Tiled Flash Linear Attention: More Efficient Linear RNN and xLSTM Kernels

NeurIPS 2025

Maximilian Beck, ✉ beck@ml.jku.at 𝕏 maxmbeck 🌐 maxbeck.ai

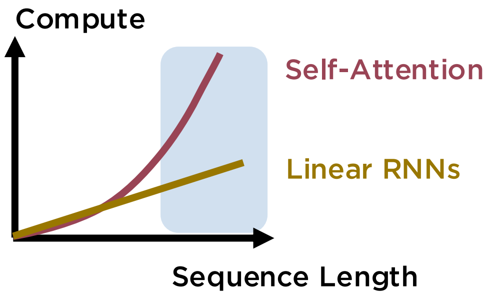
Korbinian Pöppel, Phillip Lippe, Sepp Hochreiter

November 2025

Motivation

- Recently, Linear RNNs with gating have become competitive to Transformers with softmax attention
  - Gated Linear Attention, Mamba, GatedDeltaNet
  - mLSTM (xLSTM with matrix memory)
- First frontier labs scale up (hybrid) attention alternatives, e.g. Qwen-Next or Kimi-Linear
- Two main drivers of this success:

Linear scaling in compute



AND

fast kernel implementations (Flash Linear Attention) with runtime benefits over Flash Attention

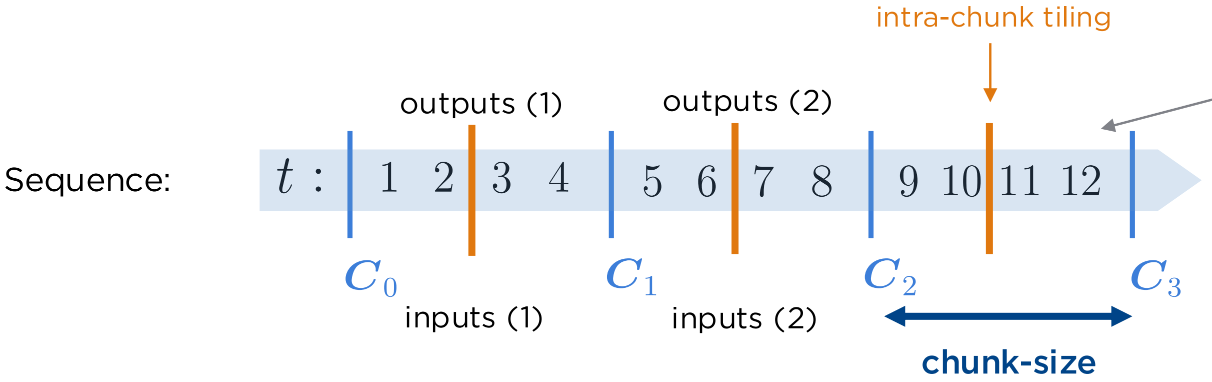
However, linear RNN kernels based on FLA often cannot fully utilize modern hardware, because they materialize too many memory states

➡ **Tiled Flash Linear Attention fully utilizes the GPU by introducing an additional parallelization dimension**

Yang, Songlin, et al. "Gated Linear Attention Transformers with Hardware-Efficient Training." *Forty-first International Conference on Machine Learning*.

Chunkwise-parallel computation & Limited chunk size

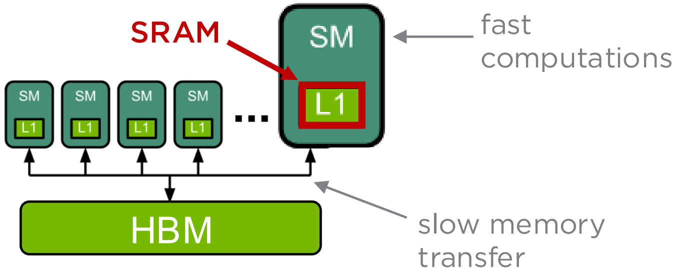
FLA kernels leverage a chunkwise-parallel formulation of linear RNNs:



compute scales quadratically with chunk size, but linear overall

chunk-size determines the number of stored and loaded memory states

Simplified model of the GPU



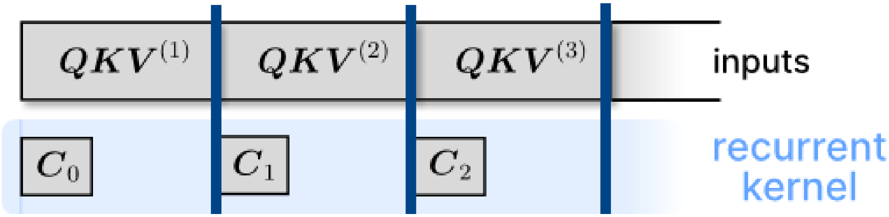
Problem: Chunk size in FLA kernels is limited by physical SRAM size.
(all inputs & outputs per chunk must fit in SRAM)

