Cuda Thread Divergence

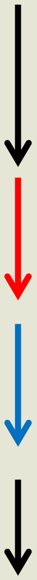
1

Minimizing thread divergence

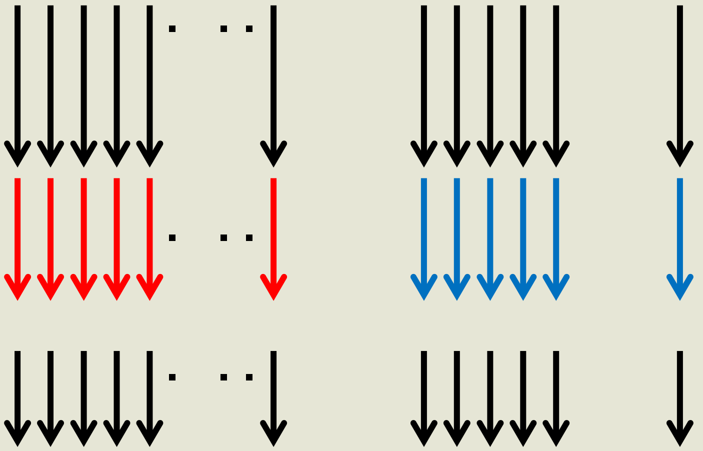
- Warp – Set of threads that execute the same instruction at a time.
- SIMD – Single Instruction, Multiple Data
- SIMT – Singe instruction, Multiple Threads

2

```
Blah;
Blah;
 …
If ( . . .)
{
 //then do smth
}
Else
{
//else do smth
}
Blah;
Blah;
```

