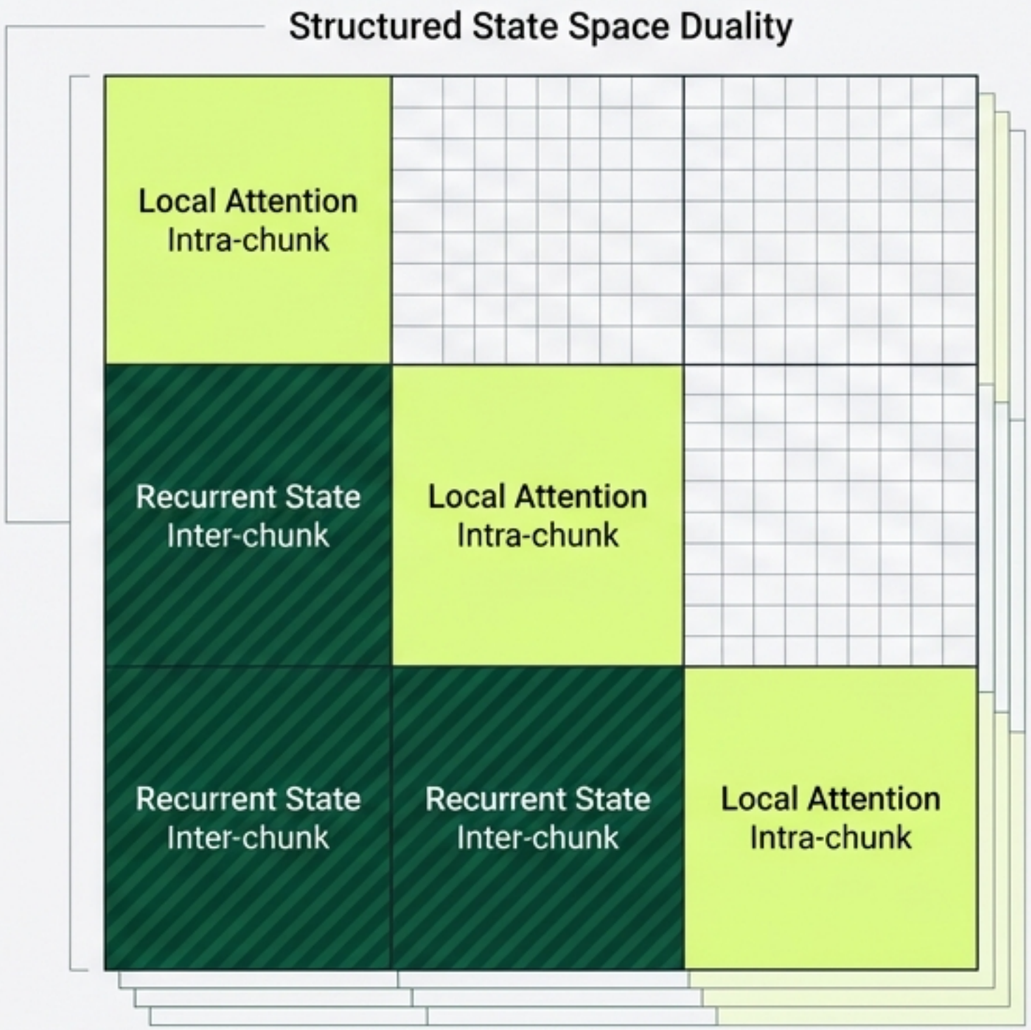
การค้นพบความเชื่อมโยงทางคณิตศาสตร์ระหว่าง SSMs และ Linear Attention

