HPC.NRW

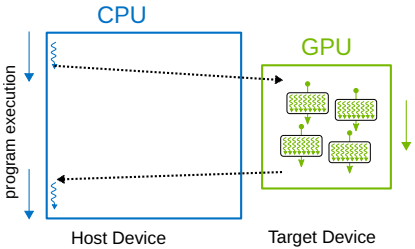
SEVERAL WAYS TO SAXPY

OpenMP GPU Offloading

Marius Neumann, Xin Wu

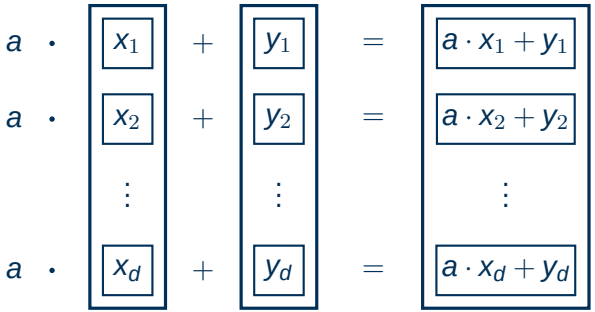
WHAT IS OPENMP FOR GPU OFFLOADING?

- directive-based parallel programming interface
- OpenMP 4.0 and later: offload computations to GPU
- open source and commercial compilers:



THE SAXPY PROBLEM

- linear combination of two float arrays
- results written to third array



Typical C/C++ code:

```
int main(){
    int N=6;
    float a=3.1415;
    float x[N]={1,2,3,4,5,6};
    float y[N]={7,8,9,0,1,2};
    float z[N];

    for(int i=0; i<N; i++){
        z[i]=a*x[i]+y[i];
    }
}
```

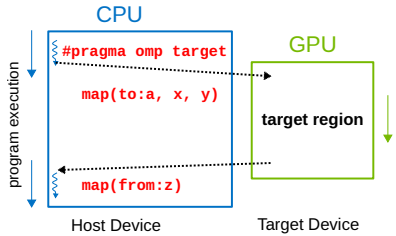
SAXPY WITH OPENMP GPU OFFLOADING

HPC.NRW

```
// OpenMP header file
#include <omp.h>

#pragma omp target \
    map(to:a, x, y) map(from:z)
#pragma omp teams default(shared)
#pragma omp distribute parallel for
    for (int i = 0; i < N; ++i) {
        z[i] = a * x[i] + y[i];
    }
```

- `#pragma omp target` create target region on GPU
- `map(to:...)`: maps variables to device data environment *before* execution
- `map(from:...)`: maps variables to host data environment *after* execution



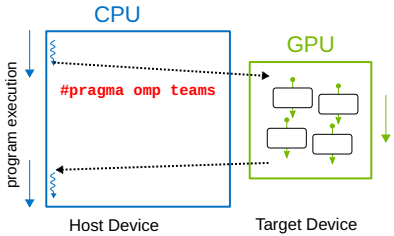
SAXPY WITH OPENMP GPU OFFLOADING

HPC.NRW

```
// OpenMP header file
#include <omp.h>

#pragma omp target \
    map(to:a, x, y) map(from:z)
#pragma omp teams default(shared)
#pragma omp distribute parallel for
    for (int i = 0; i < N; ++i) {
        z[i] = a * x[i] + y[i];
    }
```

- `#pragma omp teams` initialize a league of teams for execution on GPU
- `default(shared)`: variables are implicitly shared amongst the teams



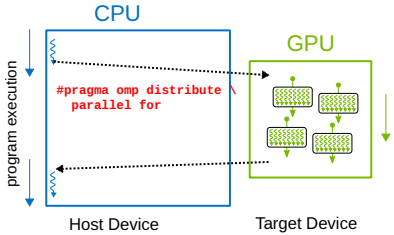
SAXPY WITH OPENMP GPU OFFLOADING

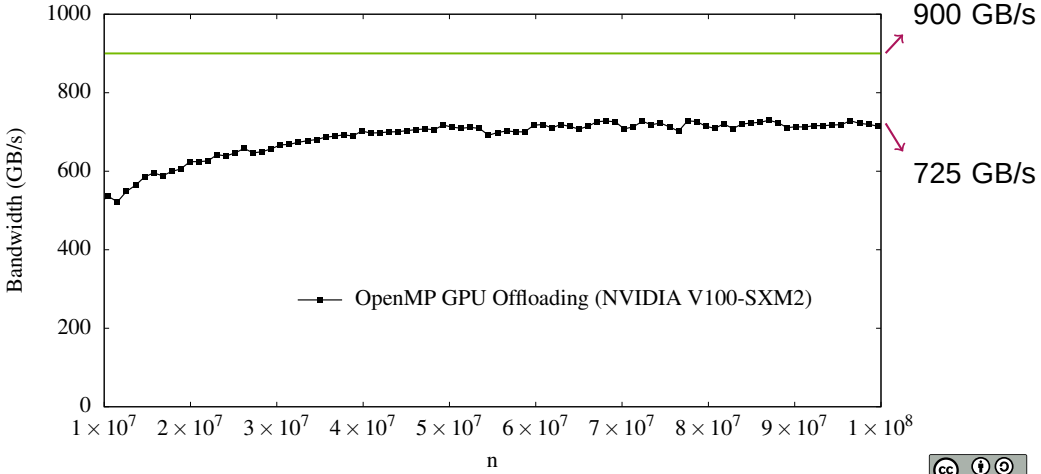
HPC.NRW

```
// OpenMP header file
#include <omp.h>

#pragma omp target \
    map(to:a, x, y) map(from:z)
#pragma omp teams default(shared)
#pragma omp distribute parallel for
    for (int i = 0; i < N; ++i) {
        z[i] = a * x[i] + y[i];
    }
```

– `#pragma omp distribute parallel for`
  1. all iterations of the `for`-loop are distributed into chunks for the teams
  2. the iterations in each chunk are then distributed across the threads within the teams




1000
800
600
400
200
0
Bandwidth (GB/s)
OpenMP GPU Offloading (NVIDIA V100-SXM2)
1×10^7
2×10^7
3×10^7
4×10^7
5×10^7
6×10^7
7×10^7
8×10^7
9×10^7
1×10^8
n
900 GB/s
725 GB/s

- SAXPY kernel can be implemented by using OpenMP GPU offloading.
  - ⇒ create target region on GPU and map variables:
    `#pragma omp target map(...)`
  - ⇒ initialize a league of teams for execution:
    `#pragma omp teams`
  - ⇒ distribute iterations across GPU threads in the teams:
    `#pragma omp distribute parallel for`
- Memory bandwidth performance
  - ⇒ 725 GB/s on NVIDIA V100
  - ⇒ 80% of theoretical bandwidth