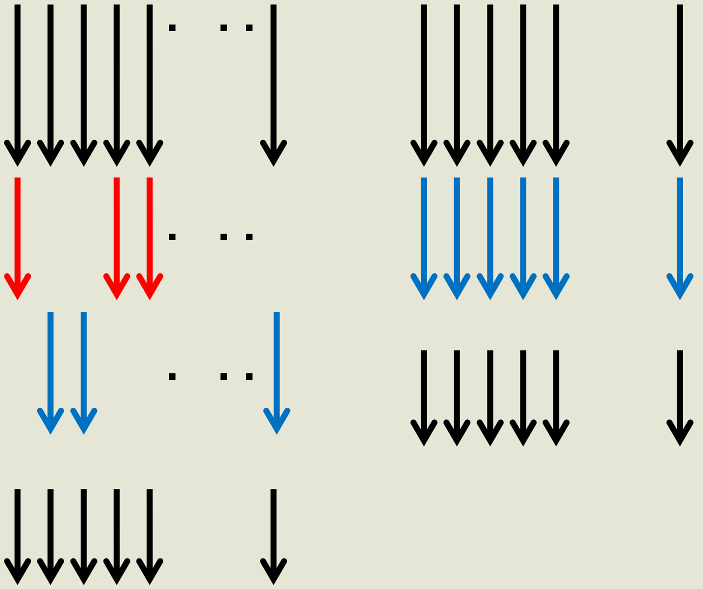


3



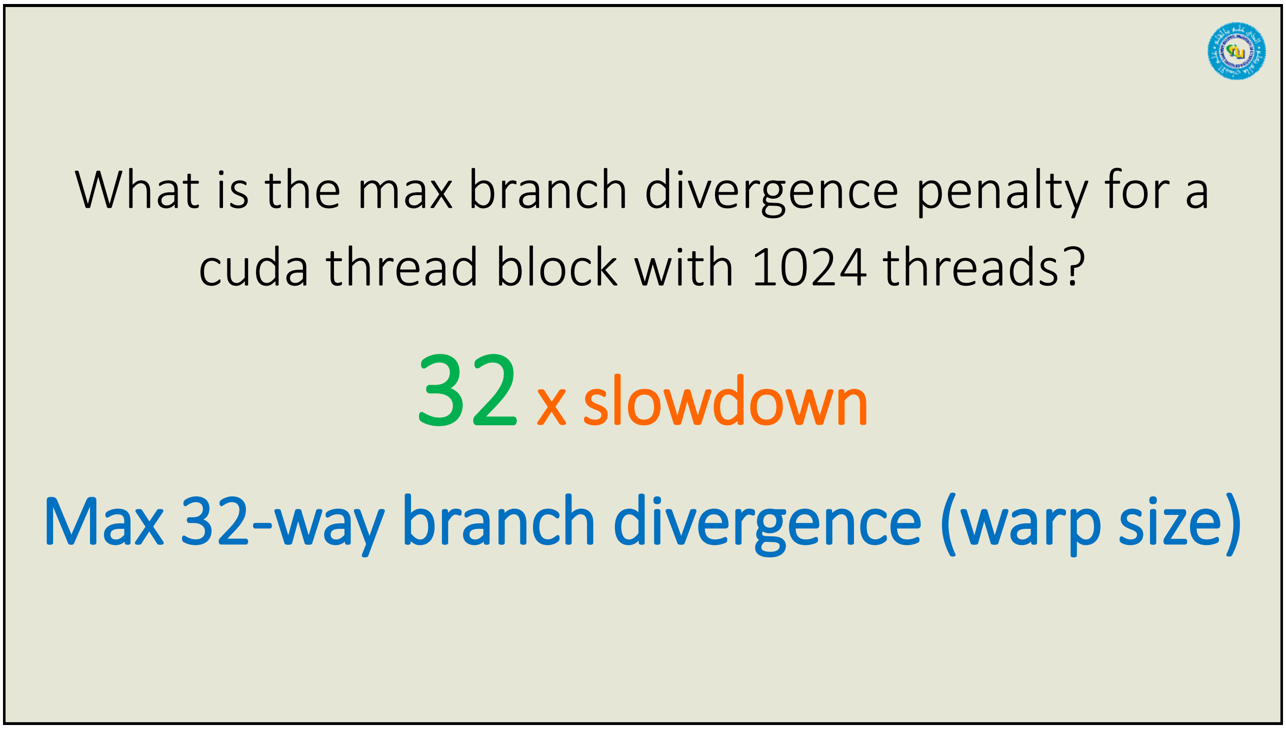
4

2-Way branch divergence



What is the max branch divergence penalty for a cuda thread block with 1024 threads?

