

Wprowadzenie: Ukryty Związek Między Transformerami a SSM

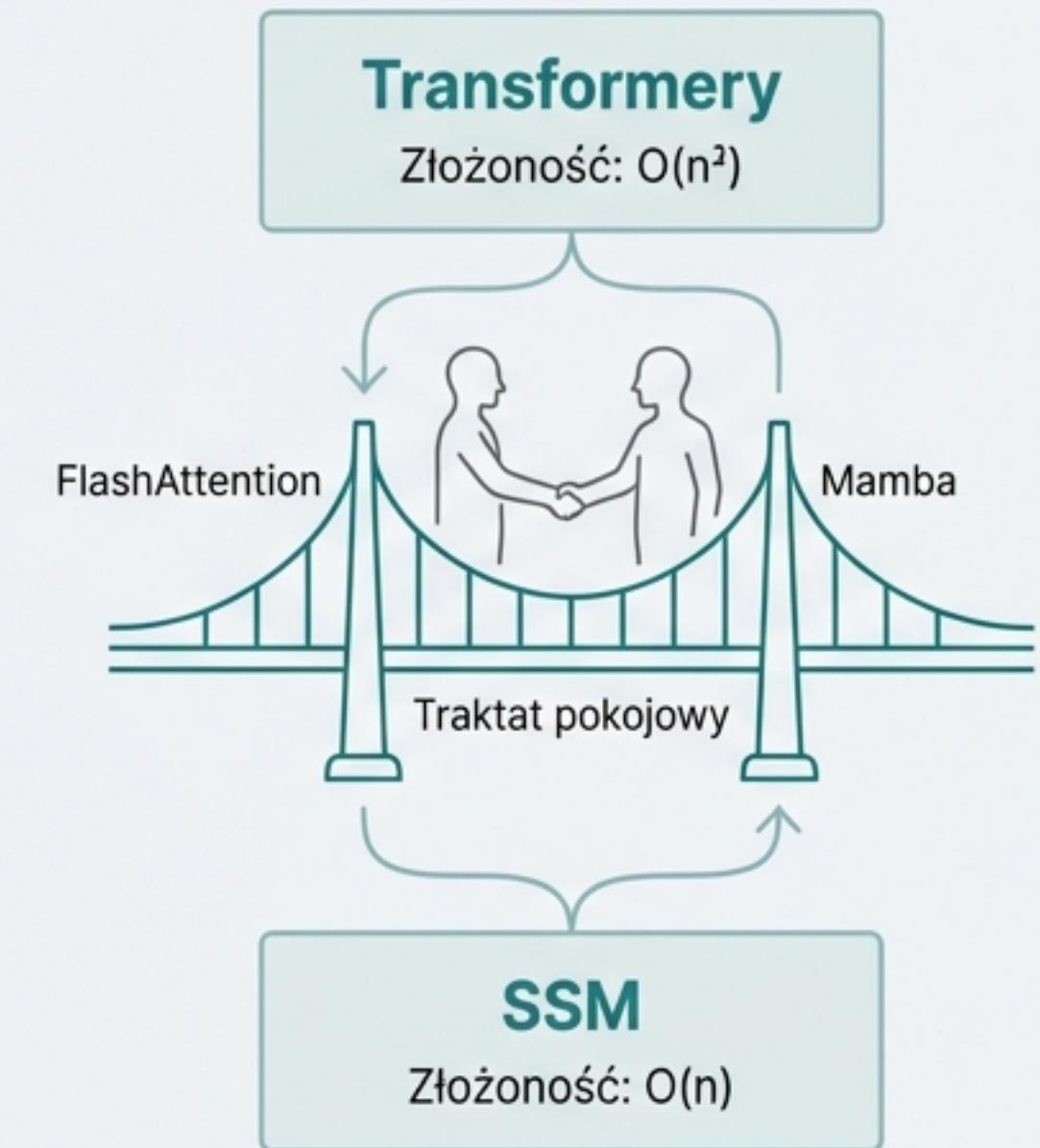
Transformery dominują: Architektury oparte na mechanizmie uwagi (attention) osiągają wyniki state-of-the-art, ale ich koszt obliczeniowy rośnie kwadratowo ($O(n^2)$) wraz z długością sekwencji.

SSM jako alternatywa: Modele Przestrzeni Stanów przetwarzają sekwencje w sposób liniowy ($O(n)$), oferując znacznie lepszą skalowalność.

"Traktat pokojowy": Publikacja udowadnia, że te dwie rodziny modeli nie są rywalami, a matematycznymi krewnymi. To fundamentalne odkrycie jednoczy dwa odrębne dotąd obszary badań.

Autorzy-eksperci: Tri Dao (twórca FlashAttention) i Albert Gu (architekt Mamba) – autorytety w obu dziedzinach, co daje im unikalną perspektywę do zbudowania tego mostu.

Kluczowa intuicja: Podwojenie długości kontekstu w Transformerach oznacza 4-krotny wzrost kosztów. SSM obiecują skalowanie liniowe, co otwiera drogę do przetwarzania znacznie dłuższych sekwencji.

