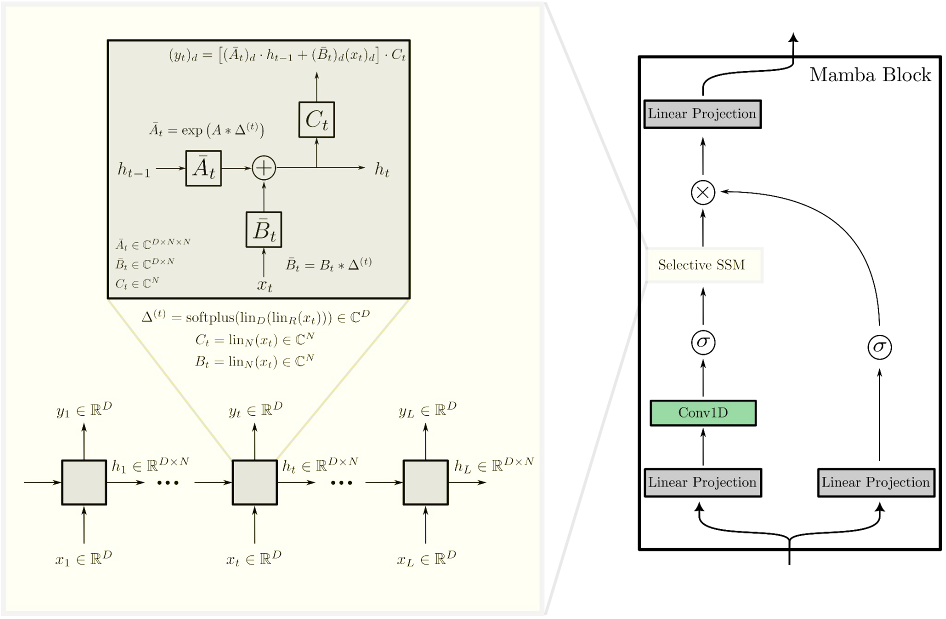

Mamba vs. Transformer:
Long Range Arena

Mamba Architecture: SSM with Selection Mechanism [1]



- SSM with **selection mechanism** (i.e. matrices A, B and C become input dependent)

$$h_t = \mathbf{\bar{A}_t} h_{t-1} + \mathbf{\bar{B}_t} x_t \,,$$
$$y_t = \mathbf{C_t} h_t \,.$$

  ⇒ added predictive power.

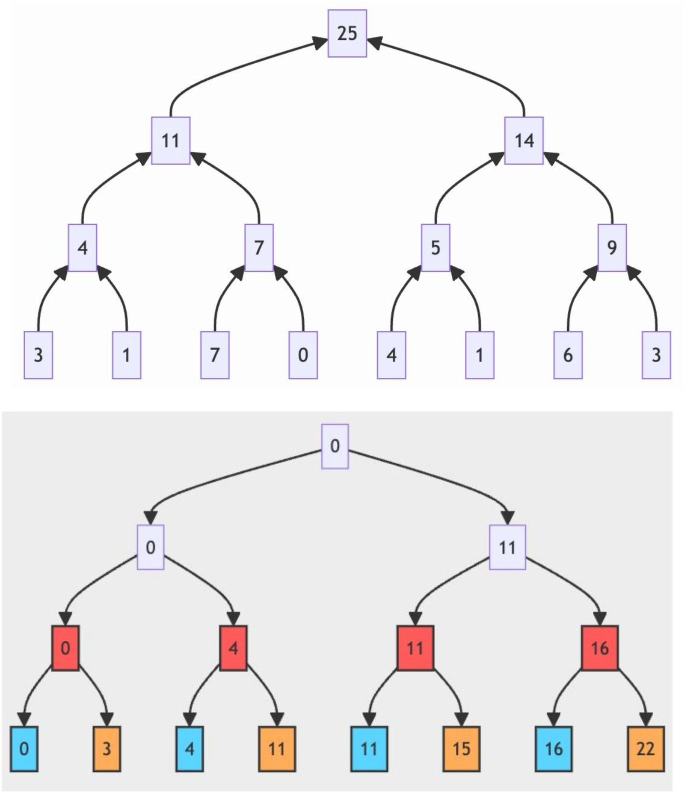
- **Selective SSM** similar to RNN: Achieves **linear time complexity** over input length (similar to an RNN), however **parallelizable**.