_______ x slowdown

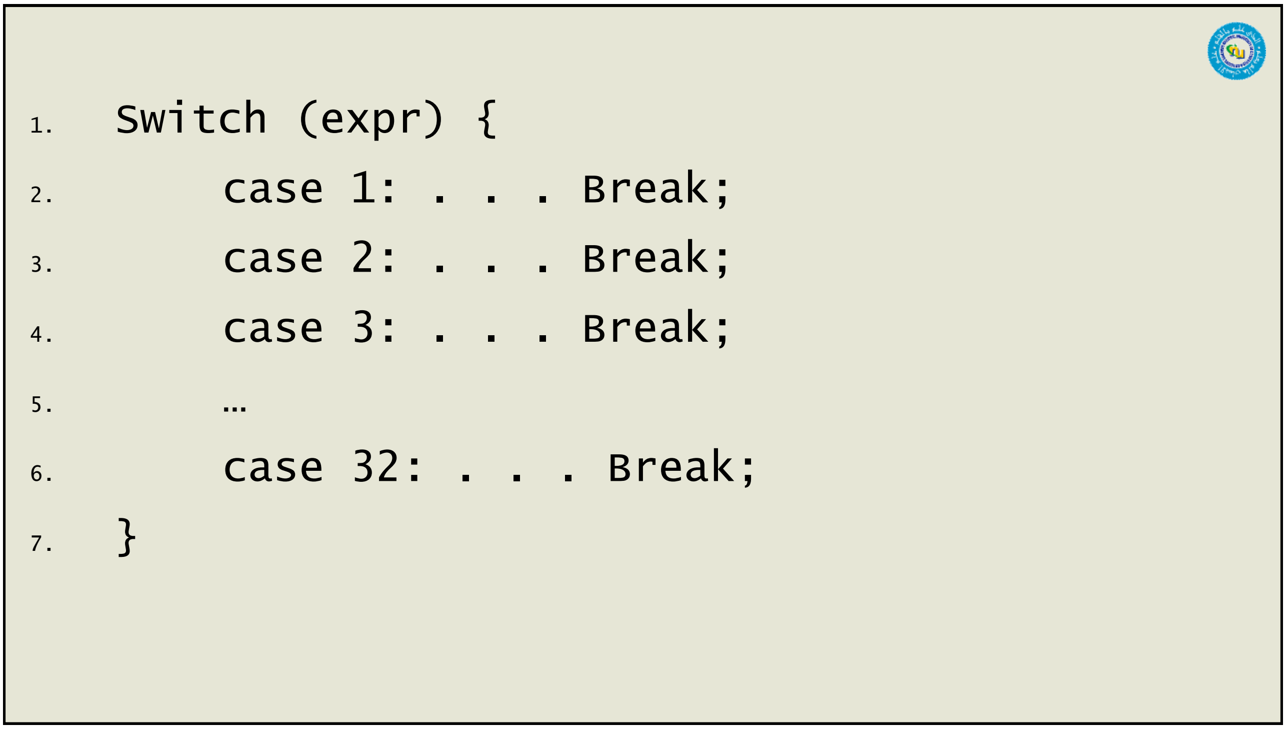


What is the max branch divergence penalty for a cuda thread block with 1024 threads?

32 x slowdown

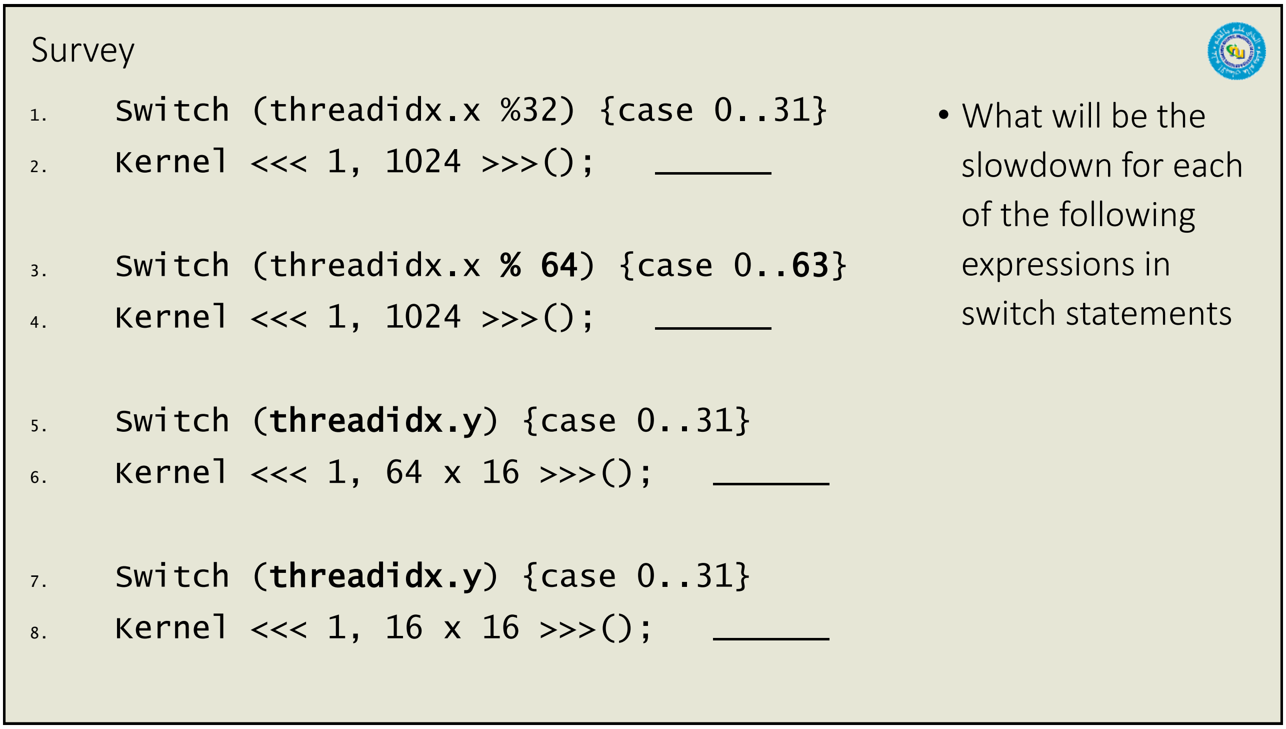
Max 32-way branch divergence (warp size)

