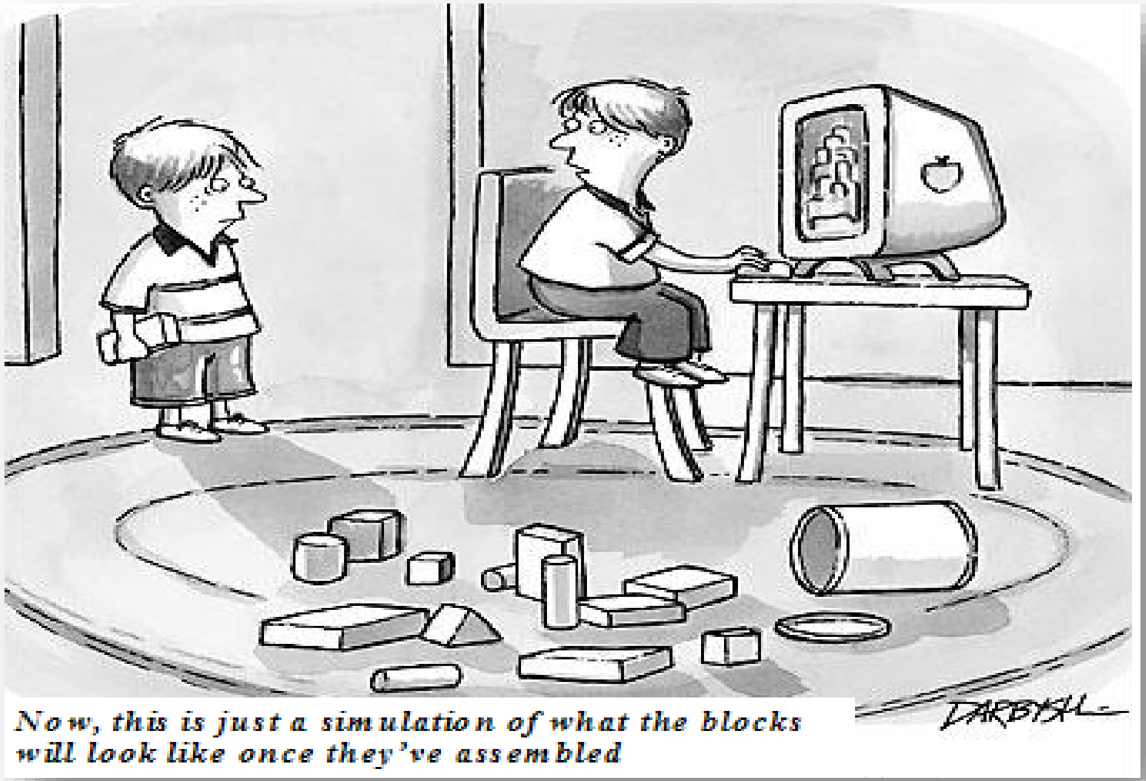
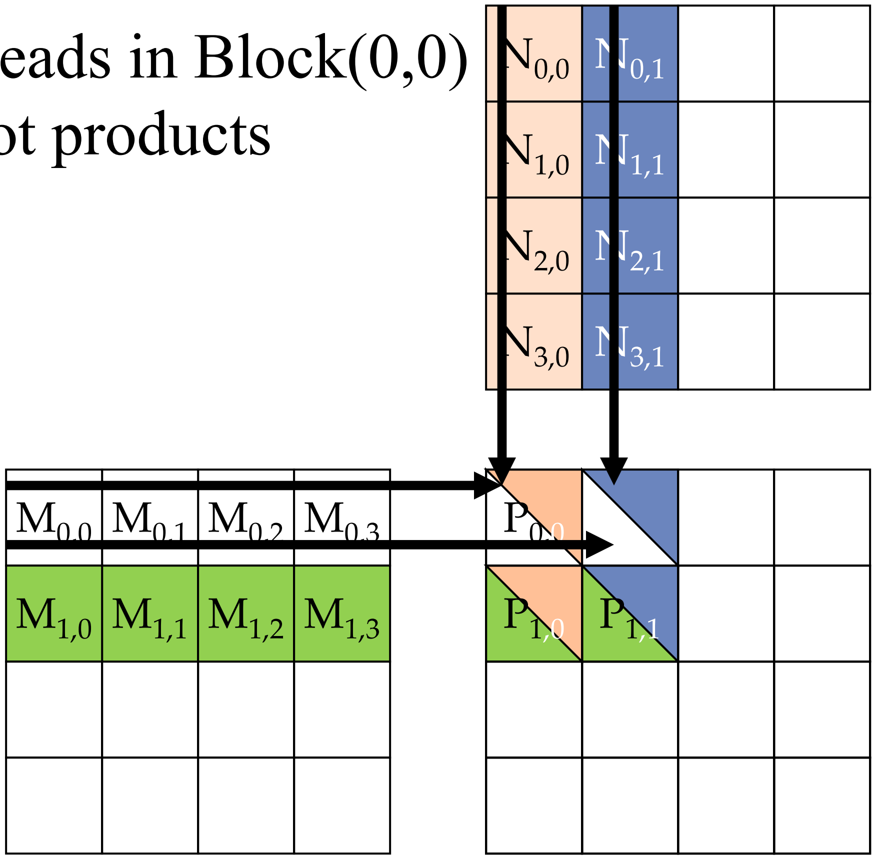
ECE569
Module 22


