center for
excellence in parallel
programming

CUDA

Georges-Emmanuel Moulard
Paul Karlshöfer

Bull
atos technologies

Warp-level operations

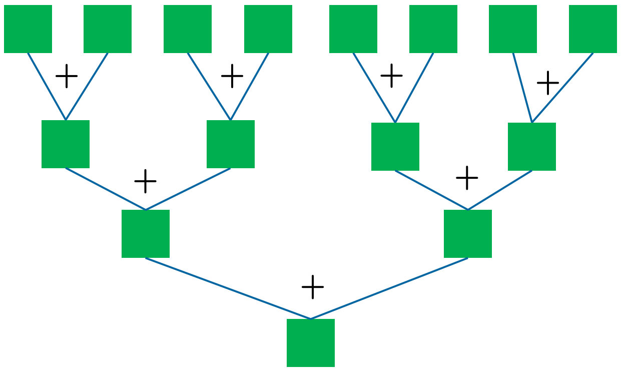
05/08/2020

Recap: Communication between Threads

Communication	Threads in distinct threadblocks	Threads in same threadblock	Threads in same warp
Level of synchronization	Grid-level	Block-level	Warp-level
Reflect in code	Distinct kernel launches	__syncthreads()	warp-level primitives

Warp-level operations

- CUDA provides some operations for data exchange, synchronization and querying the execution flow on the warp level.
- Warp-level operations are important, if data dependencies between threads of a warp exists.
- Example: Reducing 32 values (one per thread) held in registers



Warp-level operations

- Example:

```
T __shfl_sync(unsigned mask, T var, int srcLane, int width=warpSize);
```

- T can be any primitive numeric type (double included)
- Different shuffle operations exist. (e.g. shuffle_down, shuffle_xor)

- Since CUDA 9.0, all warp-level functions take a mask and are suffixed "sync"
  - Non sync functions are deprecated and especially on cc 7.x not valid!

- The shuffle functions allow exchange a variable (4 or 8 byte) between threads within a warp without having to pass by shared memory

Warp-level operations

- Threads within a warp a called **lanes**. (index from `0 .. (warpsize – 1)`)

- The **mask** specifies which threads are participating in the call
  - If a lane is not marked in the mask, but participating, the result is undefined

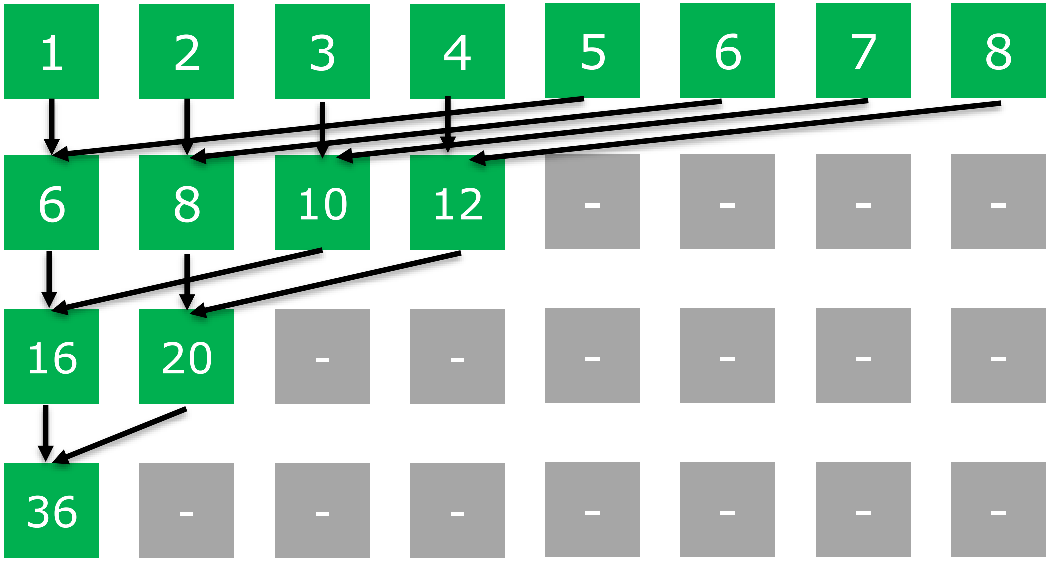
```
#define FULL_MASK 0xffffffff
__global__ void warp_broadcast(int *ret){
     int val = 0;
     if(threadIdx.x == 0)
          val = 42;
     val = __shfl_sync(FULL_MASK, val, 0, 32);

     if(threadIdx.x == 11)
          printf(“%d \n”, val);   //prints 42
}
```