- Early results show outstanding performance in NLP, vision, etc. [1].

- **Thm** [2]**:** *One channel of the Mamba layer can express all functions that a single transformer head can express. Conversely, a single Transformer layer cannot express all functions that a single selective SSM layer can.*

[1] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
[2] Ameen Ali, Itamar Zimerman and Lior Wolf. The Hidden Attention of Mamba Models. *arXiv preprint arXiv:*2403.01590, 2024.

Mamba Architecture: Selective Scan



- Based on Blelloch Parallel Scan [1] → **Parallelize** computation of **prefix sum** → O(n/t) complexity for input of size n and t workers.

- Same idea can be applied to SSM's and Mamba [2,3] for computing $h_t = \bar{\mathbf{A}}_{\mathbf{t}} h_{t-1} + \bar{\mathbf{B}}_{\mathbf{t}} x_t \,,$

  via
$$\begin{aligned} h_1 &= \bar{\mathbf{B}}_{\mathbf{1}} x_1 \,, \\ h_2 &= \bar{\mathbf{A}}_{\mathbf{2}} \bar{\mathbf{B}}_{\mathbf{1}} x_1 + \bar{\mathbf{B}}_{\mathbf{2}} x_2 \,, \\ h_3 &= \bar{\mathbf{A}}_{\mathbf{3}} \bar{\mathbf{A}}_{\mathbf{2}} \bar{\mathbf{B}}_{\mathbf{1}} x_1 + \bar{\mathbf{A}}_{\mathbf{3}} \bar{\mathbf{B}}_{\mathbf{2}} x_2 + \bar{\mathbf{B}}_{\mathbf{3}} x_3 \,, \\ h_4 &= \bar{\mathbf{A}}_{\mathbf{4}} \bar{\mathbf{A}}_{\mathbf{3}} \bar{\mathbf{A}}_{\mathbf{2}} \bar{\mathbf{B}}_{\mathbf{1}} x_1 + \bar{\mathbf{A}}_{\mathbf{4}} \bar{\mathbf{A}}_{\mathbf{3}} \bar{\mathbf{B}}_{\mathbf{2}} x_2 + \bar{\mathbf{A}}_{\mathbf{4}} \bar{\mathbf{B}}_{\mathbf{3}} x_3 + \bar{\mathbf{B}}_{\mathbf{4}} x_4 \,, \\ &\ldots \end{aligned}$$

  ~ type of prefix "sum" and Parallel scan can be applied.

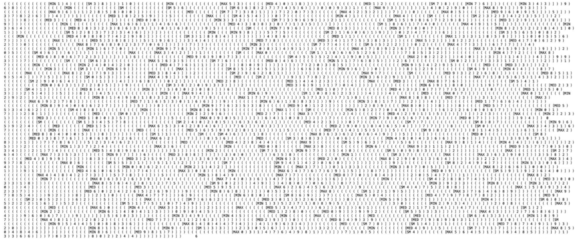
- Mamba further adds **hardware-aware** implementation akin to *FlashAttention* → significant speedups [3].

[1] Guy E Blelloch. Prefix sums and their applications. In *Sythesis of parallel algorithms*, pages 35—60. Morgan Kaufmann Publishers Inc., 1990
[2] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022
[3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Long Range Arena (LRA) Tasks [1]:

[MAX [MED [MED 1 [SM 3 1 3] 9] 6] 5]

