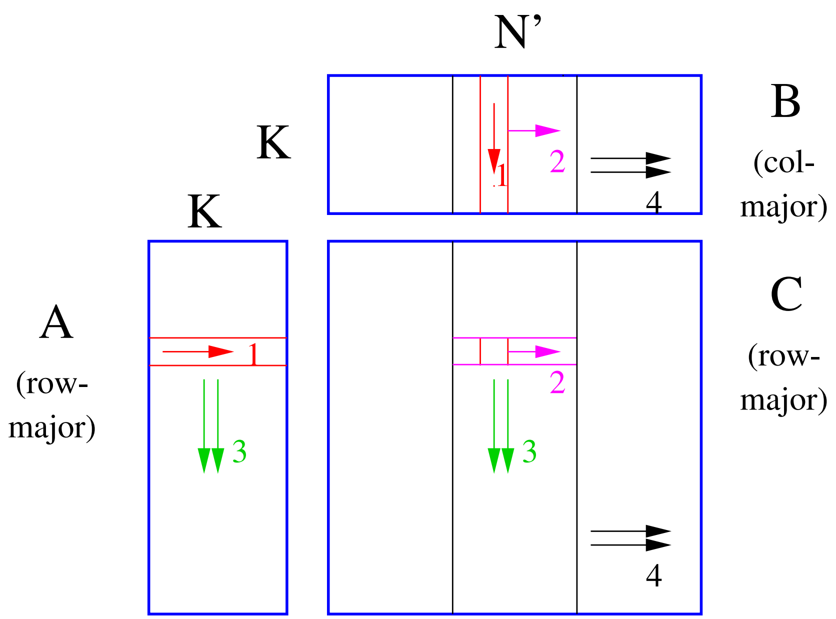
3 Memory Access Pattern Issues: local Matrix Multiply

- consider the multiply $C+=AB$, where C is $N\times N$ and A is $N\times K$ on a U170
- inner loop kernel updates (a 4×4 sub-matrix of) C using $4\times K$ ($K\times 4$) strip of A (B)
  - uses cache lookahead to minimize top-level cache misses
  - achieves 300–320 MFLOPs for $K\approx 128$ and A & B in top-level cache
- for large N, the following alg. ensures unit stride along the 2 innermost iterations:



- partition B st. $KN'\approx\min(|$level-2 cache$|, |$TLB$| * |$page$|/2)$
  - ie. $K\times N'$ sub-block of B takes no more than 1/2 the TLB entries
- on UltraSPARC I, optimal $K=88, N'=256$, sustained 260–280 MFLOPs for $512\le N<6144$