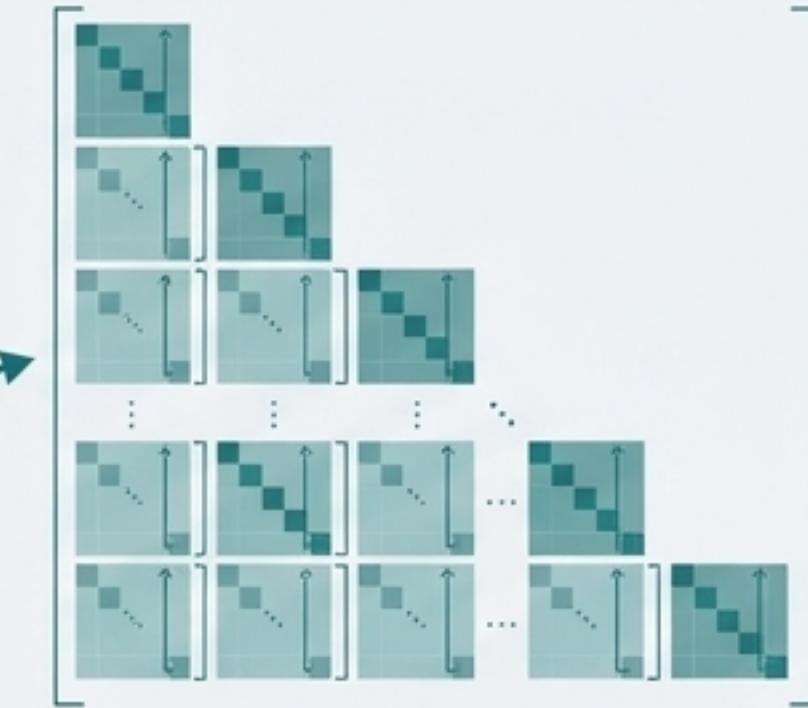


Dualność Przestrzeni Stanów (SSD): Kluczowe Odkrycie

SSM
(forma rekurencyjna)



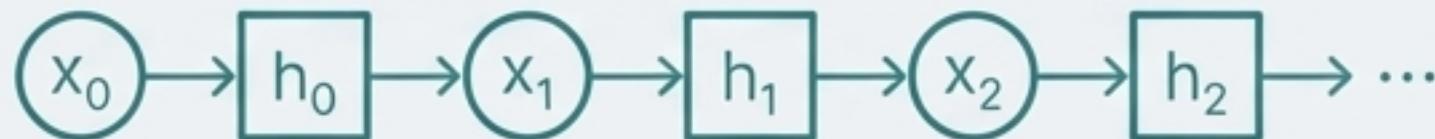
SSM
(forma macierzowa)
 $\mathbf{Y} = \mathbf{M} \times \mathbf{X}$



- Każdy Model Przestrzeni Stanów (SSM) można wyrazić jako prostą operację mnożenia macierzy:
 $\mathbf{Y} = \mathbf{M} \times \mathbf{X}$.
- Sekret tkwi w macierzy **M**. Nie jest ona losowa – posiada specjalną strukturę wewnętrzną zwaną **macierzą semi-separowalną**.
- Ta struktura oznacza, że w macierzy istnieją ukryte, powtarzalne wzorce, a nie chaotyczne wartości.
- Dzięki tej właściwości, wynik $\mathbf{Y} = \mathbf{MX}$ można obliczyć na dwa fundamentalnie różne sposoby.
- **Dualność** oznacza, że ten sam SSM może przyjąć dwie różne "maski obliczeniowe" – jedną szybką i rekurencyjną, a drugą przypominającą mechanizm uwagi.

Dwie Formy Obliczeniowe: Liniowa kontra Kwadratowa

Forma Liniowa (rekurencyjna)



Forma Kwadratowa (naiwna)

$$\mathbf{M} \times \mathbf{X} = \mathbf{Y}$$

Standardowy, szybki sposób obliczania SSM.
Złożoność liniowa $\mathcal{O}(n)$, przetwarzanie 'token po tokenie'.

Jawne zbudowanie całej macierzy \mathbf{M} i
przemnożenie jej przez \mathbf{X} .
Złożoność kwadratowa $\mathcal{O}(n^2)$, wolniejsza,
ale odsłaniająca głębszą prawdę.

Przełomowa obserwacja: Forma kwadratowa SSM wygląda niemal identycznie jak mechanizm uwagi (**attention**). Debata "SSM kontra Transformery" opierała się na fałszywej dydaktyce.

Ustrukturyzowana Uwaga Maskowana (SMA): Druga Strona Medalu

- **Początek:** Liniowa Uwaga (Linear Attention) unikała złożoności $O(n^2)$ poprzez zmianę kolejności mnożenia.
- **Uogólnienie autorów:** **Ustrukturyzowana Uwaga Maskowana (SMA).**
- **Kluczowa idea:** Zamiast prostej, trójkątnej maski, SMA używa ustrukturyzowanej macierzy \mathbf{L} (macierzy semi-separowalnej).



