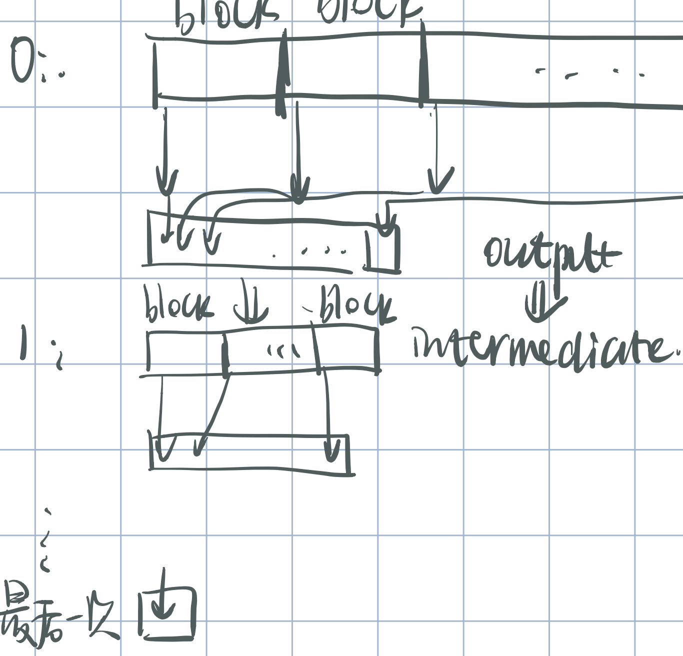
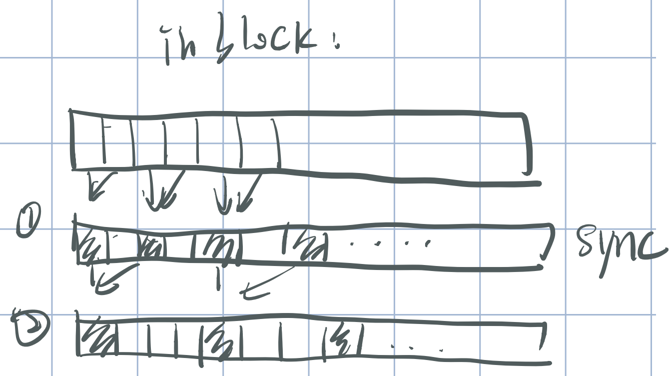
⚠ 这里所有的 idx 都不对. 注意后期看如何改一下[illegible]

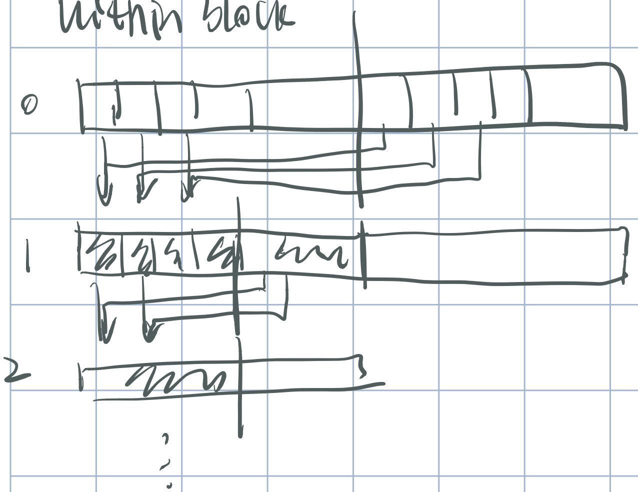
Reduction 0&1.



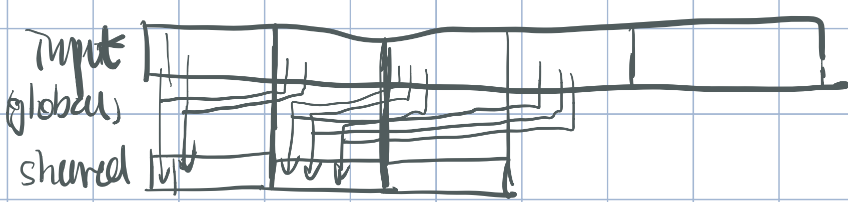


Reduction 2.

within block



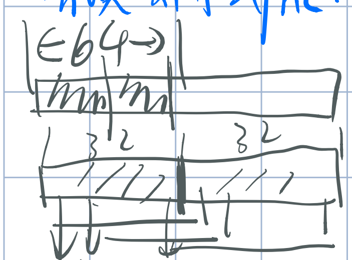
Reduction 3



每个 block 里 shared mem 存储 2 个 block 的 global mem 里的内容.
所以 kernel 可以少一半.

Reduction 4

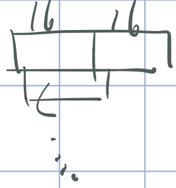
原理: 同一 warp 里的是同步指令.
所以不用 sync. 展开循环.



可以保证一定按顺序执行.
因为 SIMT

Reduction 5

用模板参数 告知 block 信息.
让 compiler 完成 loop unrolling.
block size 的最大是 512 threads,
所以可以有限循环展开.



Reduction 6

应该是每个grid间规约起来!