➡ **We need to load & store many memory states! → Slow!**

➡ **Solution: TFLA introduces an additional tiling dimension within the chunks!**
(only inputs & outputs per tile must fit in SRAM, use more tiles for larger chunk size)

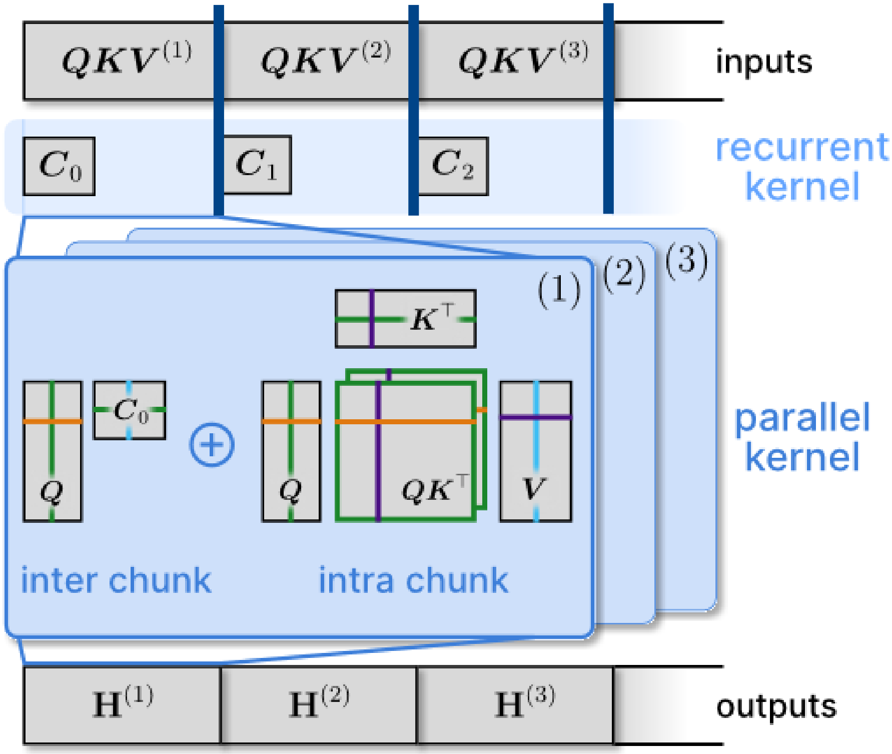
JKU | NXAI

Tiled Flash Linear Attention: Two levels of sequence parallelism



- **TFLA divides the sequence into chunks** and parallelizes over chunks **(1st level of sequence parallelism)**
- The recurrent kernel materializes the (first) memory state for each chunk

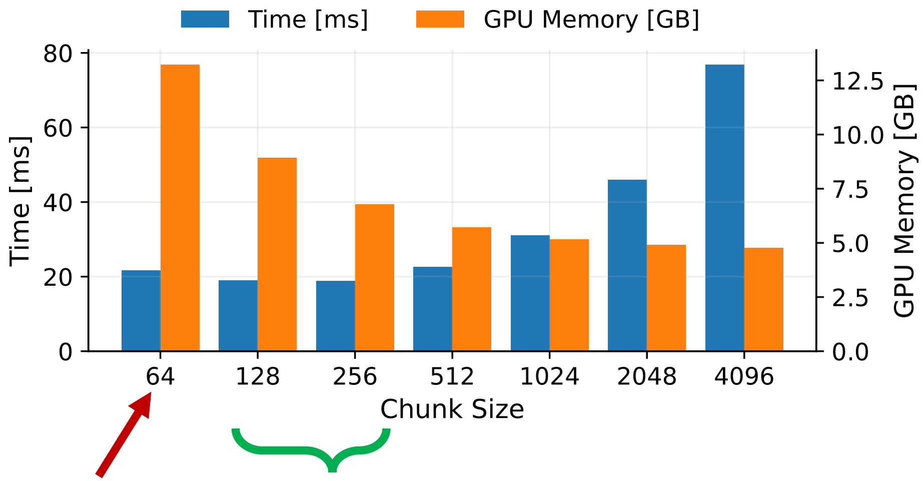
Tiled Flash Linear Attention: Two levels of sequence parallelism



- **TFLA divides the sequence into chunks** and parallelizes over chunks **(1st level of sequence parallelism)**
- The recurrent kernel materializes the (first) memory state for each chunk
- The parallel TFLA kernel **divides every chunk further** into smaller tiles
- **TFLA additionally parallelizes over the chunk tiles (2nd level of sequence parallelism)**

