$[B, A] \longrightarrow [A]$

1. Warp Read Warp Reduce

* Basic Method
  1. Each warp reads 32 consecutive values.
     A block reads 32 lines. ? (32x32 or 16x64)
  2. Save values in shared memory.
     Each warp reads a column.
  3. Use warp reduce.

* Requirement for Efficiency
  $B \geq 32$

* Different cases
  1. $\lfloor A/32 \rfloor \geq N_{sm}$
     Each warp reads multiple lines and reduce first.
  2. $\lfloor A/32 \rfloor < N_{sm}$
     Each warp reads 1 line.
     Different blocks reduce different part of the same column.
     Use atomic add for final result.
  3. combination of 1 and 2

* Corner Case / Margin Part ($a = A < 32$ or $a = A \bmod 32$)
  let $n = \lfloor 32/a \rfloor$
  1. 32 warps, each warp reads n lines (margin part ?)
     or na warps, each warp reads 32 values (corner case)
  2. write and then read in shared memory
  3. each warp reduce 32 values, atomic add at last
     ?

$[B, A] \rightarrow [A]$

2. Thread Reduce

* Basic Method
  ① Each block reads 1024 consecutive columns.
  ② Each thread reduce the values it reads.

* Requirement for Efficiency
  $H > 1024$

* Different Cases
  ① same as basic
  ② different blocks reduce the same column
  use atomic reduce for final result

* Corner Case / Margin Part
  $n = \lfloor a/1024 \rfloor$
  ① threads in a block read consecutive na values
  ② each thread add up the values
  ③ atomic reduce

3. Warp Read Part Warp Reduce

Can't be implemented in Doce ?

HIGH DIMENSION

* let $R_1 \to R_2$ to be a basic reduction with ≥ 2 DIMs
  - ① $AR_1 \to AR_2$
    use AN blocks
  - ② $BR_1 \to R_2$
    2.1) use BN blocks, reduce atomically
    2.2) each thredd reads Bn values and reduce first
    2.3) mixture
    * if $N_{blocks} > N_{sm}$ use 2.1
      if $N_{blocks} < N_{sm}$
      use 22 or 23 to let $N_{blocks} > N_{sm}$

* ① the basic method depends on last two dimensions
  ($B_2 A_0 \to A_0$ or $A_1 B_0 \to B_0$)
  ② discussion with A^*, B^*, A_0, B_0

$$A^* = \prod_{i=1}^{N_A} A_i \qquad B^* = \prod_{i=1}^{N_B} B_i$$

for 2-D case, just let $A^* = B^* = 1$

① $B<32$ $A>1024$: TR

② $B>32$ $A<1024$: URWR

③ $B>32$ $A>1024$: to be tested

④ A^*B^* large $B<32$ $A<1024$: TR?

⑤ corner case: A^*B^* large $B \cdot A < 32 \cdot 32$ (optional)