7



```
Switch (expr) {
      case 1: . . . Break;
      case 2: . . . Break;
      case 3: . . . Break;
      …
      case 32: . . . Break;
}
```

8

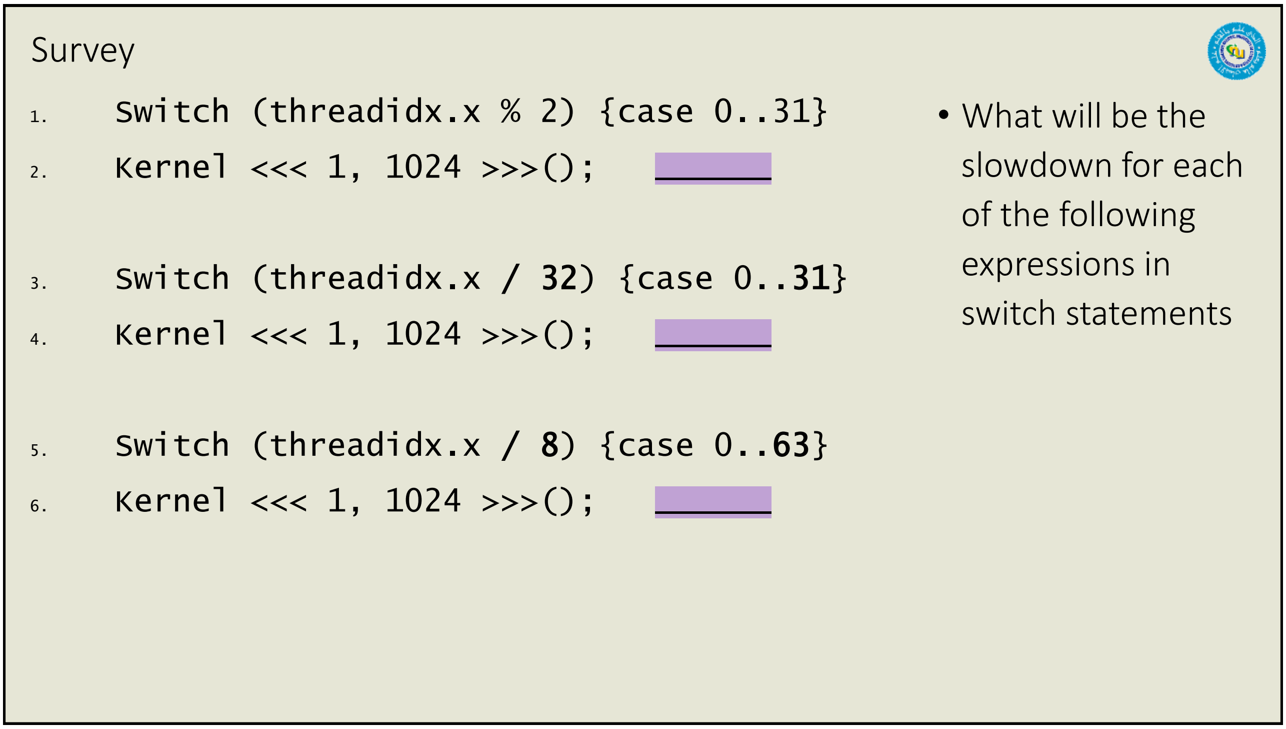


Survey

```
1.  Switch (threadidx.x %32) {case 0..31}
2.  Kernel <<< 1, 1024 >>>();     ______

3.  Switch (threadidx.x % 64) {case 0..63}
4.  Kernel <<< 1, 1024 >>>();     ______

5.  Switch (threadidx.y) {case 0..31}
6.  Kernel <<< 1, 64 x 16 >>>();     ______

7.  Switch (threadidx.y) {case 0..31}
8.  Kernel <<< 1, 16 x 16 >>>();     ______
```

- What will be the slowdown for each of the following expressions in switch statements

9



Survey

```
1.  Switch (threadidx.x % 2) {case 0..31}
2.  Kernel <<< 1, 1024 >>>();     ______

3.  Switch (threadidx.x / 32) {case 0..31}