Zaskakujący rezultat: Jeśli jako \mathbf{L} użyjemy macierzy 1-separowalnej, cały mechanizm uwagi staje się szczególnym przypadkiem SSM. To zamyka pętlę dualności.

Algorytm SSD: Inżynieryjny Przełom

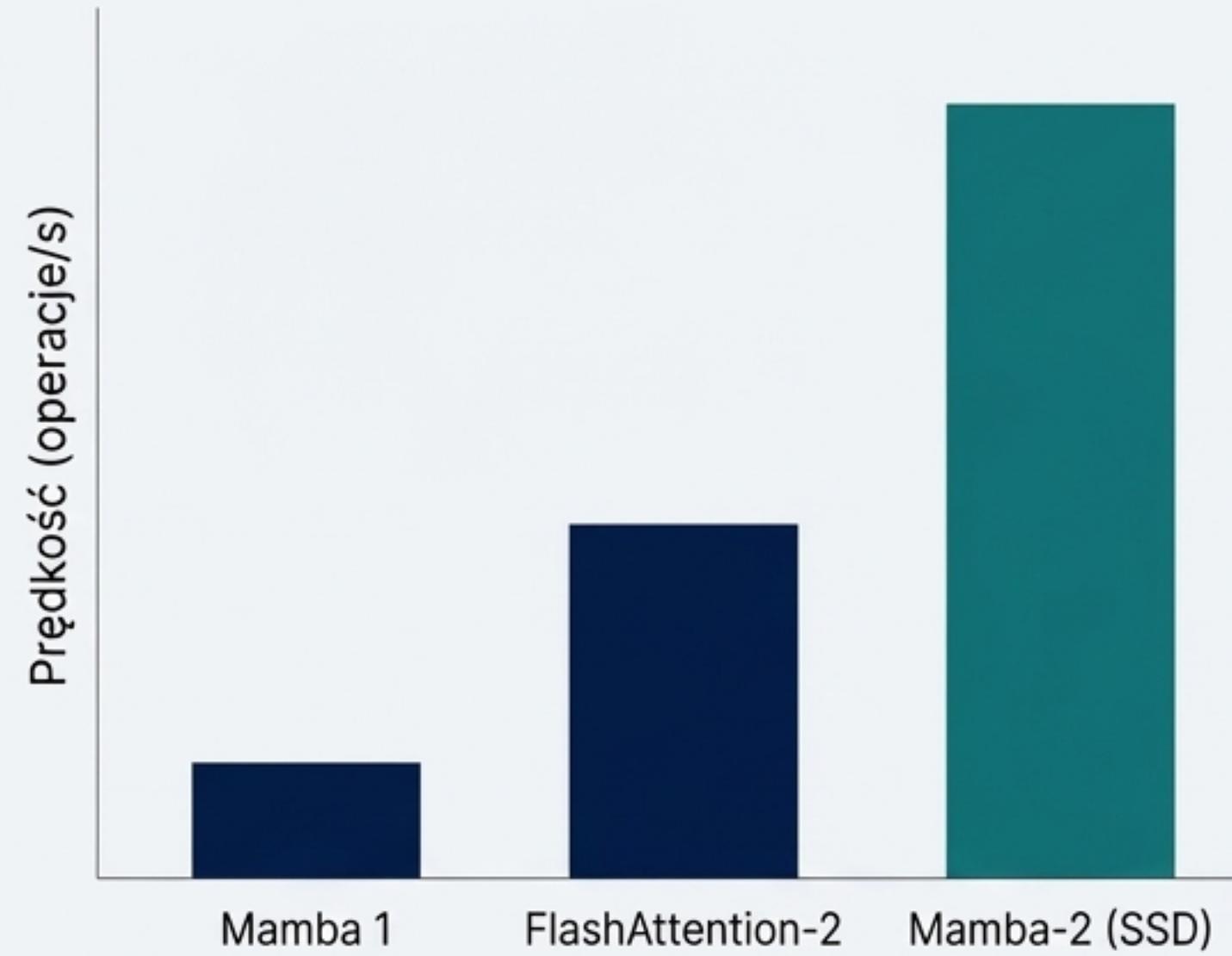
- Algorytm SSD to serce Mamba-2 – nowa metoda obliczeniowa, która wykorzystuje dualność.
- **Cel:** Połączyć wydajność sprzętową uwagi z liniowym skalowaniem SSM.
- **Podejście 'Chunking':** Sekwencja wejściowa jest dzielona na małe bloki (chunki).
- **Kluczowa korzyść:** Koszt kwadratowy $O(n^2)$ jest ograniczony do małych okien, dzięki czemu ogólna złożoność pozostaje liniowa.