Results

Trade-off between memory & runtime

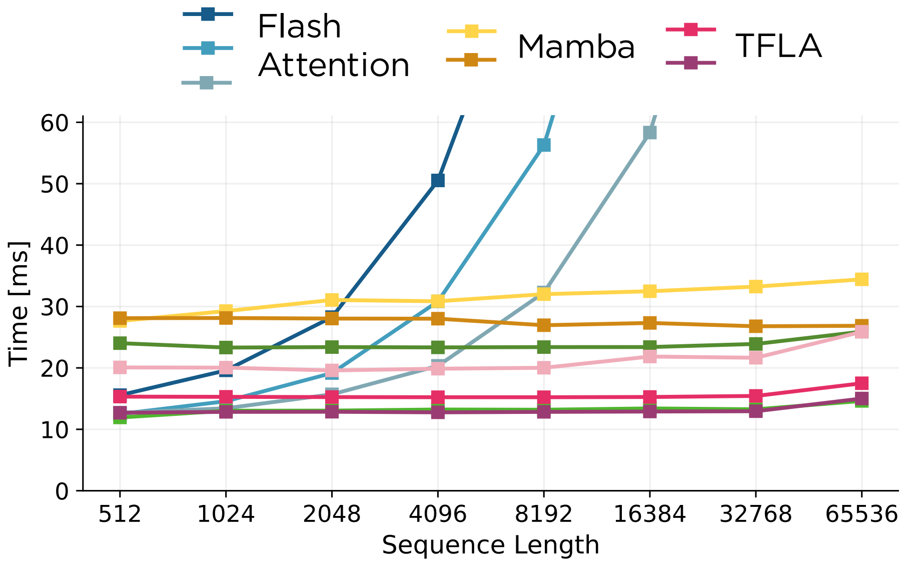


previous limit

runtime optimal

&

State of the art training kernel runtimes



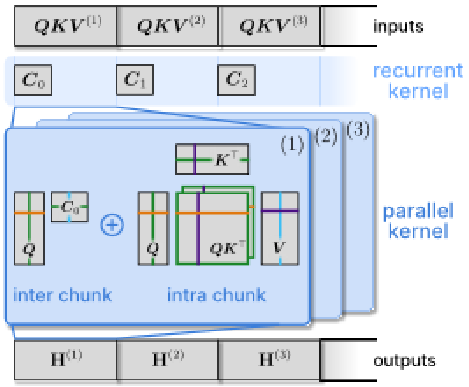
Training kernel runtime

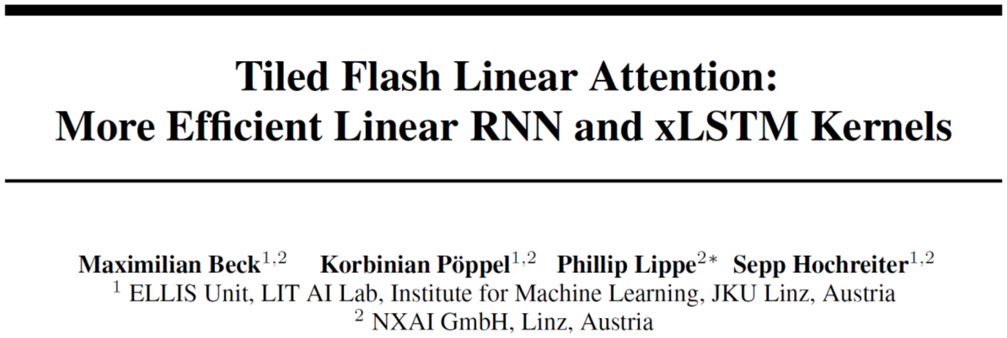
➡ **TFLA kernels are faster than FlashAttention & 2x faster than Mamba**

JKU | NXAI

Come and see us at our poster for more details!

- Application to xLSTM & faster mLSTM variant
- More kernel benchmarks
- Theoretical runtime and arithmetic intensity calculations





Tiled Flash Linear Attention:
More Efficient Linear RNN and xLSTM Kernels

Maximilian Beck[1,2] **Korbinian Pöppel**[1,2] **Phillip Lippe**[2*] **Sepp Hochreiter**[1,2]
[1] ELLIS Unit, LIT AI Lab, Institute for Machine Learning, JKU Linz, Austria
[2] NXAI GmbH, Linz, Austria

arxiv.org/abs/2503.14376

github.com/NX-AI/mlstm_kernels

- Recording link: https://recorder-v3.slideslive.com/?share=105189&s=0b7d7709-45dd-4507-9a44-2851afc0be33