4.  Kernel <<< 1, 1024 >>>();     ______

5.  Switch (threadidx.x / 8) {case 0..63}
6.  Kernel <<< 1, 1024 >>>();     ______
```

- What will be the slowdown for each of the following expressions in switch statements

10

Branch Divergence in real world

- Assume a 1024 x 1024 image.
- Requiring Special handling of pixels on the boundary

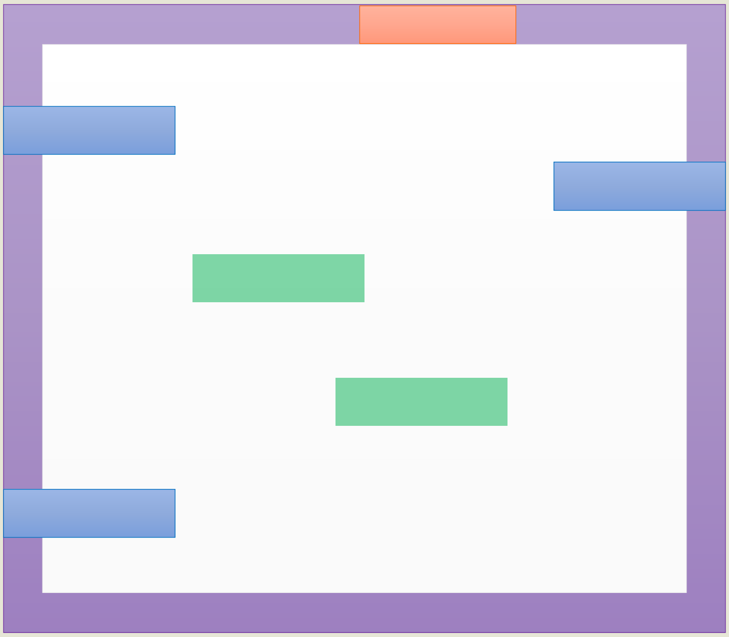
11

Maximum branch divergence of any warp? __________ - way

```
__global__
If(threadIdx.x == 0 ||
   threadxIdx.x == 1024
   threadIdx.y==0 ||
   threadIdx.y==1024){
  //deal with boundary cond.
}else {
  //do smth
}
```

12

Branch divergence in convolution



13

Branch divergence in real life

- Be aware of branch divergence
- Don’t panic if there are if statements
- No real strategy in reducing branch divergence