Wzrost Wydajności: Liczby Mówią Same za Siebie

2-8x szybszy niż Mamba 1 **Szybszy niż FlashAttention-2** dla sekwencji >2K tokenów

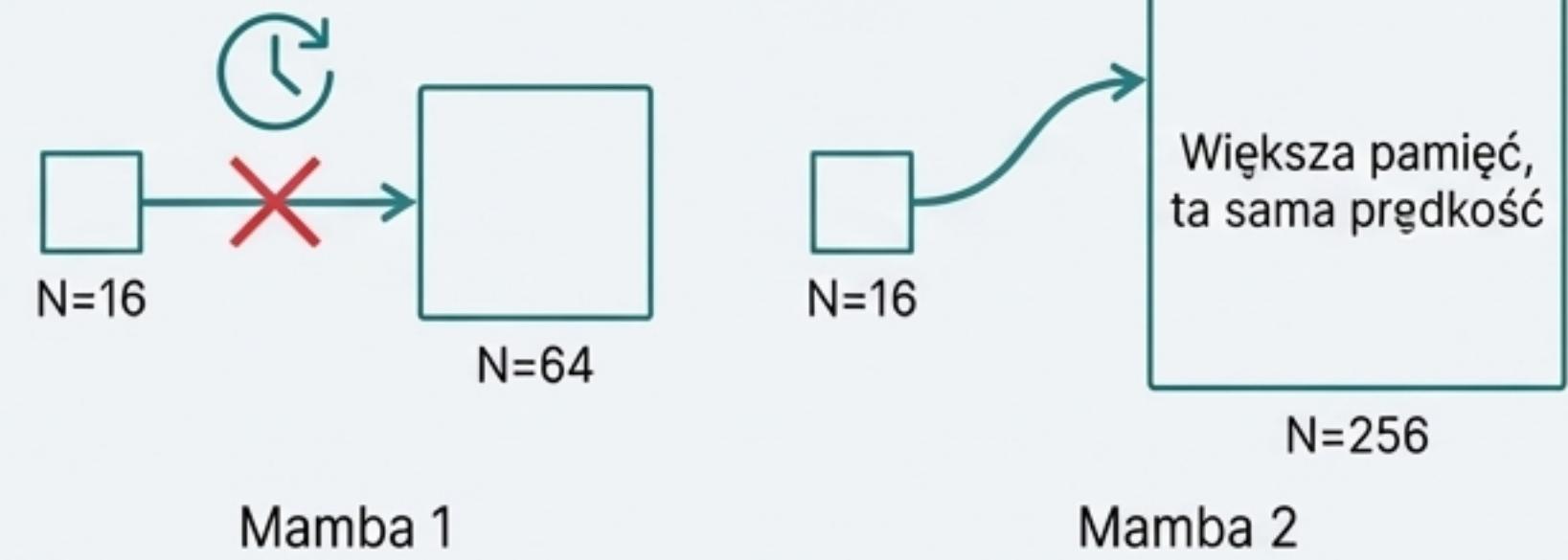
Prędkość przetwarzania (długość sekwencji 16K)



Przełom w skalowaniu stanu

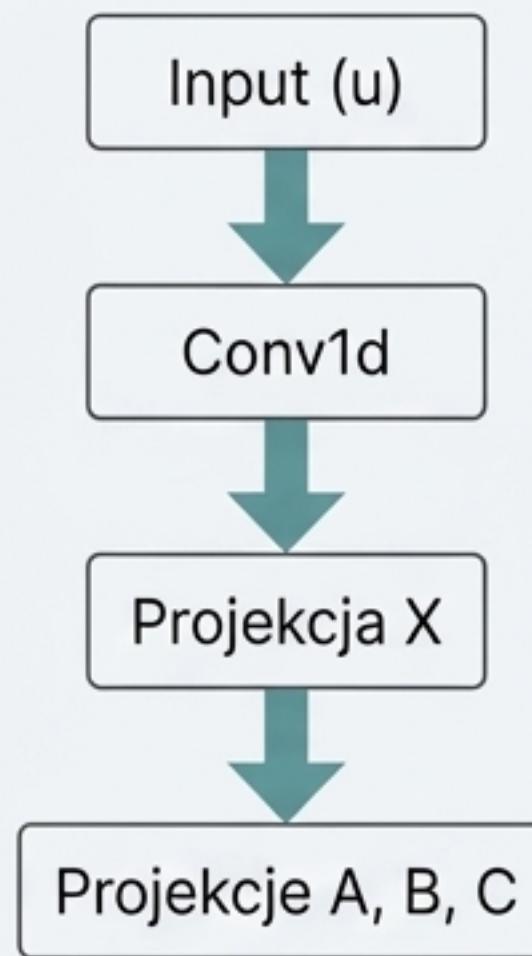
Przełamanie bariery "Pamięci Roboczej"

Mamba-2 pozwala na użycie znacznie większego
Współczynnika Ekspansji Stanu (N) bez spowolnienia.