Matrix Transformation

เปลี่ยนการคำนวณแบบ Recurrent ให้กลายเป็นการคูณ Matrix แบบ Block Decomposition

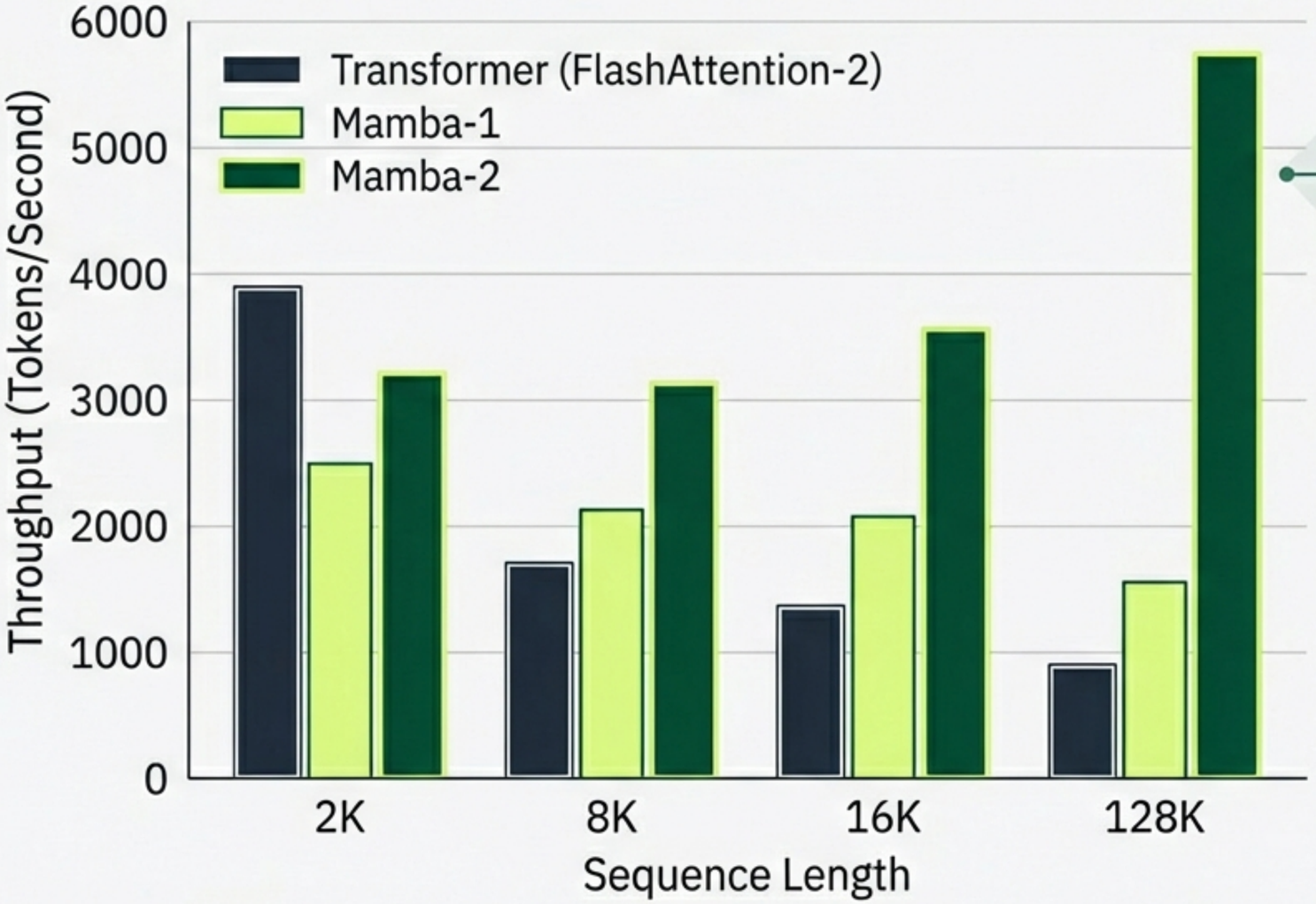
Hardware Efficiency

ออกแบบมาเพื่อ Tensor Cores บน GPU ยุคใหม่ Train เร็วขึ้น 2-8 เท่า



NotebookLM

Speed & Scale: เร็ว แรง ไร้ขีดจำกัด

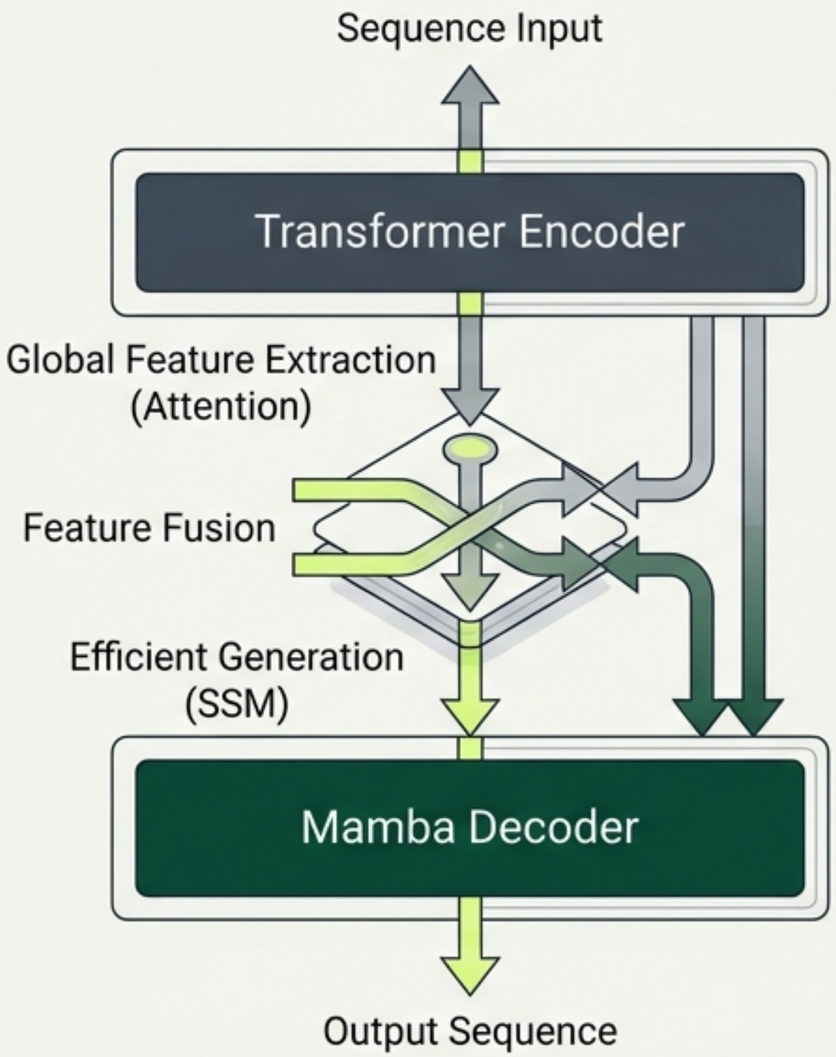


Massive Throughput
Mamba-2 เร็วกว่าและรองรับ Context ยาวกว่า โดยประสิทธิภาพไม่ตก

- **Linear Time $O(L)$:** เวลาประมวลผลเพิ่มขึ้นเป็นเส้นตรง ไม่ทวีคูณ
- **Constant Inference Cost:** ใช้เวลาคงที่ต่อ 1 คำ ($O(1)$) ไม่ว่าจะคุยมานานแค่ไหน
- **Long Context King:** รองรับ State ขนาดใหญ่ได้ดีเยี่ยม