Truth: 6; Pred: 5



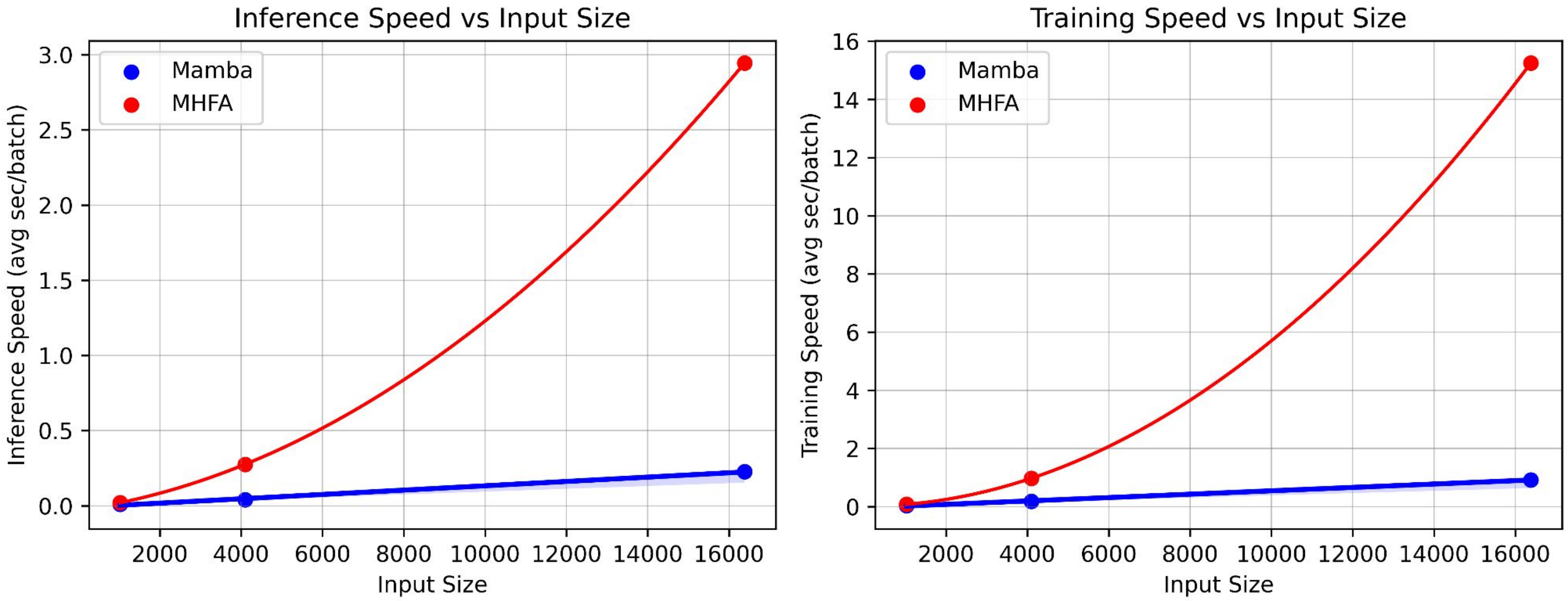
32x32

64x64

- **Long ListOps:** Up to 2k-length, 10-way classification task.
- **Text (IMDB):** Byte/Char-level text sentiment analysis task for the IMDB dataset; fixed length of 2k characters.
- **Retrieval (AAN):** Byte-level document retrieval; determine if two (scientific) papers are related or not. Up to 4k+4k char length.
- **Image (CIFAR-10):** Unravel 32x32 CIFAR images into a 1028k length list and classify it into 10 categories.
- **Pathfinder:** Unravel different resolution images 32x32 → 256x256 and determine whether 2 dots connected.

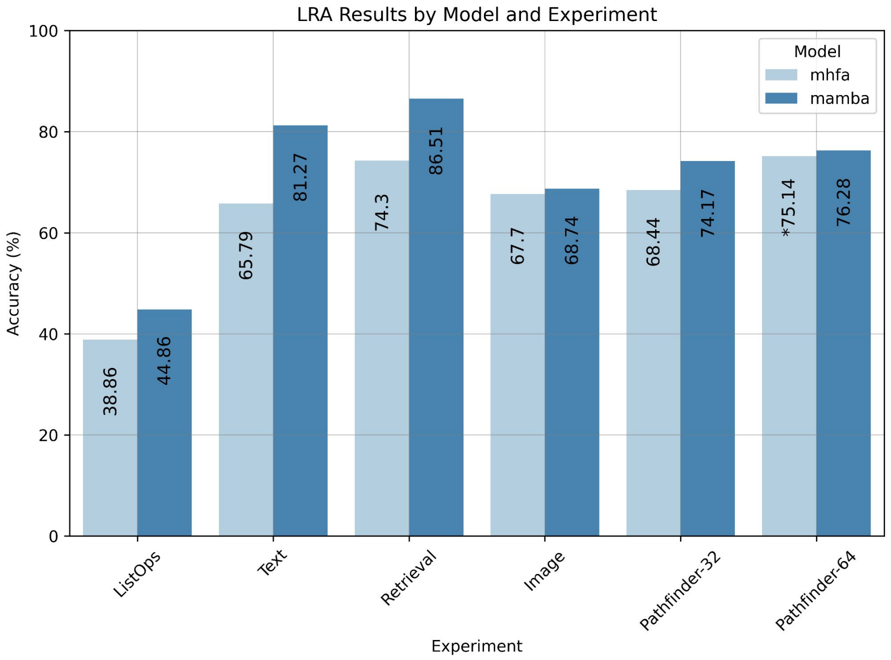
[1] Yi Tay, et al.. Long range arena: A benchmark for efficient transformers.