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

- Tiling Concept

Multiple accesses to the same address

Threads in Block(0,0)
4 dot products

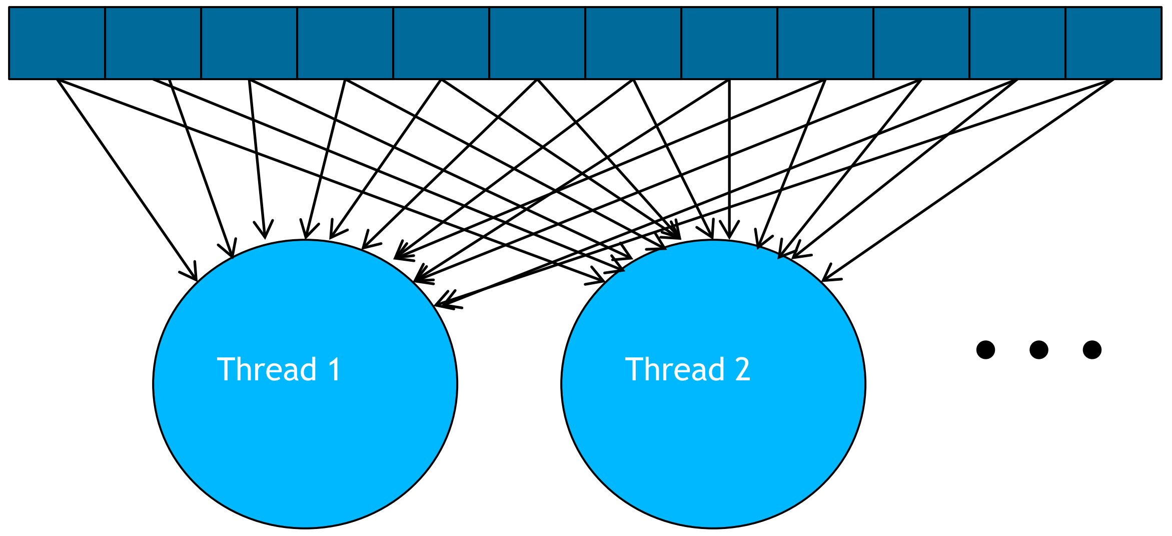


- **For both P0,0 and P0,1**
  - row 0 of M is common!
- **For both P0,0 and P1,0**
  - Col 0 of N is common
- **In fact each element in M is accessed twice**
  - Same for N

- If threads accessing the same address collaborate
  - Operate on shared memory for the common addresses
- We would reduce the global memory access by 2x

