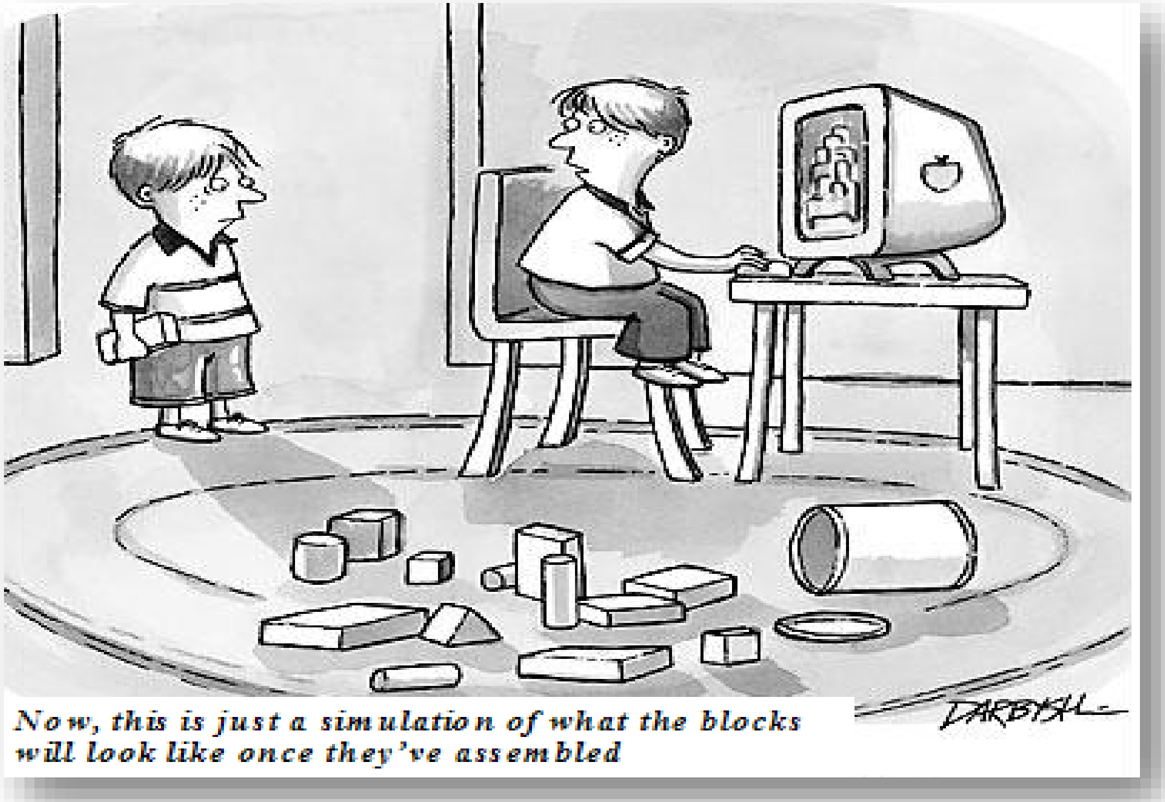
ECE569
Module 27



Now, this is just a simulation of what the blocks will look like once they've assembled

- Matrix Multiplication – Thread Divergence

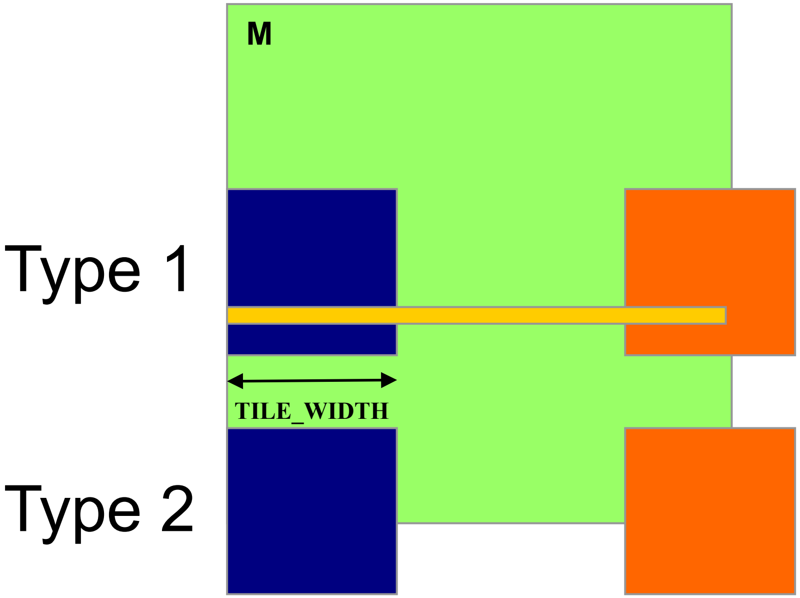
Performance Impact of Control Divergence

```
if(Row < Width && t * TILE_WIDTH+tx < Width)
 ds_M[ty][tx]=M[Row*Width+p*TILE_WIDTH+tx];
else
  ds_M[ty][tx] = 0.0;

if (p*TILE_WIDTH+ty < Width && Col < Width)
 ds_N[ty][tx]=N[(p*TILE_WIDTH +ty)*Width+Col];
else
 ds_N[ty][tx] = 0.0;
```