14

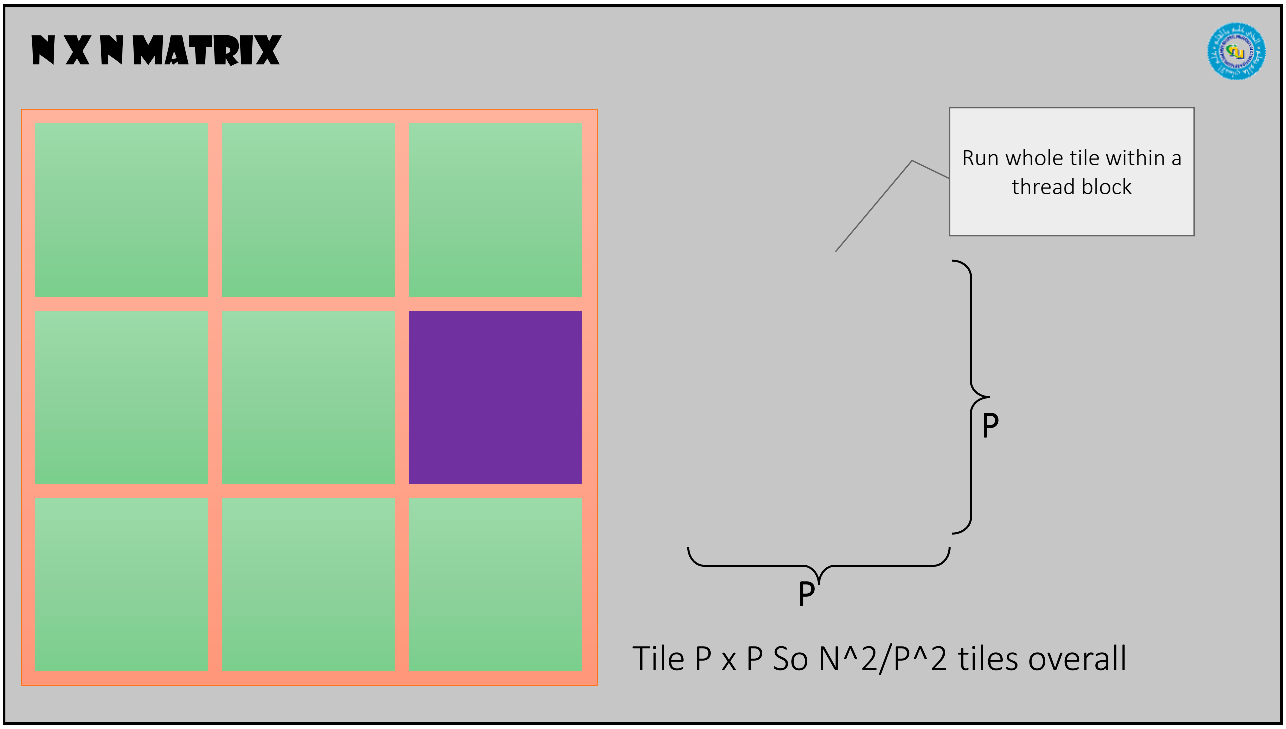
Major guidelines

- Avoid code with too many branches
- Be aware of large imbalance in thread workloads
  - For loops with variable terminating statements.

