**Exercise 6**

For the simple matrix multiplication (P = M\*N) based on row-major layout,

which input matrix will have coalesced accesses?

A. M

B. N

C. M, N

D. Neither

**Answer:** B. N

Let us assume

M and N to be 4x4 matrices

block dimension 2x2

For Block(0.0) the elements accessed by each thread for k=0 is given in the following table. As seen matrix N has coalesced access.

|  |  |  |
| --- | --- | --- |
|  | M | N |
| Thread(0,0) | M[0] | N[0] |
| Thread(0,1) | M[0] | N[1] |
| Thread(1,0) | M[4] | N[0] |
| Thread(1,1) | M[4] | N[1] |

**Exercise 7**

For the tiled matrix–matrix multiplication (M\*N) based on row-major layout,

which input matrix will have coalesced accesses?

A. M

B. N

C. M , N

D. Neither

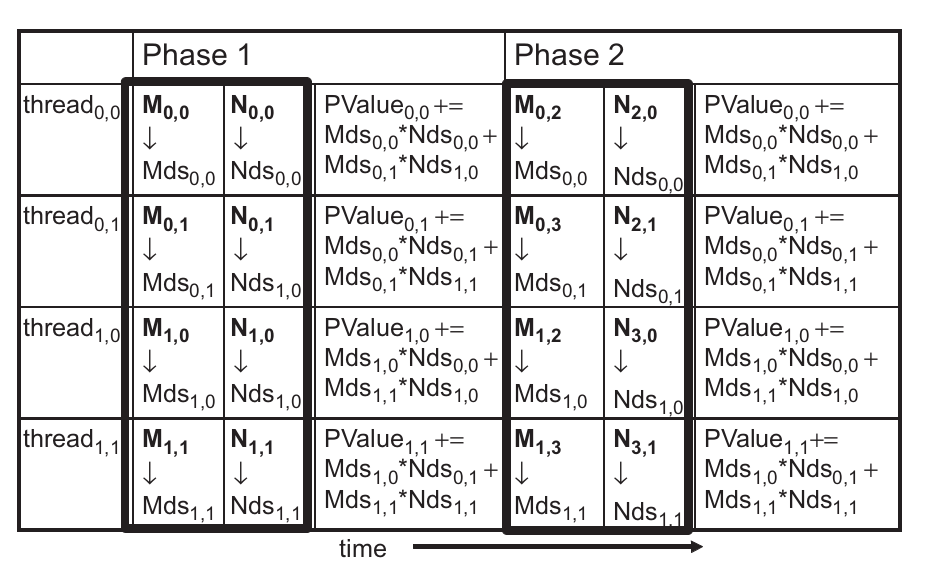
**Answer:** C. M, N

Linear representation of matrices:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| M(0,0) | M(0,1) | M(0,2) | M(0,3) | M(1,0) | M(1,1) | M(1,2) | M(1.3) | M(2,0) | M(2.1) | M(2,2) | M(2,3) | M(3.0) | M(3.1) | M(3,2) | M(3,3) |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| N(0,0) | N(0,1) | N(0,2) | N(0,3) | N(1,0) | N(1,1) | N(1,2) | N(1.3) | N(2,0) | N(2.1) | N(2,2) | N(2,3) | N(3.0) | N(3.1) | N(3,2) | N(3,3) |

Memory access:



**Exercise 8:**

For the simple reduction kernel, if the block size is 1024 and warp size is 32,

how many warps in a block will have divergence during the 5 th iteration?

A. 0

B. 1

C. 16

D. 32

**Answer:** D. 32

For a block size of 1024 there would be 32 warps. In the 5th iteration stride value would be 16 and threads whose indices are multiples of 32 will perform addition. 32 threads (0, 32, 64, ... , 928, 960, 992) will perform addition in iteration 5 while the rest of the threads would have been skipped. However, as the indices of these threads are spread out there would be one thread per warp which performs addition and the remaining threads are skipped. Thus all 32 warps will have divergence.

**Exercise 9:**

For the improved reduction kernel, if the block size is 1024 and warp size is

32, how many warps will have divergence during the 5 th iteration?

A. 0

B. 1

C. 16

D. 32

**Answer:** A. 0

In the 5 th iteration there will be 32 threads that perform addition and 32 threads that skip it. So, there won’t be any divergence in any of the warps.

**Exercise 11:**

For tiled matrix multiplication out of the possible range of values for

BLOCK\_SIZE , for what values of BLOCK\_SIZE will the kernel completely avoid

un-coalesced accesses to global memory? (You need to consider only square

blocks.)

**Answer:** For BLOCK\_SIZE of 1 there would be neither coalesced nor un-coalesced access to global memory. For all other values of BLOCK\_SIZE both input matrices will have coalesced access to global memory.