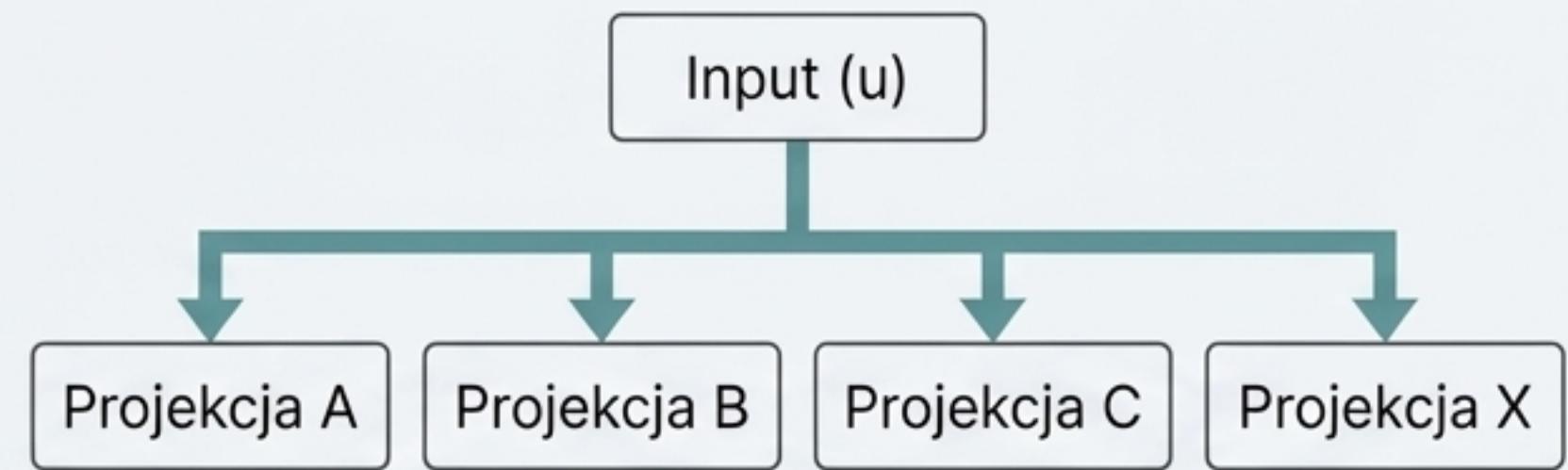
Innowacje Architektoniczne z Świata Transformerów

Mamba 1: Zależności Sekwencyjne

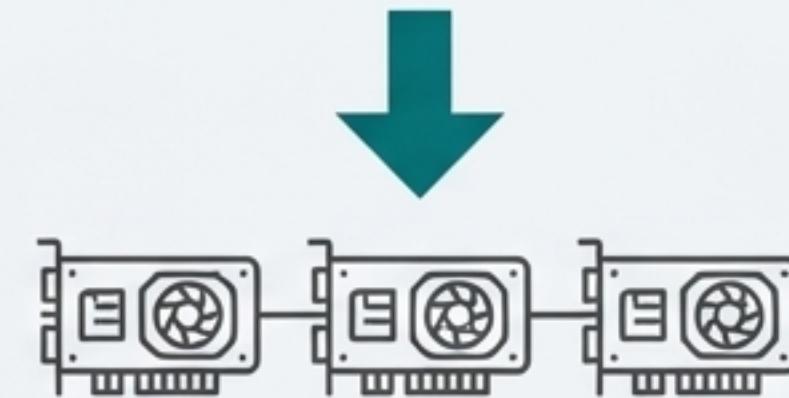


Zależności sekwencyjne utrudniały efektywną paralelizację.

Mamba 2: Równoległe Projekcje



Kluczowa korzyść: Równoległe projekcje umożliwiają efektywną Równoległość **Tensorową** (Tensor Parallelism).



Stabilność treningu: Dodanie warstwy **Group Norm** na końcu bloku stabilizuje proces uczenia w większych modelach.

