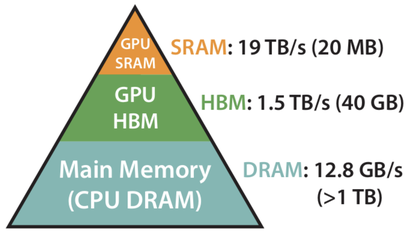
[Re] FlashAttention Three Years On

Kia Ghods*, Jishnu Roychoudhury*, Addison Wu*
Princeton University, *equal contribution

Motivation

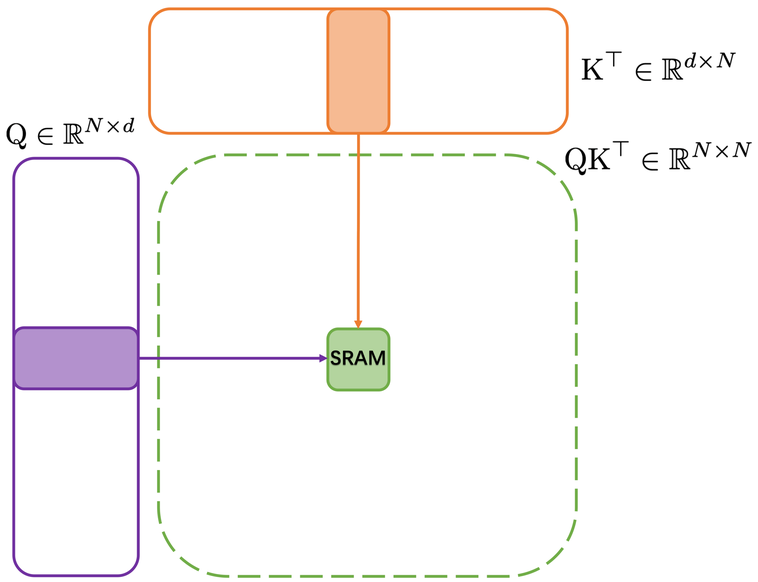
- **Attention** is **slow** on long sequences, O(N^2)
- Limitations to deep-learning operations:
  - Compute — # of operations
  - Memory — HBM accesses

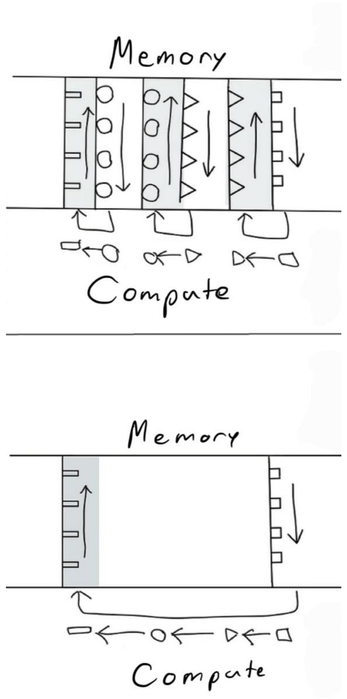
- **Key Observation:** Attention is **memory-bound**
- Can reduce # HBM accesses to improve speed



Memory Hierarchy with Bandwidth & Memory Size

Strategy: Tiling, Fusion





- **Tiling:** Split Q, K, V into blocks to fit into SRAM
- **Streaming:** $x = \begin{bmatrix} x^{(1)} & x^{(2)} \end{bmatrix} \in \mathbb{R}^{2B} \quad m(x) = m(\begin{bmatrix} x^{(1)} & x^{(2)} \end{bmatrix}) = \max(m(x^{(1)}), m(x^{(2)}))$

$$f(x) = \begin{bmatrix} e^{m(x^{(1)})-m(x)} f(x^{(1)}) & e^{m(x^{(2)})-m(x)} f(x^{(2)}) \end{bmatrix}$$

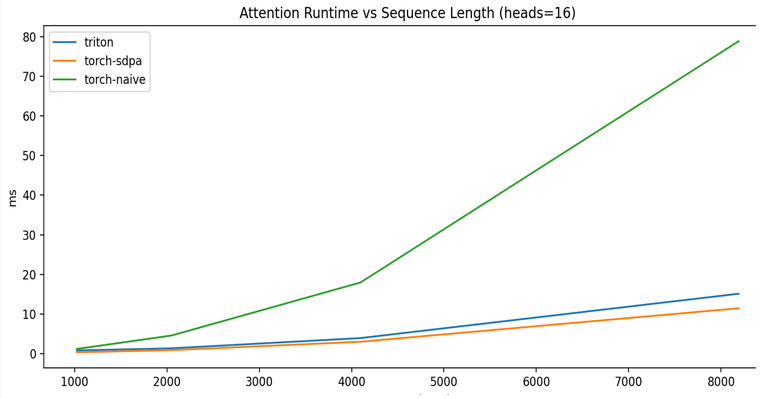
$$\ell(x) = \ell(\begin{bmatrix} x^{(1)} & x^{(2)} \end{bmatrix}) = e^{m(x^{(1)})-m(x)}\ell(x^{(1)}) + e^{m(x^{(1)})-m(x)}\ell(x^{(2)})$$

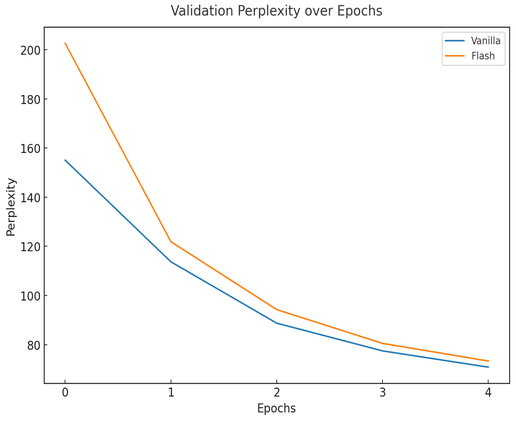
$$\mathrm{softmax}(x) = \frac{f(x)}{\ell(x)}$$

- **Fusion:** Combine scaling, masking, softmax into one kernel to minimize HBM traffic
- **Reduces HBM access** from O(Nd + N^2) to O(N^2d^2/M) where M is SRAM memory size

Experiments

Triton Kernel Performance





Extension: Multi-Query Attention