Reduction (warp-level)

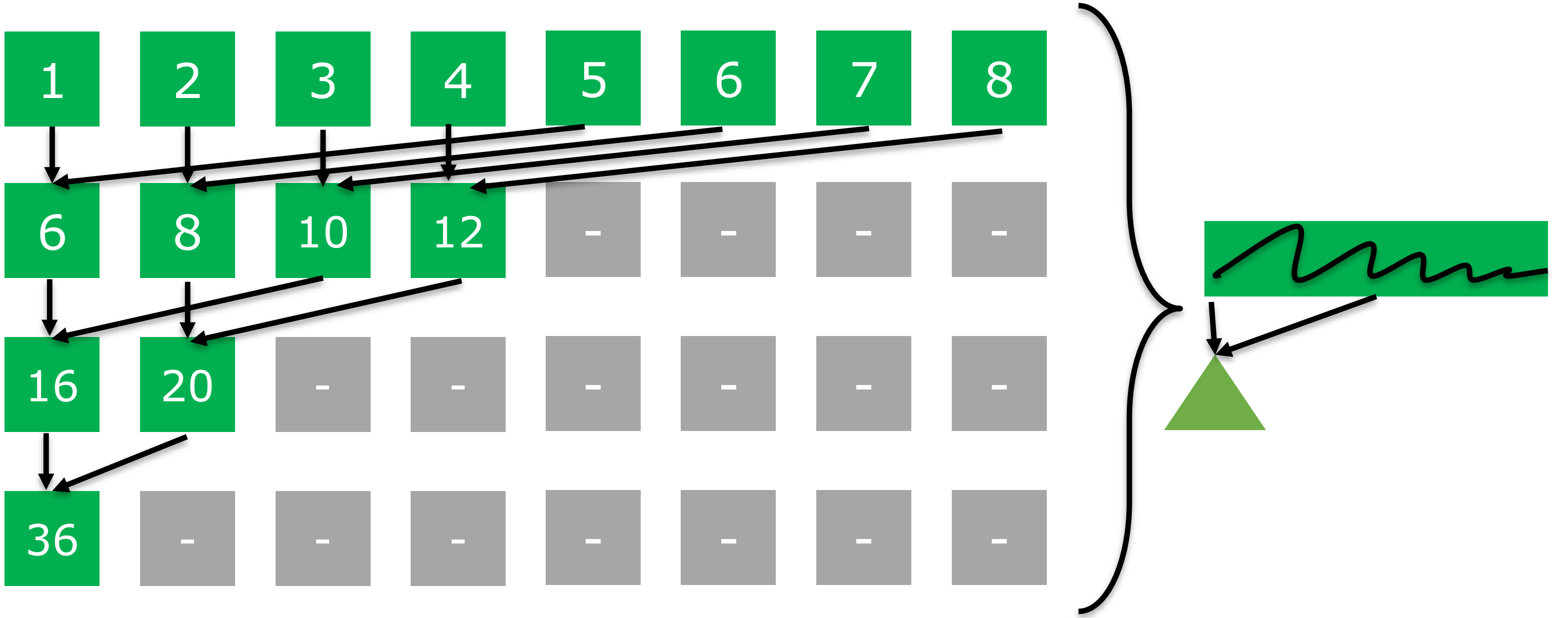
```
#define FULL_MASK 0xffffffff
__inline__ __device__
void warp_reduction(int *val) {
     for (int offset = WARPSIZE/2; offset > 0; offset /= 2)
          *val += __shfl_down_sync(FULL_MASK, *val, offset);
}
```

`__inline__` similar to C/C++ `inline`. A hint to the compiler.