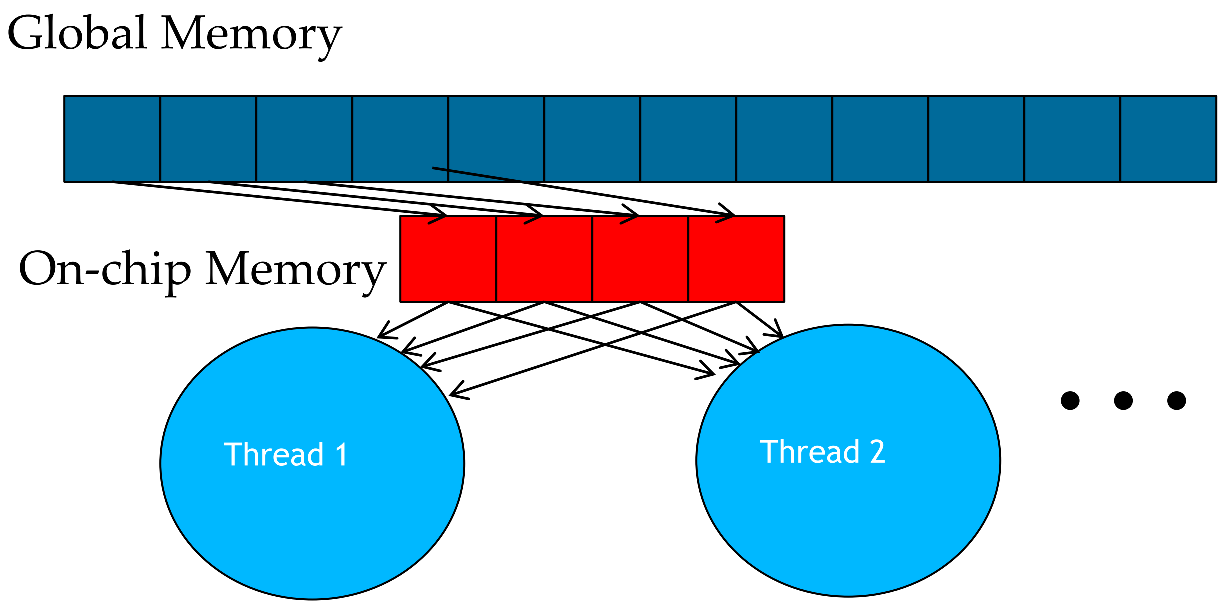
Memory Access Pattern

- **Global Memory Access Pattern of the Basic Matrix Multiplication Kernel**



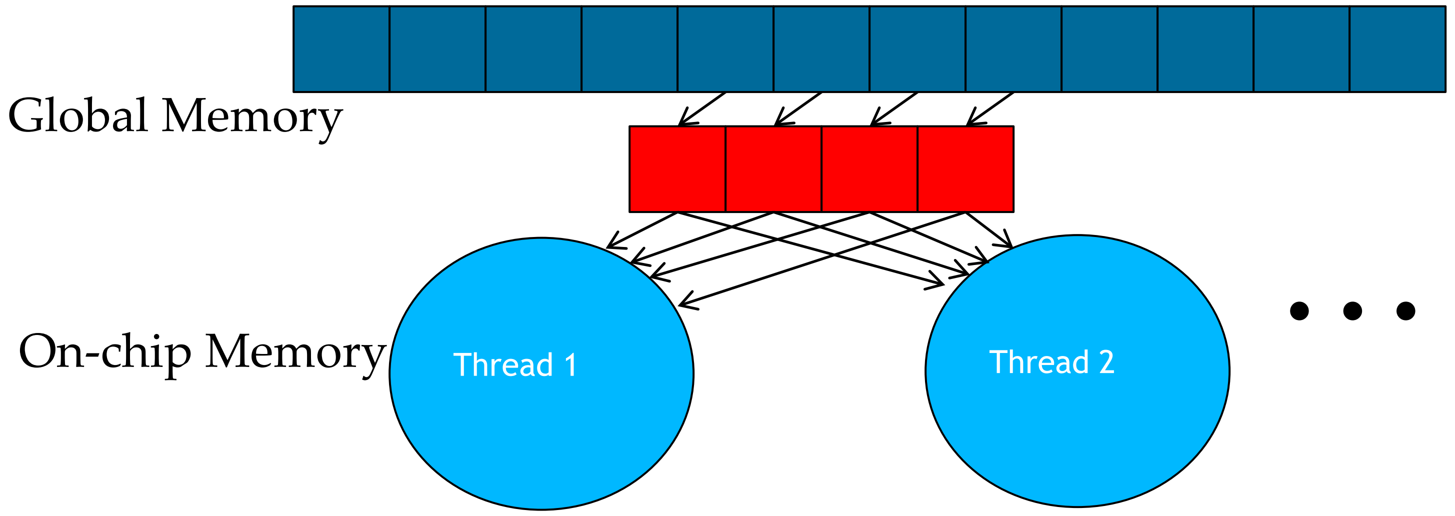
Tiling/Blocking - Basic Idea

- If threads are accessing the same global memory address space
- Divide the global memory content into tiles
- Focus the computation of threads on one or a small number of tiles at each point in time



