## Paper

<https://arxiv.org/abs/2205.14135>

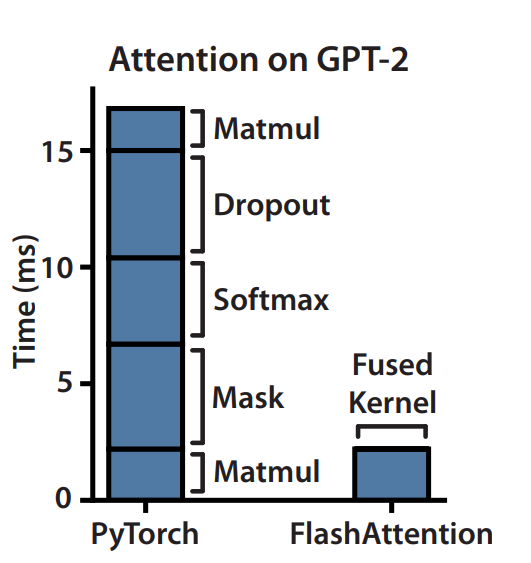
## Goal

* Attention is slow because of reads and writes
* Use tiling and recomputation to reduce IO (inputs/outputs)
* Application: transformers that can be faster and have longer context

## Background papers

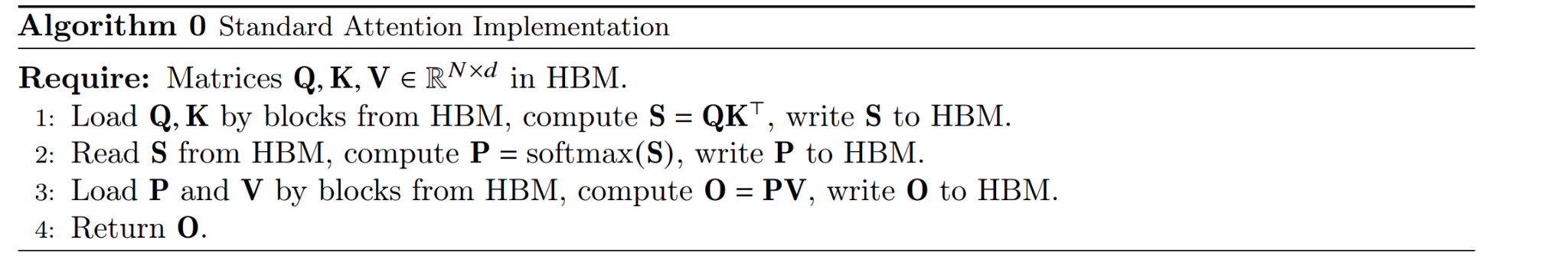
* Approximate attention
  + Sparse
  + Low rank
  + Trade quality for speed
  + Not widely adapted

