

Computing $\text{LT}(A \cdot B^T) \cdot C$

• slow for long contexts

Not utilizing full memory bandwidth available between HBM and MEM

- Outer products are not compute intensive

• bottlenecked by bandwidth

• Leads to pretty poor performance

We use a block-based algorithm to greatly improve the performance

3

4

Computing $\text{LT}(A \cdot B^T) \cdot C$

- Slow for long contexts
 - Not utilizing full memory bandwidth available between HBM and VMEM
 - Outer products are not compute intensive
 - bottlenecked by bandwidth
 - Leads to pretty poor performance
- We use a block-based algorithm to greatly improve the performance

Block-wise Algorithm