NotebookLM

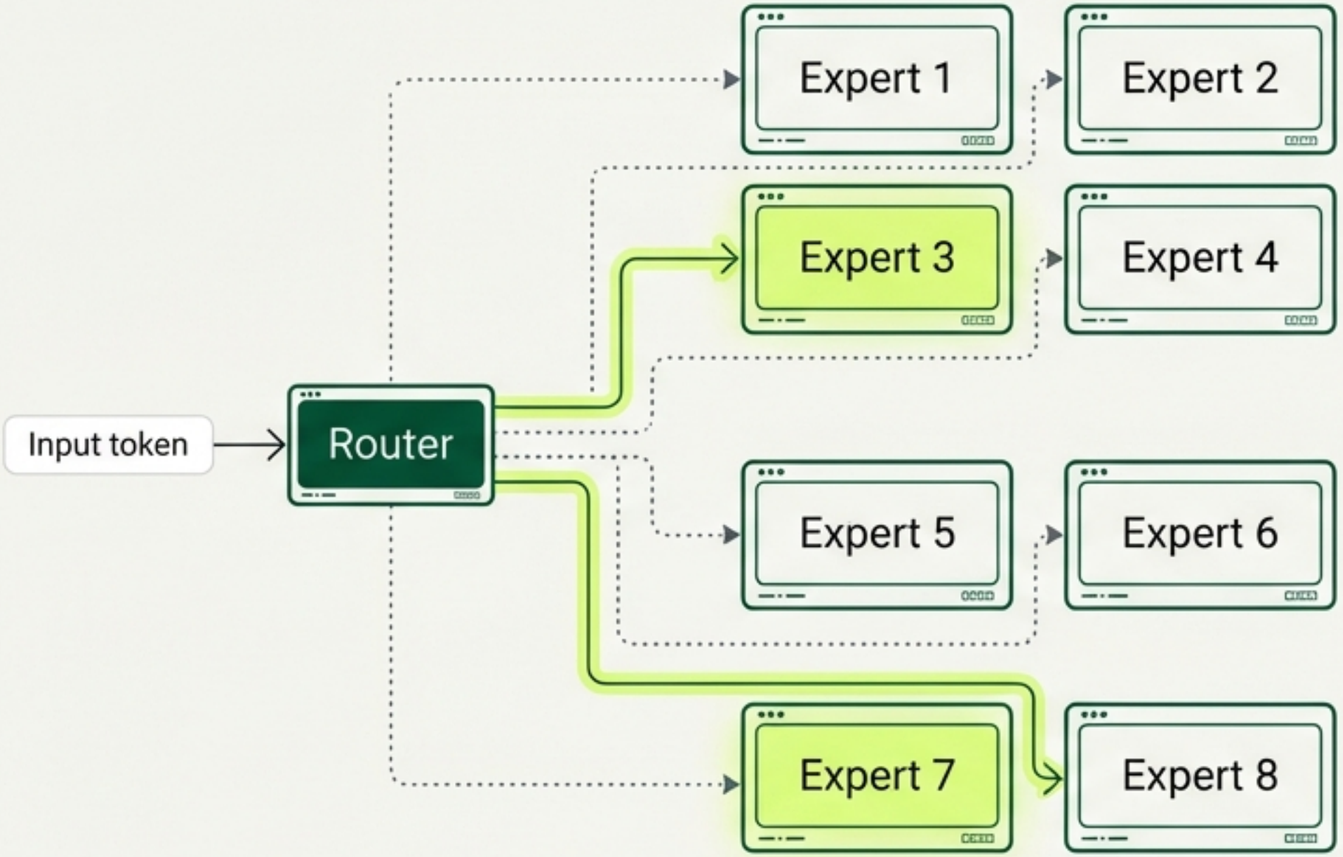
Hybrid Power: TransMamba และการผสานจุดแข็ง



- **The Concept:** ใช้ Transformer เป็น Encoder เพื่อเข้าใจภาพรวม และ Mamba เป็น Decoder เพื่อสร้างคำตอบที่รวดเร็ว
- **Feature Fusion:** เทคนิคการรวม Hidden States จาก Mamba เข้ากับ Features จาก Transformer
- **Result:** ได้ความฉลาดของ Transformer และความเร็วของ Mamba ในโมเดลเดียว

NotebookLM

Scaling with BlackMamba: Mixture of Experts (MoE)



Sparse Activation: Only selected experts are used per token.