## Tldr



* Attention takes way too much time

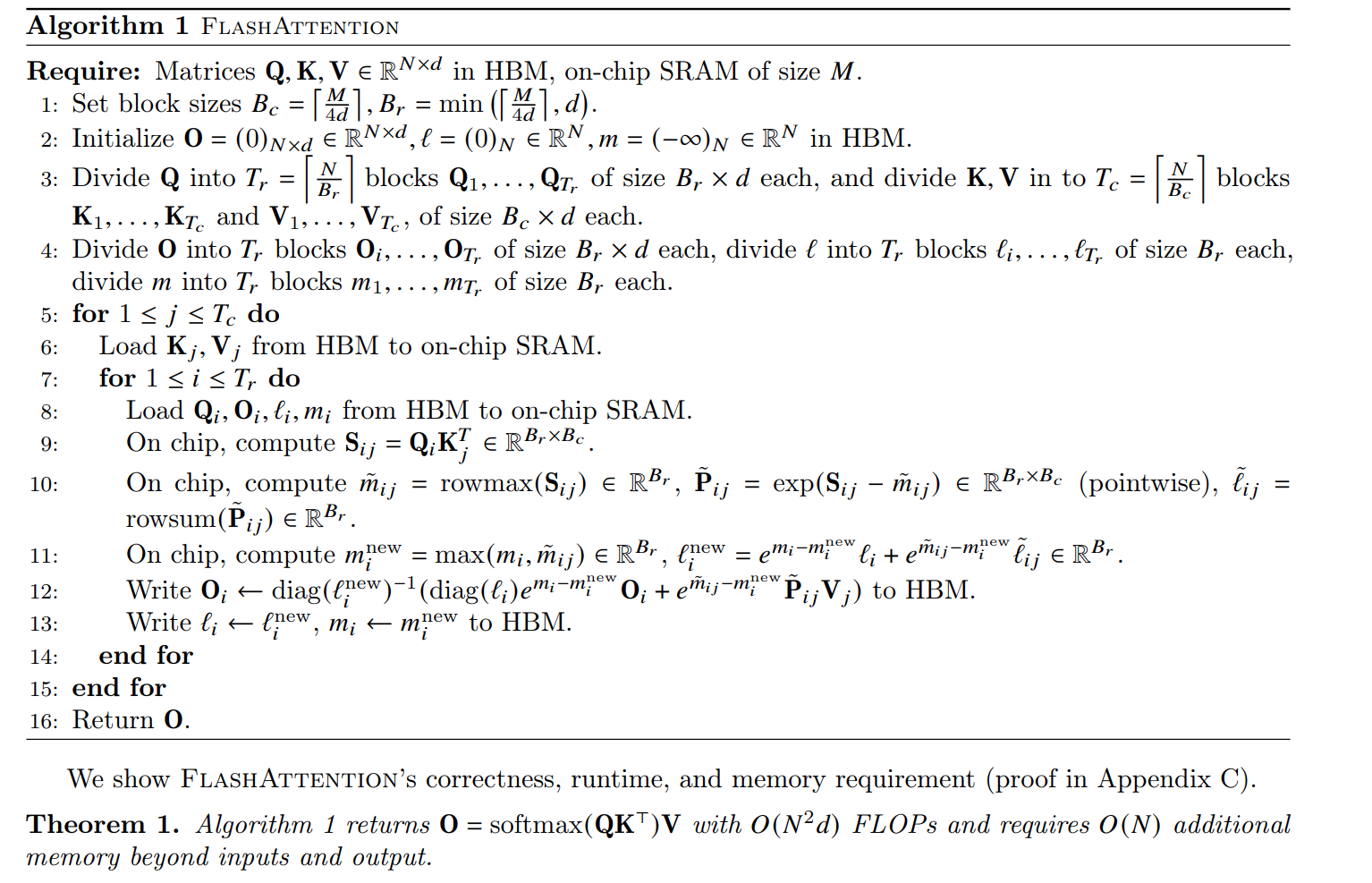
## Original Attention Algorithm



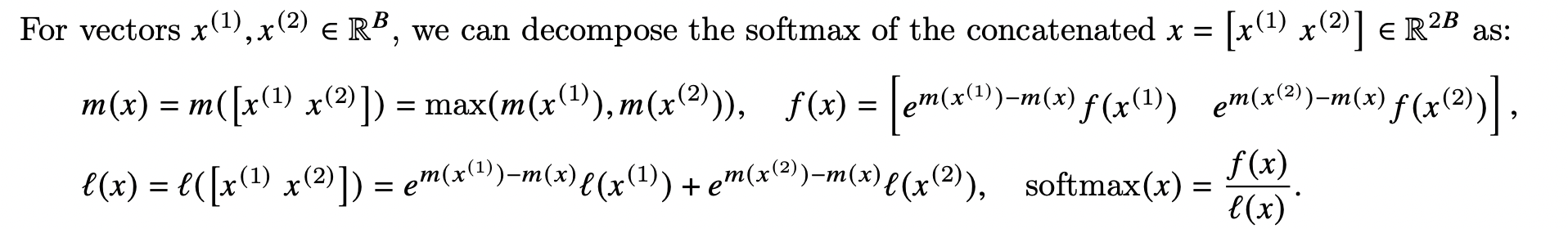
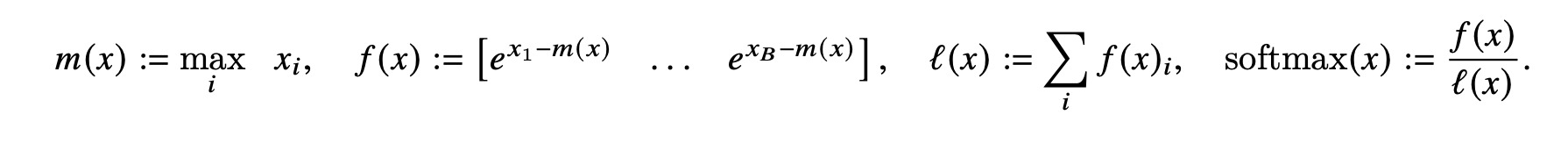
* Given Q,K, V (since attention deals with Query, key value vectors)
  + Want to get an O
* S = QK^T
* P = Softmax(S)
* O = PV
* HBM = high bandwidth memory

## Main idea of flash attention

* we split the inputs Q, K, V into blocks [**Tiling**]
  + enables us to implement our algorithm in one CUDA kernel
* load them from slow HBM to fast SRAM
* compute the attention output with respect to those blocks [**Recomputation]**
* By scaling the output of each block by the right normalization factor before adding them up, we get the correct result at the end.



## Tiling



* But what is x^1 and x^2

## Recomputation

* Incrementally calculates the soft max instead of storing intermediate matrices
* Uses kernel fusion

## How they prove their faster algorithm

* Proved asymptotically
* We discussed this extensively last week

## Key idea

* Breaks matrix apart in order to be able to put smaller matrices in SRAM

## Useful youtube videos

* [Flash Attention 2.0 with Tri Dao (author)! | Discord server talks](https://youtu.be/IoMSGuiwV3g?si=FvqmG9kY5bRBEgES)
* <https://www.youtube.com/live/gMOAud7hZg4?si=1KZU1AV3IMm6_4It>
* [MedAI #54: FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness | Tri Dao](https://youtu.be/FThvfkXWqtE?si=WSobjVAZZCff9-yl)

## Code

* <https://github.com/Dao-AILab/flash-attention>

## Code notes

## Meeting Oct 8, 2023

* Theorem 2: if you assume a certain component is 1/n^2 then it cancels out the original runtime complexity and it makes it linear
* Fun fact: big bird is flash sparse attention
* Why is runtime connected to sequence length?
  + Sequence length is the n variable
* Why can’t they also tile S and P?
  + Imo i think they can do this but they also might need to have the matrices together for softmax (because it
* What is gradient checkpointing
  + Can find this in other paper