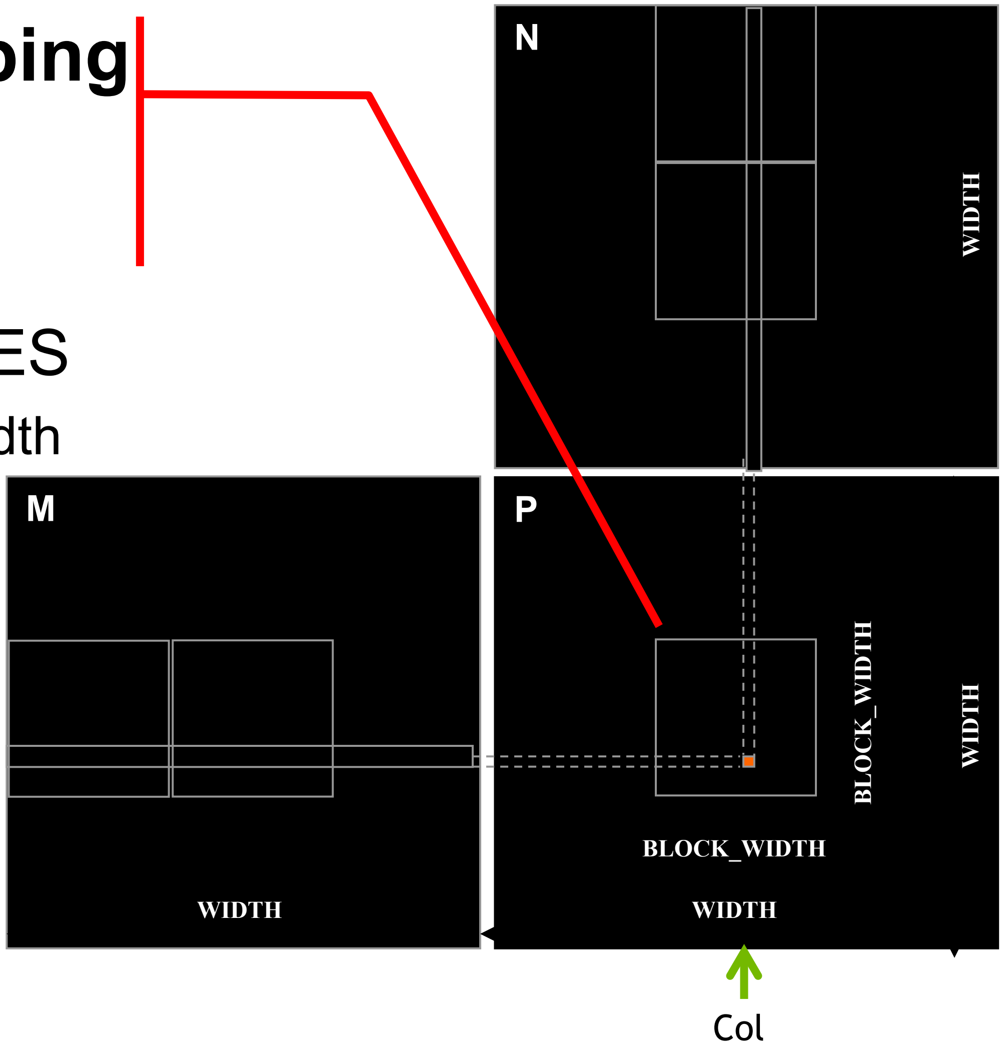
Tiling/Blocking - Basic Idea

- **Program transformation technique**
  - Localizes memory locations accessed among threads and the timing of their accesses.
  - Divides the execution into phases to the scope of the shared data



Thread Blocks: Natural Tiling of the Matrix

- **Thread-to-data mapping**
  - divides P into tiles
- **Thread Block**
  - Matrix divided into TILES
    - Block_WidthxBlock_Width



- Explore data reuse opportunities across threads in a block
- Divide operation into phases rather than working on the entire row/column of data!