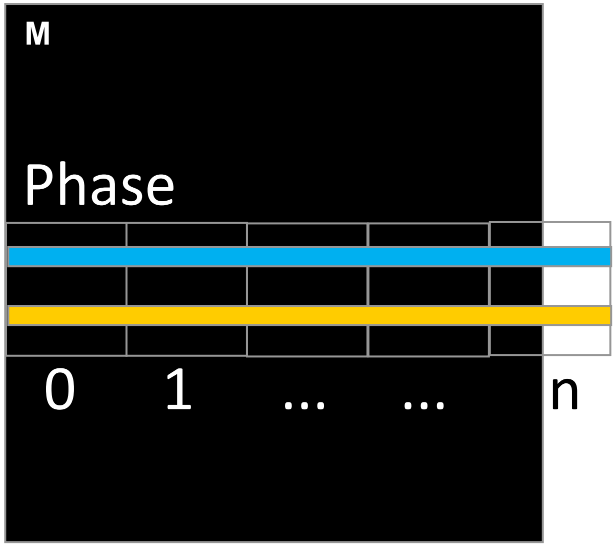
Two types of blocks in loading M Tiles

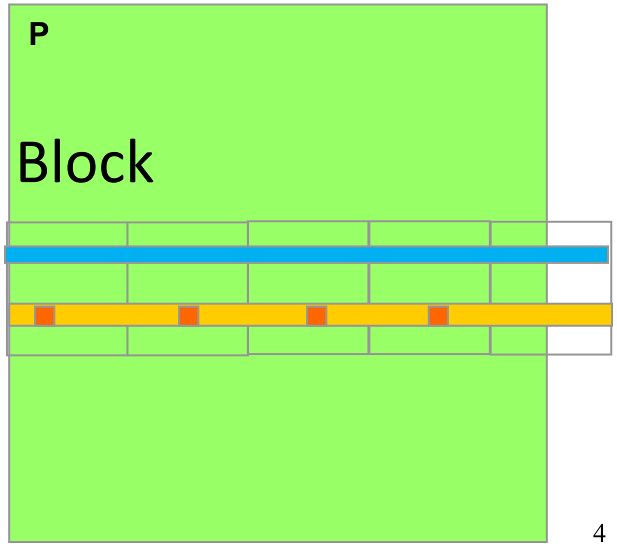
- Type 1
  - Blocks whose tiles are all within valid range until the last phase.
- Type 2
  - Blocks whose tiles are partially outside the valid range all the way



Analysis of Control Divergence Impact

- Assume 16x16 tiles and thread blocks, and matrices of 100x100. How many thread blocks are allocated? How many phases will each thread go through?





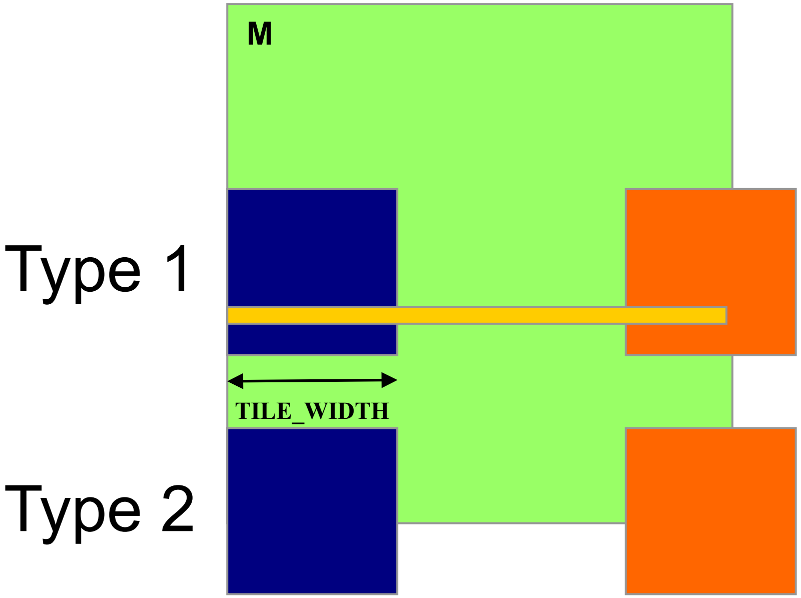
Analysis of Control Divergence Impact

- Assume 16x16 tiles and thread blocks
- Assume square matrices of 100x100
- How many thread blocks are allocated?
  - There are 49 thread blocks (7 in each dimension)

- How many phases will each thread go through?
  - Each thread will go through 7 phases (ceiling of 100/16)