Speed Comparison: Mamba Vs MHFA



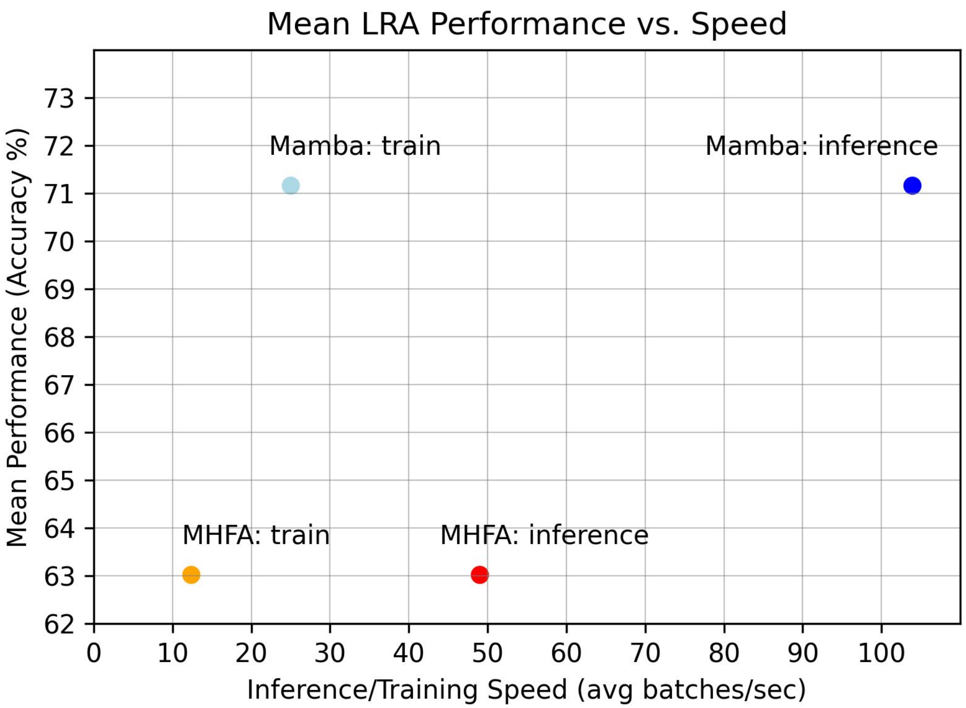
Both models contain ~600k parameters and we replace the MHFA part with a Mamba block. Values are obtained from the LRA pathfinder task with batch size = 32, ran on A100 in Google Colab.

Performance Analysis: Mamba vs MHFA + others



Model (input length)	ListOps 2000	Text 2048	Retrieval 4000	Image 1024	Pathfinder 1024	Pf-64 4096	PathX 16384	Avg*
Mamba	44.86	81.27	86.51	68.74	74.17	76.28	N/A	71.11
MHFA+RoPE	38.86	65.79	74.30	67.70	68.44	N/A	N/A	63.02
Transformer [24]	36.37	64.27	57.46	42.44	71.40	N/A	N/A	54.39
Performer [24]	18.01	65.40	53.82	42.77	77.05	N/A	N/A	51.41
RoPE [1]**	47.90	79.08	82.31	75.04	76.64	N/A	84.72	72.19
S4 [10]	59.60	86.82	90.90	88.65	94.20	N/A	96.35	84.03
Diagonal [12]	60.6	84.8	87.8	85.7	84.6	N/A	87.8	80.7
S5 [23]	62.15	89.31	91.40	88.00	95.33	N/A	98.58	85.24

LRA Performance Vs Speed



Conclusions

- Mamba is a **powerful** new architecture that is competitive (or superior) to MHFA (and its variants).
- Mamba has **faster** training/inference.
- Our results indicate significantly **better performance** than RoPE MHFA for comparable (or smaller) model sizes on LRA.
- Previous SSM’s (S4/S5/..) still reign supreme on LRA, but seem lacking in generalizing beyond benchmark tasks.
- We further experiment with modifications to Mamba (results in our report).