Concept of Tiling

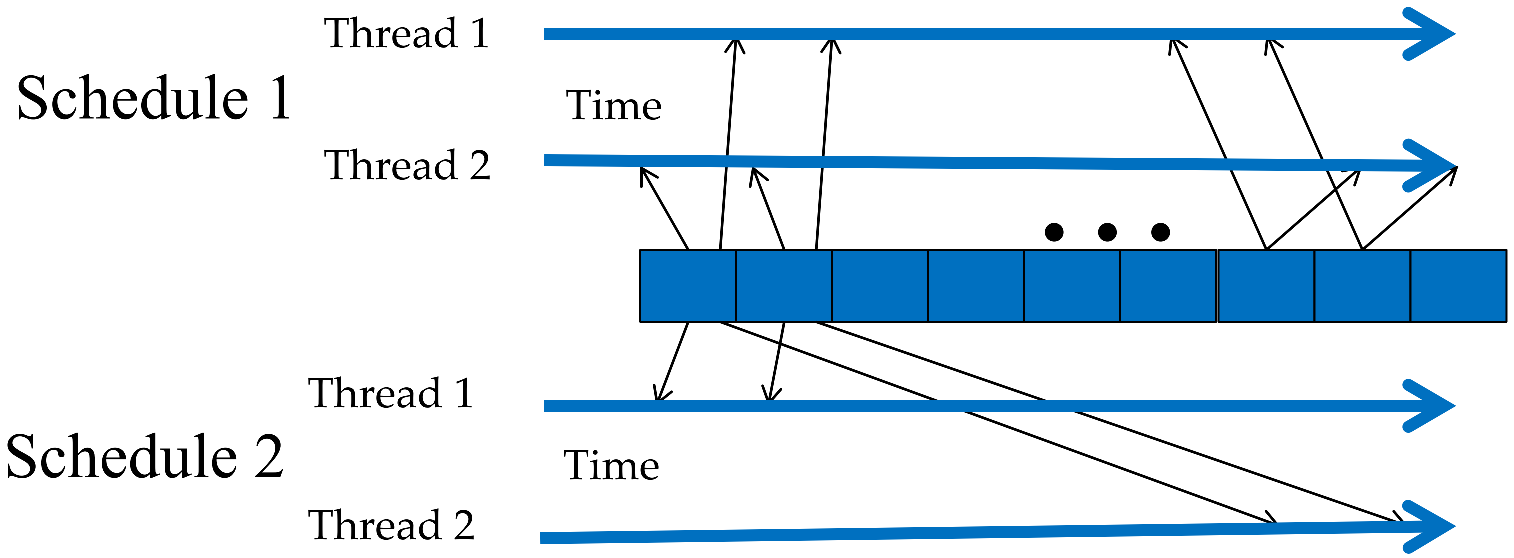
- **In a congested traffic system, significant reduction of vehicles can greatly improve the delay seen by all vehicles**
  - Carpooling for commuters
  - Tiling for global memory accesses
    - drivers = threads accessing their memory data operands
    - cars = memory access requests

Challenges of Tiling

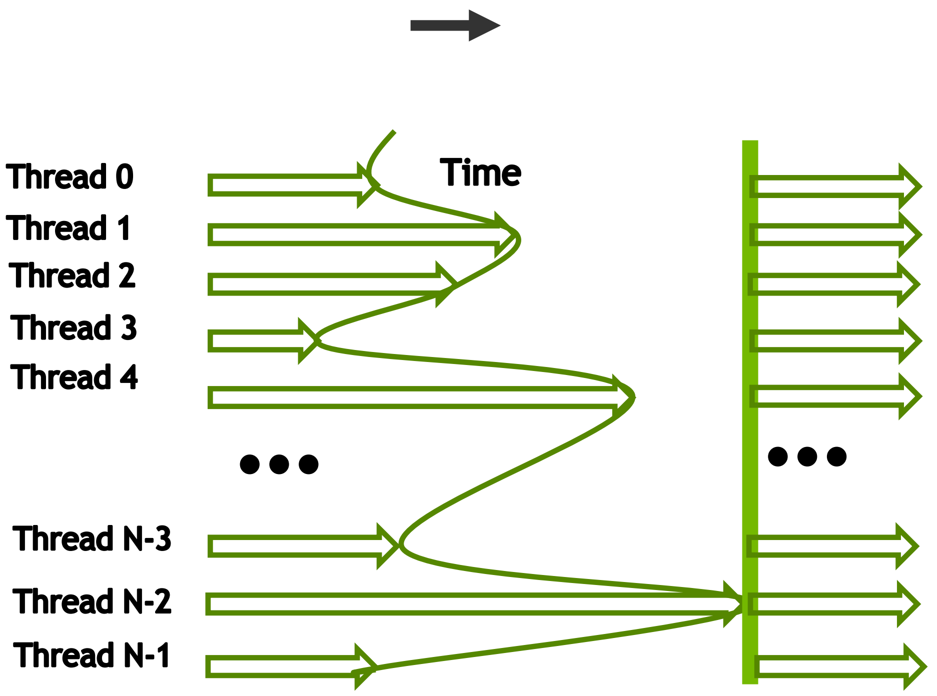
- **Some carpools may be easier than others**
  - Car pool participants need to have similar work schedule
  - Some vehicles may be more suitable for carpooling
- **Similar challenges exist in tiling**

Need a similar schedule

- **Good: when threads have similar access timing**
- **Bad: when threads have very different timing**



Barrier Synchronization for Tiling



Next

- Tiling for matrix multiplication