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

- 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

34

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

- 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