



GPU Teaching Kit
Accelerated Computing



Module 5.2 – Thread Execution Efficiency

Performance Impact of Control Divergence

Objective

- To learn to analyze the performance impact of control divergence
 - Boundary condition checking
 - Control divergence is data-dependent

Performance Impact of Control Divergence

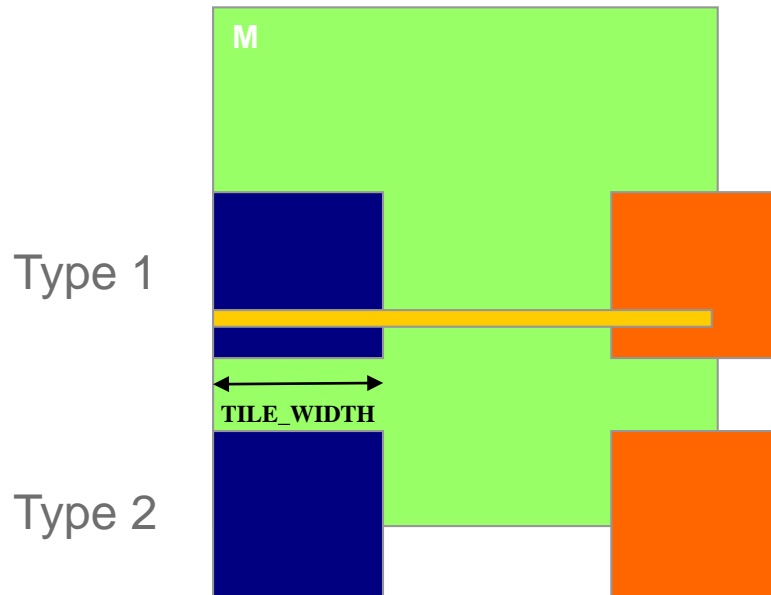
- Boundary condition checks are vital for complete functionality and robustness of parallel code
 - The tiled matrix multiplication kernel has many boundary condition checks
 - The concern is that these checks may cause significant performance degradation
 - For example, see the tile loading code below:

```
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

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



Analysis of Control Divergence Impact

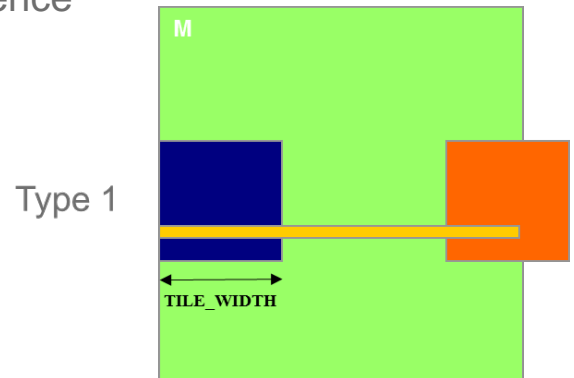
- Assume 16x16 tiles and thread blocks
- Each thread block has 8 warps (256/32)
- Assume square matrices of 100x100
- Each thread will go through 7 phases (ceiling of 100/16)

- There are 49 thread blocks (7 in each dimension)

Control Divergence in Loading M Tiles

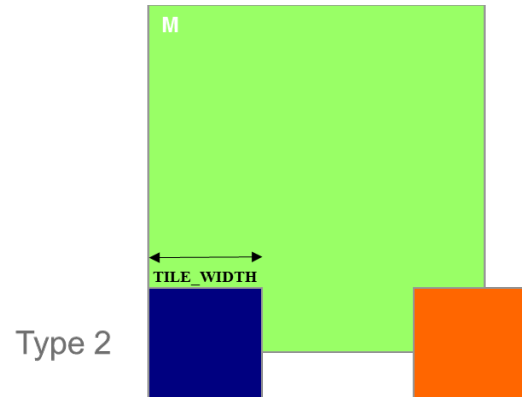
- Assume 16x16 tiles and thread blocks
- Each thread block has 8 warps (256/32)
- Assume square matrices of 100x100
- Each warp will go through 7 phases (ceiling of 100/16)

- There are 42 (6*7) Type 1 blocks, with a total of 336 (8*42) warps
- They all have 7 phases, so there are 2,352 (336*7) warp-phases
- The warps have control divergence only in their last phase
- 336 warp-phases have control divergence



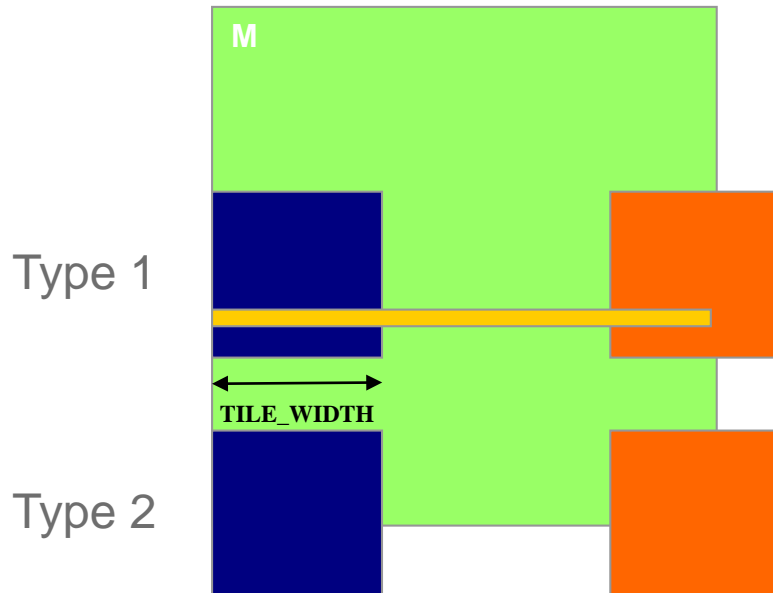
Control Divergence in Loading M Tiles (Type 2)

- Type 2: the 7 block assigned to load the bottom tiles, with a total of 56 ($8*7$) warps
- They all have 7 phases, so there are 392 ($56*7$) warp-phases
- The first 2 warps in each Type 2 block will stay within the valid range until the last phase
- The 6 remaining warps stay outside the valid range
- So, only 14 ($2*7$) warp-phases have control divergence



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% ($350/2,944$ or $(336+14)/(2352+14)$)



Additional Comments

- 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) in the Parallel Algorithm Patterns modules



GPU Teaching Kit

Accelerated Computing



The GPU Teaching Kit is licensed by NVIDIA and the University of Illinois under the [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).