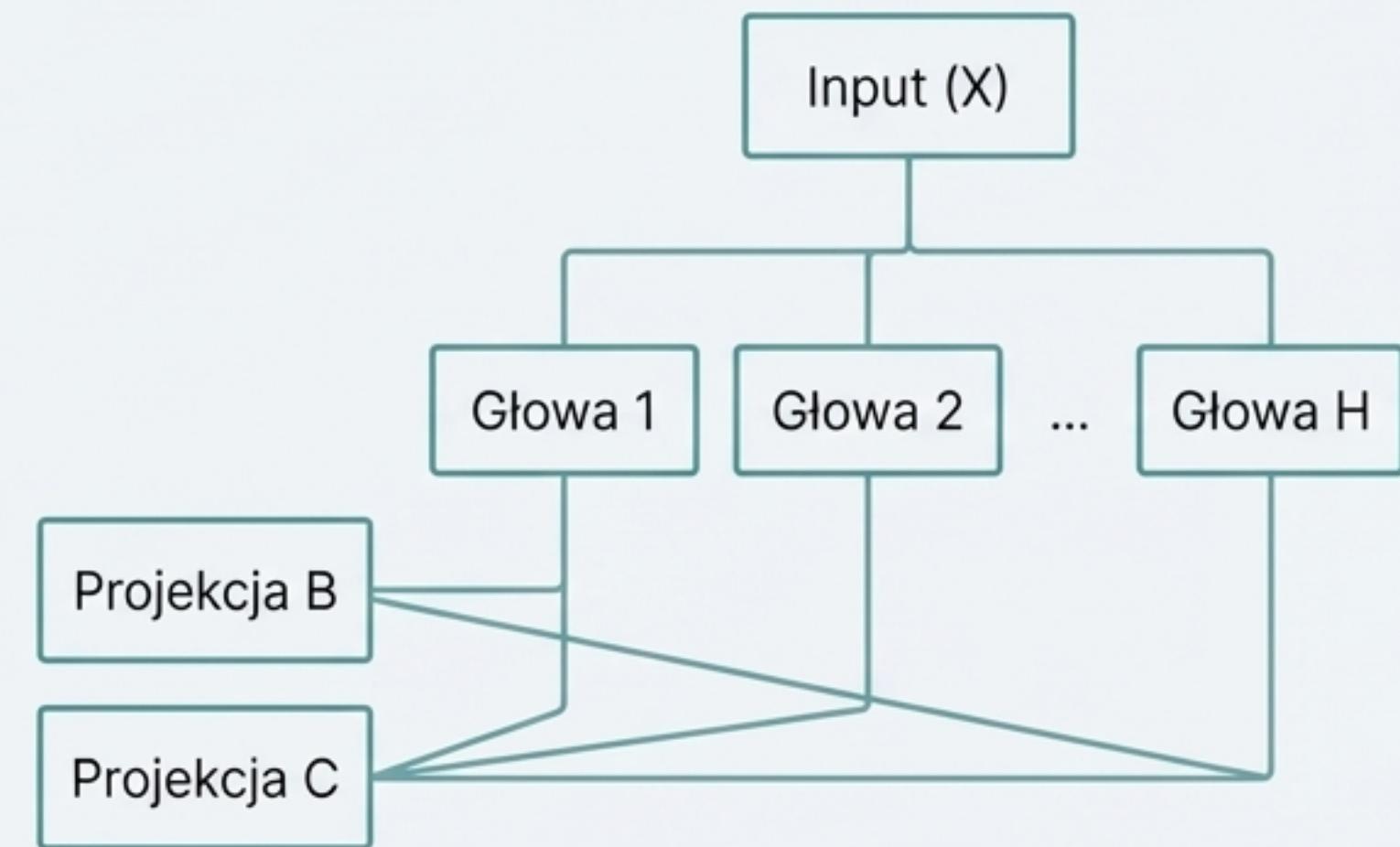
Wzorce Wielogłowicowe w Świecie SSM

Koncepcja 'multi-head attention' została przeniesiona do domeny SSM.

Mamba-2 wykorzystuje wzorzec analogiczny do **Uwagi Wielowartościowej (Multi-Value Attention, MVA)**, w pracy określany jako **Multi-input SSM (MIS)**.

Eksperymenty wykazały, że wzorce typu MVA dają najlepszą wydajność w kontekście SSM.

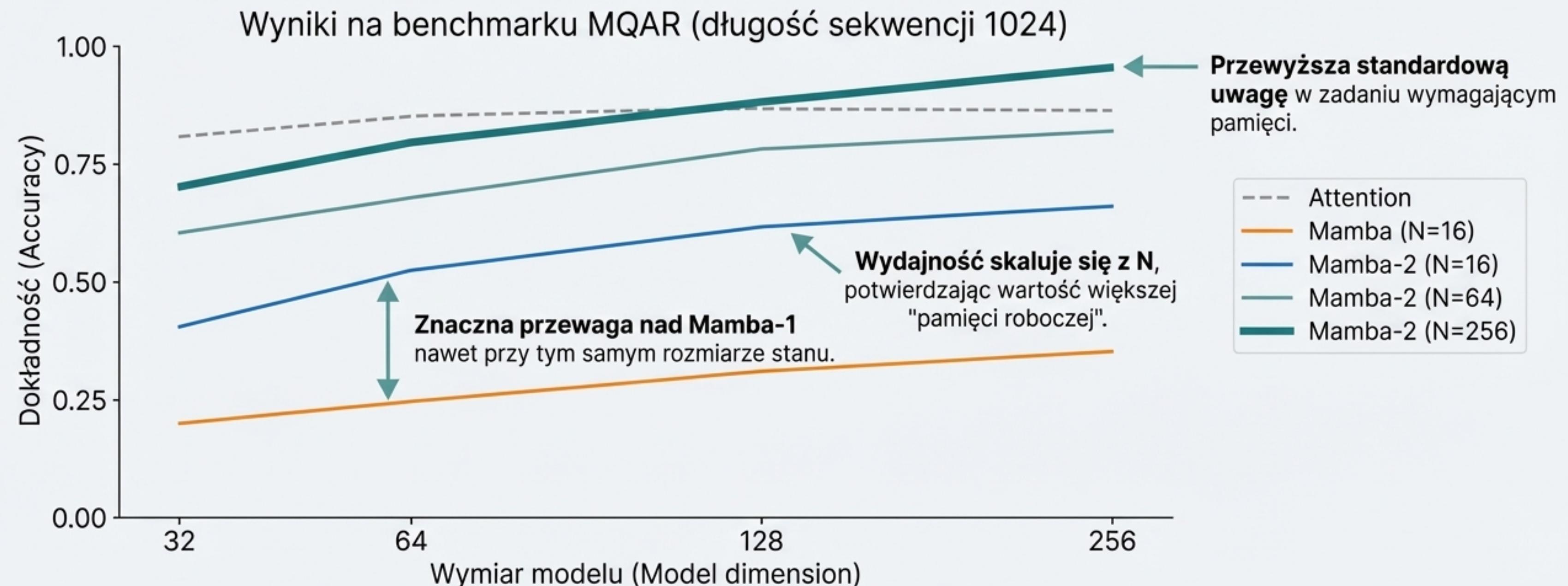
Synteza: Architektura Mamba-2 odzwierciedla wzorce projektowe Transformerów, zachowując jednocześnie korzyści wydajnościowe SSM.