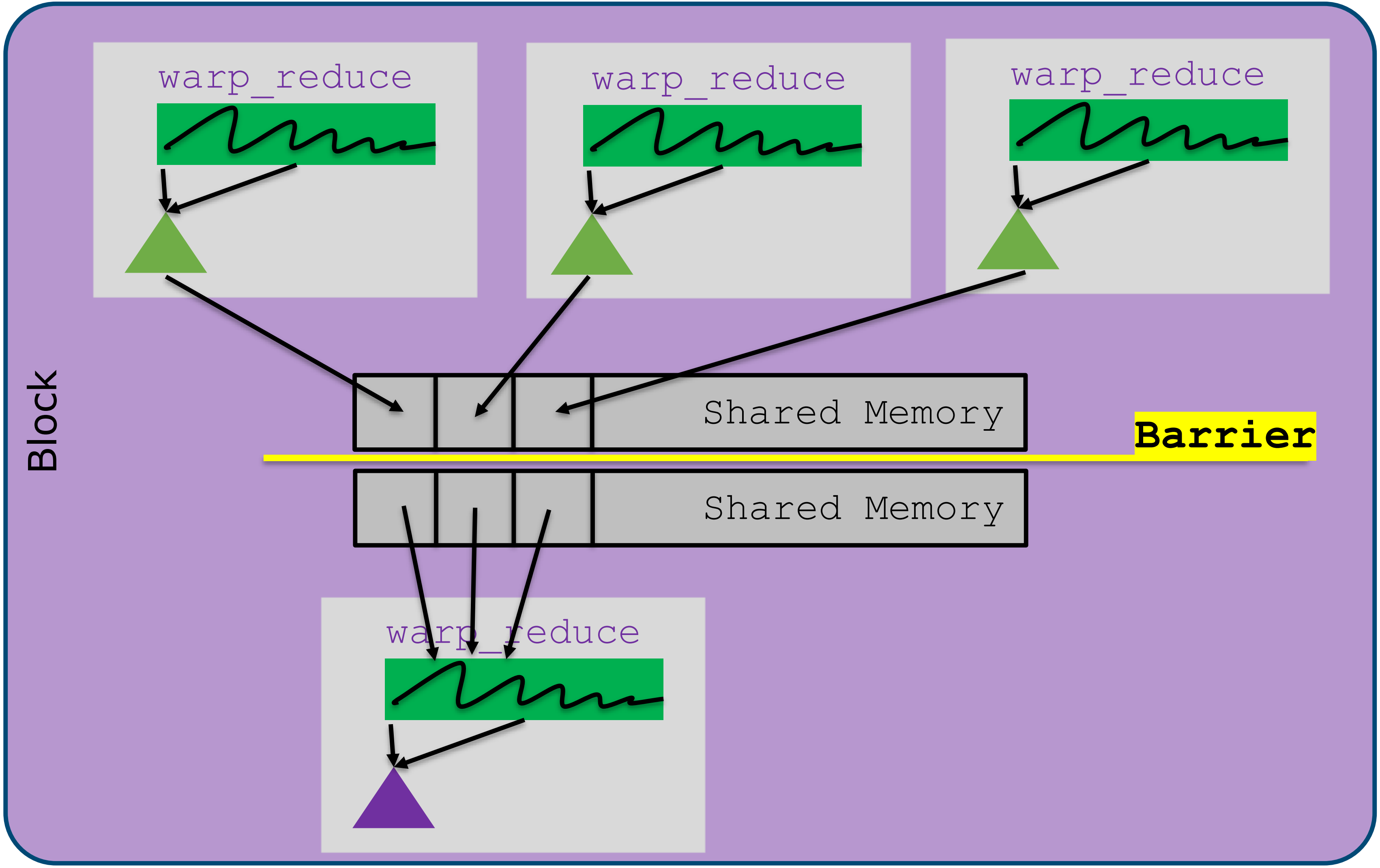
Reduction (block-level)

- Idea



Reduction (block-level)

- Idea:



Copyright

Thanks

For more information please contact:
Paul Karlshöfer
paul.karlshoefer@atos.net

29-10-2018