MoE Architecture

แทนที่ MLP layers เดิมด้วย 'Experts' (ผู้เชี่ยวชาญเฉพาะด้าน) จำนวนมาก

Routing Mechanism

ระบบเลือกใช้งานเฉพาะ Expert ที่จำเป็นต่อ Token นั้นๆ (Active Parameters ต่ำ แม้ Total Parameters สูง)

Performance

BlackMamba ชนะ Transformer ในขณะที่ใช้ **Training FLOPs** น้อยกว่า

NotebookLM

Zamba2: ผู้นำโมเดลขนาดเล็ก (SLM) ยุคใหม่

Model	Size	MMLU Score	HellaSwag Score	Speed
Zamba2-1.2B	**1.2B**	**43.1**	**70.9**	**High**
Gemma2-2B	2B	32.8	71.4	Medium
Llama3.2-1.2B	1.2B	36.8	63.7	Medium

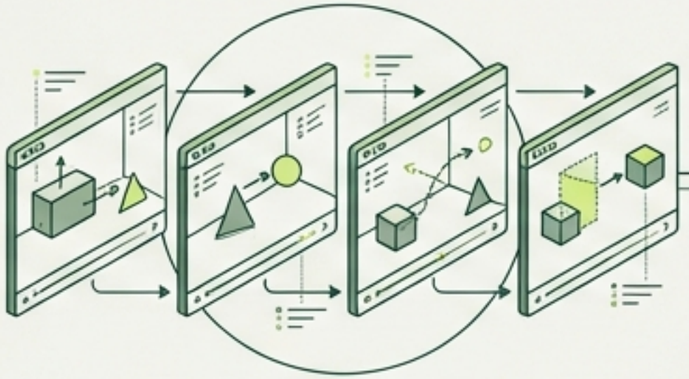
- **Hybrid Design:** ใช้ Mamba-2 Backbone ผสมกับ Shared Attention Blocks
- **State-of-the-Art:** ชนะโมเดลขนาดใหญ่กว่าอย่าง Gemma และ Llama ในหลาย Benchmark
- **Low Latency:** เหมาะสำหรับการใช้งาน On-device หรือ Edge Computing

NotebookLM

Beyond Text: ก้าวต่อไปสู่โลกแห่งข้อมูลหลากมิติ

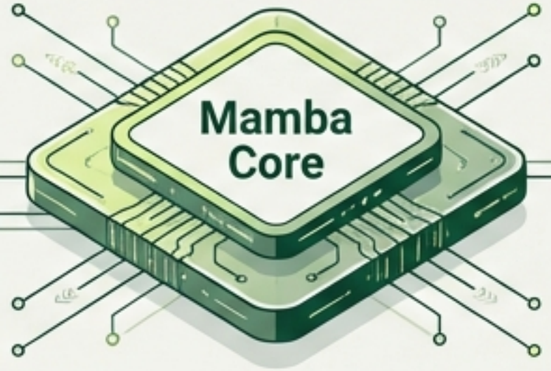
Genomics

วิเคราะห์รหัสพันธุกรรมยาวนับล้านหน่วย
(Long-sequence DNA)



Video/Vision

Vision Mamba: เข้าใจภาพเคลื่อนไหวและวิดีโอ



Audio & Signal

ประมวลผลสัญญาณเสียงต่อเนื่องความละเอียดสูง

Universal AI

Mamba ไม่ใช่แค่ Chatbot
แต่คือรากฐานใหม่ของ AI ที่เข้าใจ
Pattern ของโลกผ่านข้อมูลทุกรูปแบบ

NotebookLM

The Future is Linear: บทสรุปแห่งประสิทธิภาพ

Transformer Era ($O(L^2)$) → **Mamba Era** ($O(L)$) → ∞

Efficiency

ปลดล็อกคอขวดการคำนวณด้วย Linear Complexity

Capability

รองรับ Infinite Context และ Multimodal Data อย่างแท้จริง

Vision

กุญแจสำคัญสู่ยุคถัดไปของ AI ที่ทรงพลังและเข้าถึงได้

NotebookLM