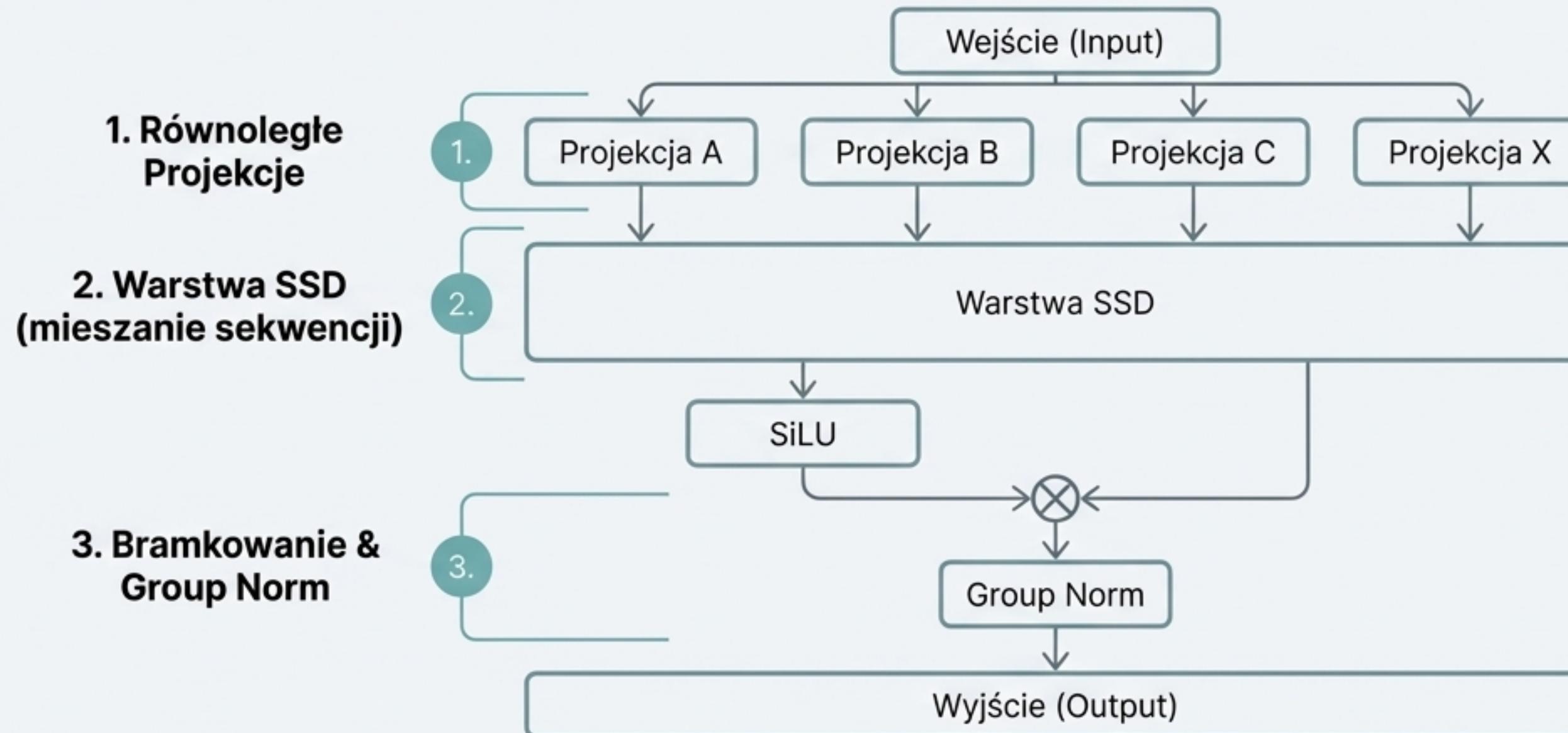
H głów dla wejścia (X), ale tylko 1 głowa dla parametrów stanu (B, C).