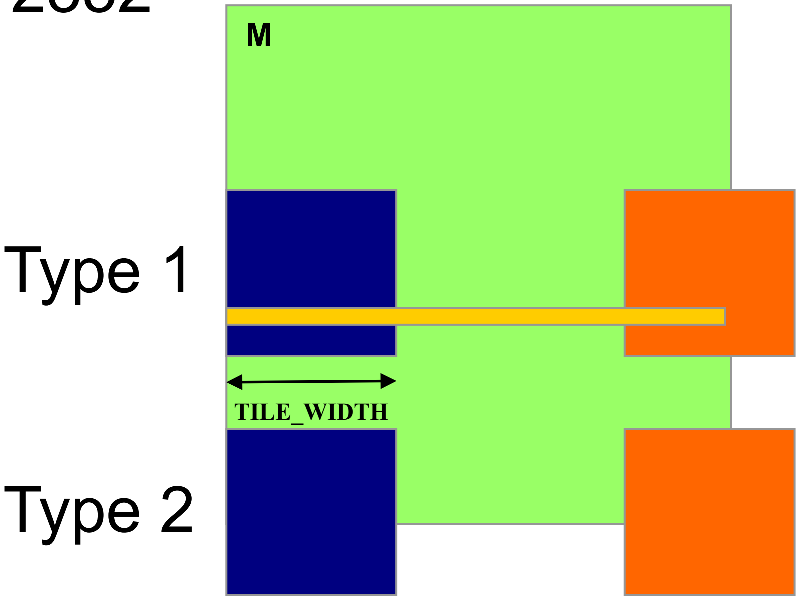
Control Divergence in Loading M Tiles

- How many Type 1 Blocks?
- How many warps in Type 1?
- How many warp phases in Type 1?
- How many warps observe control divergence in Type 1?



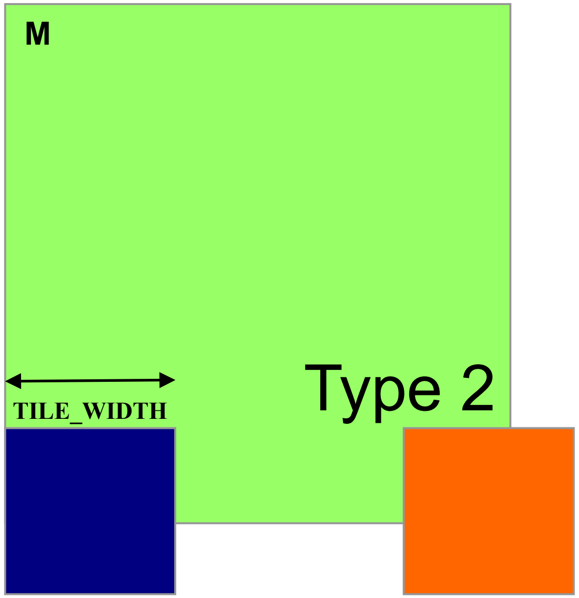
Control Divergence in Loading M Tiles

- How many Type 1 Blocks?
  - 6 rows, 7 columns => 42 blocks
- How many warps in Type 1?
  - Each block 16x16, 8 warps => 336 warps
- How many warp phases in Type 1?
  - 7 phases per warp => 7*336 = 2352
- How many warps observe control divergence in Type 1?
  - Only last phase observes divergence
    - 1*336 = **336 warps** have control divergence



Control Divergence in Loading M Tiles

- How many Type 2 Blocks?
- How many warps in Type 2?
- How many warp phases in Type 2?
- How many warps observe control divergence in Type 2?



Control Divergence in Loading M Tiles

- How many Type 2 Blocks and Warps?
  - 7 blocks , each 8 warps => 56 warps
- How many warp phases in Type 2?
  - 7 phase each => 7*56 = 392
- How many warps observe control divergence in Type 2?
  - 100x100 => last 4 rows of the matrix are processed by blocks in row 7.
    - Each row 16 elements => 64 elements
    - 2 warps/block process valid data unit last phase.
      - Each phase requires boundary check 2*7 = **14 warps**
    - 6 remaining warps outside the valid range (not a concern)

Overall Impact of Control Divergence

- **Type 1 Blocks:**
  - 336 out of 2,352 warp-phases have control divergence
- **Type 2 Blocks:**
  - 14 out of 392 warp-phases have control divergence
- **The performance impact is expected to be less than 12%**

Conclusions

- The calculation of impact of control divergence in loading **N** **tiles** is somewhat different and is left as an exercise
- The estimated performance impact is data dependent.
  - For larger matrices, the impact will be significantly smaller
- In general, the impact of **control divergence** for boundary condition checking for **large input data** sets should **be insignificant**
  - One **should not hesitate to use boundary checks** to ensure full functionality
- The fact that a kernel is full of control flow constructs **does not mean** that there will be heavy occurrence of control divergence
  - We will cover some algorithm patterns that naturally incur control divergence (such as parallel reduction)

Next

- **Atomic Operations**