15

TILED MATRIX OPERATIONS

16

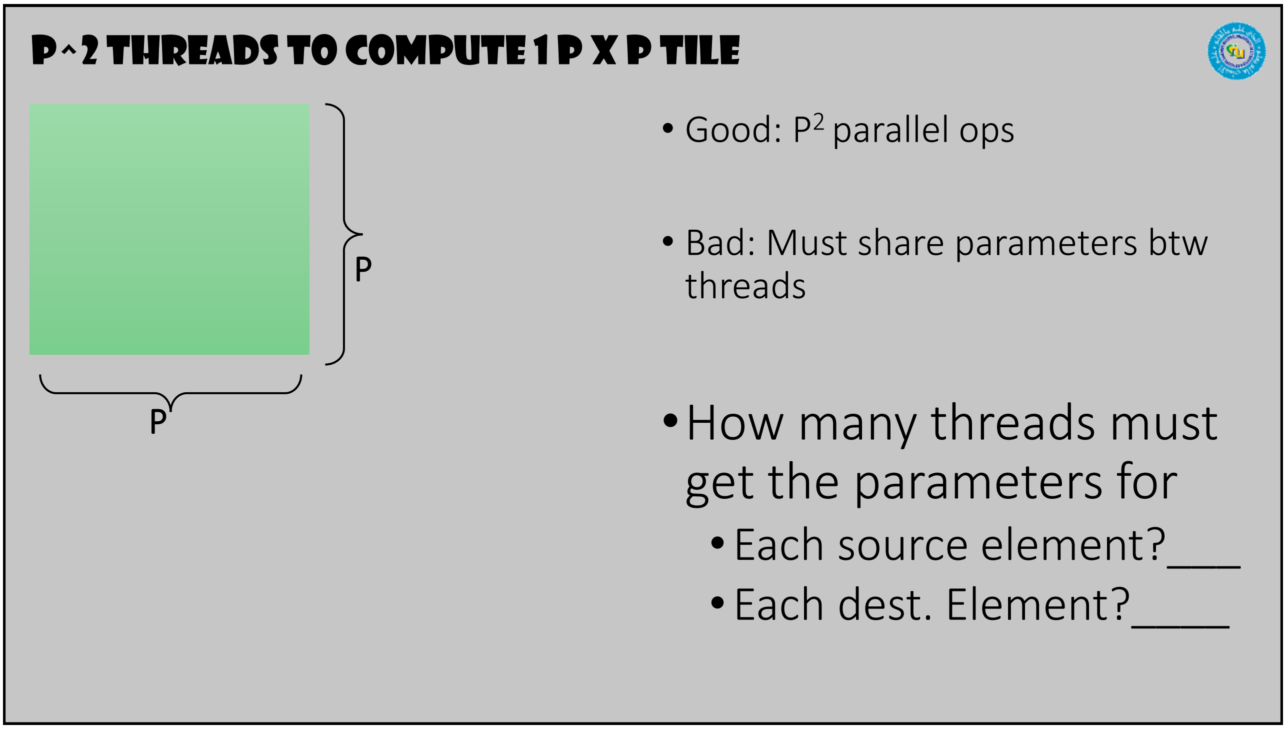


N X N MATRIX

Run whole tile within a thread block

Tile P x P So N^2/P^2 tiles overall

17

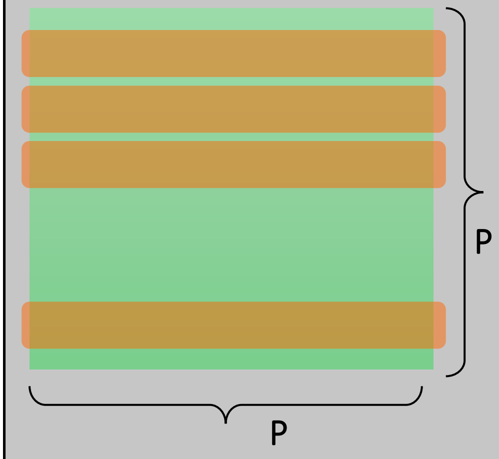


P^2 THREADS TO COMPUTE 1 P X P TILE

- Good: P^2 parallel ops
- Bad: Must share parameters btw threads
- How many threads must get the parameters for
  - Each source element?____
  - Each dest. Element?_____

18

P threads to compute 1 P x P tile



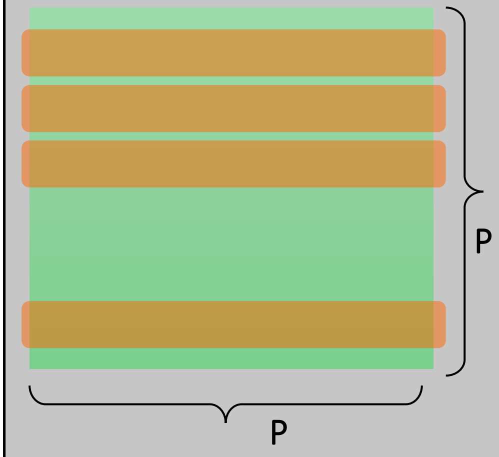
Is parallelization

increased? _____

decreased?_____

No change?_____

P threads to compute 1 P x P tile



- fewer threads
- More work per thread

- Communication
  - among threads vs within a thread