Wyniki Benchmarków: Dowód w Praktyce

Benchmark: MQAR (Multi-Query Associative Recall) - "brutalny test pamięci" polegający na odzyskiwaniu par klucz-wartość z długiego kontekstu. Jest to znana słabość tradycyjnych modeli rekurencyjnych.

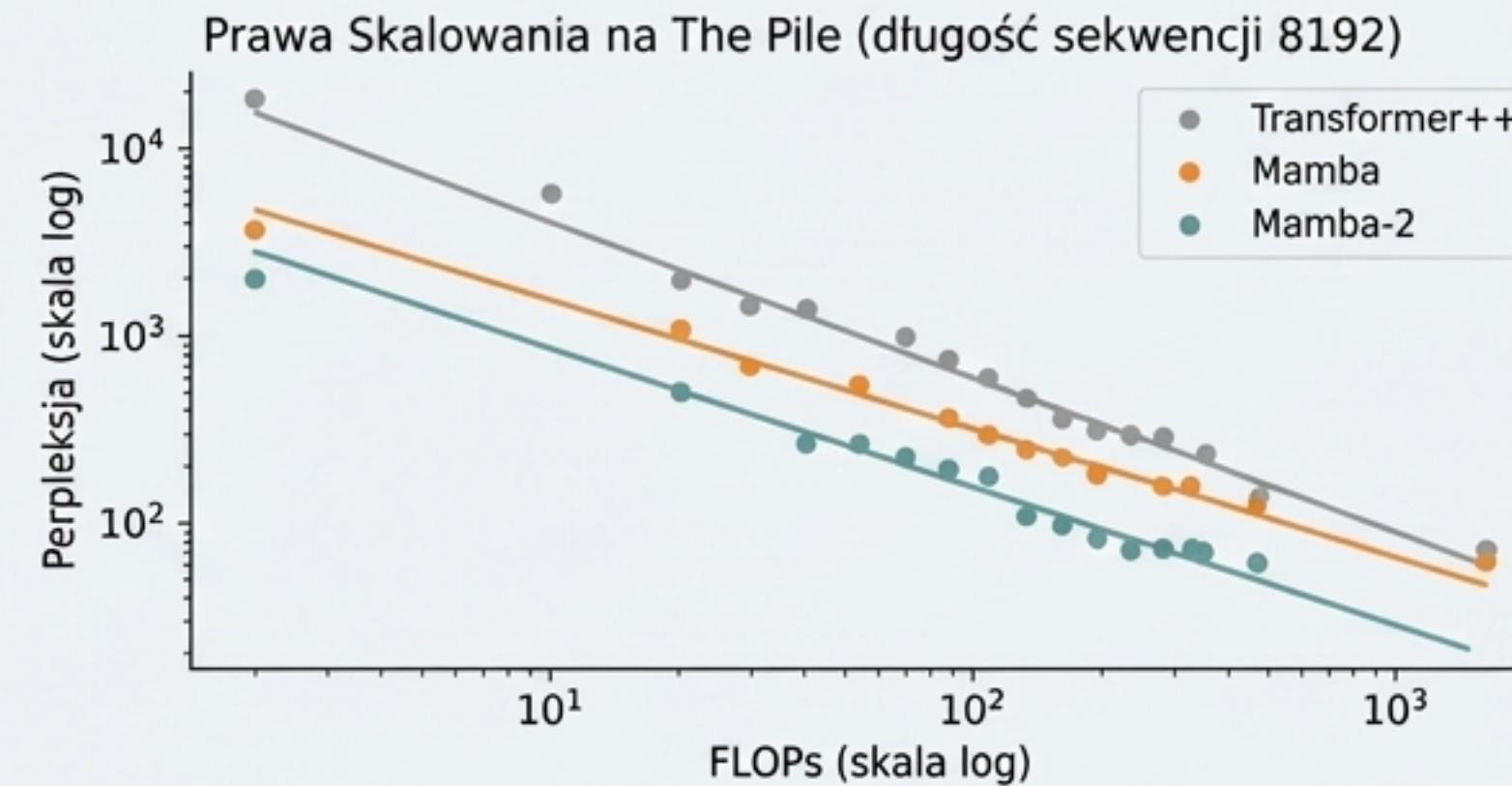


Architektura Mamba-2: Synteza Dwóch Światów



To nie jest przypadkowy zbiór komponentów, ale przemyślana synteza zasad projektowych ze światów SSM i Transformerów, zoptymalizowana pod kątem wydajności i skalowalności.

Nowa Era Skalowalności i Perspektywy



- **Lepsza wydajność:** Mamba-2 jest Pareto-optymalna – osiąga niższy błąd (perpleksję) przy tym samym budżecie obliczeniowym w porównaniu do Mamba 1 i silnego wariantu Transformer++.
- **Nowe ramy pojęciowe:** Dualność Przestrzeni Stanów (SSD) to coś więcej niż tylko jeden model. To nowe narzędzie do rozumienia i projektowania przyszłych modeli sekwencyjnych, jednoczące światy SSM i uwagi.

**Jedna struktura, dwa spojrzenia.
Zunifikowana przyszłość modeli sekwencyjnych.**