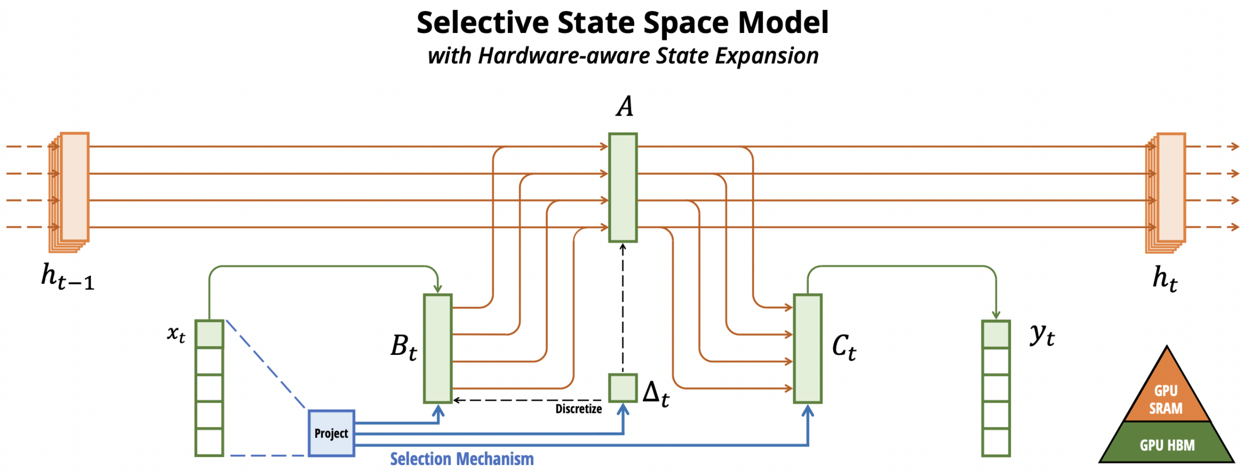
Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: ML-Engineer.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Mamba



$$\boxed{h'_{(t)} = A \cdot h_{(t)} + B \cdot x_{(t)} \qquad (1a)}$$

$$\boxed{y_{(t)} = C \cdot h_{(t)} \qquad (1b)}$$

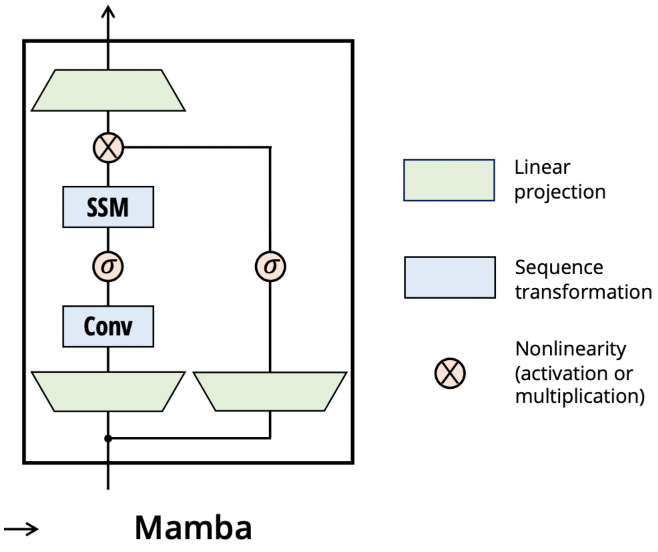
Using zero-order hold (ZOH) we get:

$$\boxed{h_{(t)} = \bar{A} \cdot h_{(t-1)} + \bar{B} \cdot x_{(t)} \qquad (2a)}$$

$$\boxed{y_{(t)} = C \cdot h_{(t)} \qquad (2b)}$$

$$\boxed{\bar{A} = \exp(\Delta A)}$$

$$\boxed{\bar{B} = (\Delta A)^{-1} \cdot [\exp(\Delta A) - I] \cdot \Delta B}$$



Main characteristics:

- Simply letting the SSM parameters be functions of the input addresses their weakness with discrete modalities, allowing the model to selectively propagate or forget information along the sequence length dimension depending on the current token.

- Mamba enjoys fast inference (5× higher throughput than Transformers) and linear scaling in sequence length.

- Mamba achieves state-of-the-art performance across several modalities such as language, audio, and genomics.