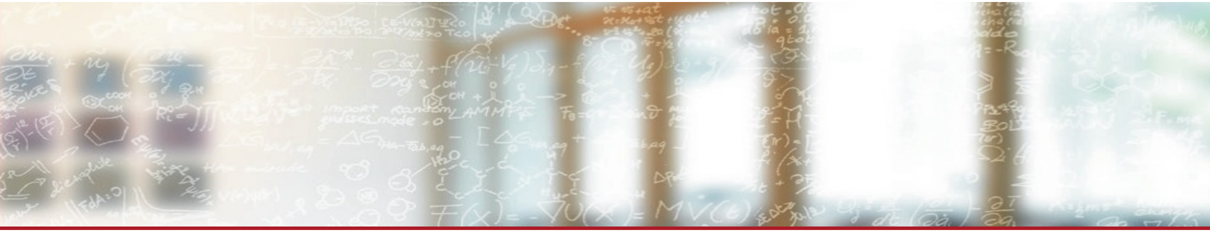
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Introduction to OpenACC

Directive Based GPU Programming

Vasileios Karakasis, CSCS

May 14–15, 2018

Overview

- Day 1 (Understand the basic concepts of OpenACC)
  - Execution and memory model
  - Basic directives
  - Profiling and debugging
  - Hands-on sessions

- Day 2 (Advanced topics)
  - Asynchronous execution and wait queues
  - Interoperability with CUDA and MPI
  - Deep copy
  - Hands-on sessions

What is OpenACC?

- Collection of compiler directives for specifying loops and regions to be offloaded from a host CPU to an attached accelerator device
- Host + Accelerator programming model
- High-level representation
- Current specification version: 2.6

Why to use OpenACC?

Because . . .

- I don’t care about all the little hardware details, I want my science done.
- I want to run on accelerators, but I still need a fast and readable code.
- I need portability across different accelerator vendors, but also be able to run on the multicore.
- I inhereted a large legacy monolithic codebase, which I don’t dare to refactor completely, but I need to get my results faster.
- My code is in Fortran.

OpenACC is not a silver bullet

- A high-level representation is not a panacea.
  - You still need to understand and adapt to the programming model.
- Does not substitute hand-tuning, but can serve as a very good starting point.
- User base not yet as large as of classic OpenMP for multicores, but it is expanding.
  - You may run into compiler bugs or specification ambiguities.

Format of directives

- C/C++
  - `#pragma acc` *directive-name [clause-list] new-line*
  - Scope is the following *block of code*
- Fortran
  - `!$acc` *directive-name [clause-list] new-line*
  - Scope is until `!$acc end` *directive-name*

Programming model

- Host-directed execution
- Compute intensive regions are offloaded to attached accelerator devices
- Host orchestrates the execution on the device
  - Allocations on the device
  - Data transfers
  - Kernel launches
  - Wait for events
  - Etc. . .

Execution model

- The device executes *parallel* or *kernel regions*
- Parallel region
  - Work-sharing loops
- Kernel region
  - Multiple loops to be executed as multiple kernels
- Levels of parallelim
  1. *Gang*
  2. *Worker*
  3. *Vector*

  - Parallelism levels are decided by the compiler but can be fine-tuned by the user

Execution model

- The device executes *parallel* or *kernel regions*
- Parallel region
  - Work-sharing loops
- Kernel region
  - Multiple loops to be executed as multiple kernels
- Levels of parallelim
  1. *Gang* → *CUDA block*
  2. *Worker* → *CUDA warp*
  3. *Vector* → *CUDA threads*

  - Parallelism levels are decided by the compiler but can be fine-tuned by the user
  - Mapping to CUDA blocks/warps/threads is implementation defined

Execution model

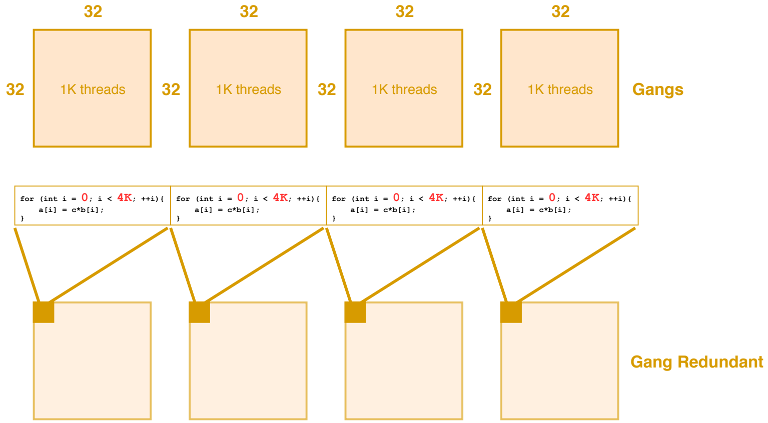
Modes of execution

- Gang
  - Gang-redundant (GR)
  - Gang-partioned (GP)
- Worker
  - Worker-single (WS)
  - Worker-partitioned (WP)
- Vector
  - Vector-single (VS)
  - Vector-partitioned (VP)

Execution model

Modes of execution

```
for (int i = 0; i < 4096; ++i) {
    a[i] = c*b[i];
}
```



```
for (int i = 0; i < 4K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 0; i < 4K; ++i){
    a[i] = c*b[i];
}
```

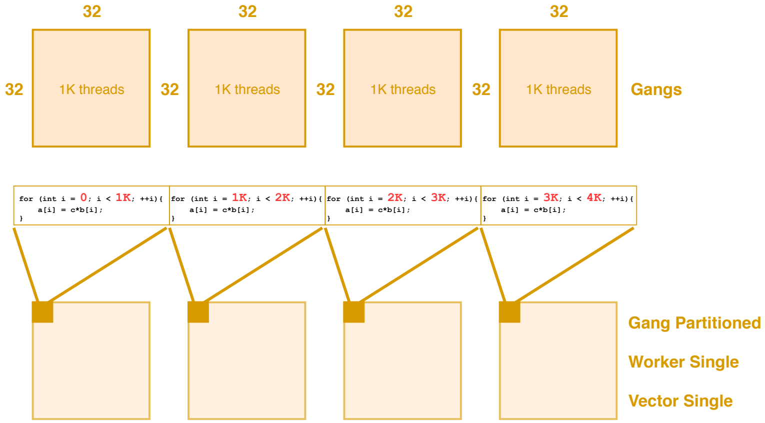
```
for (int i = 0; i < 4K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 0; i < 4K; ++i){
    a[i] = c*b[i];
}
```

Execution model

Modes of execution

```
for (int i = 0; i < 4096; ++i) {
    a[i] = c*b[i];
}
```



```
for (int i = 0; i < 1K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 1K; i < 2K; ++i){
    a[i] = c*b[i];
}
```

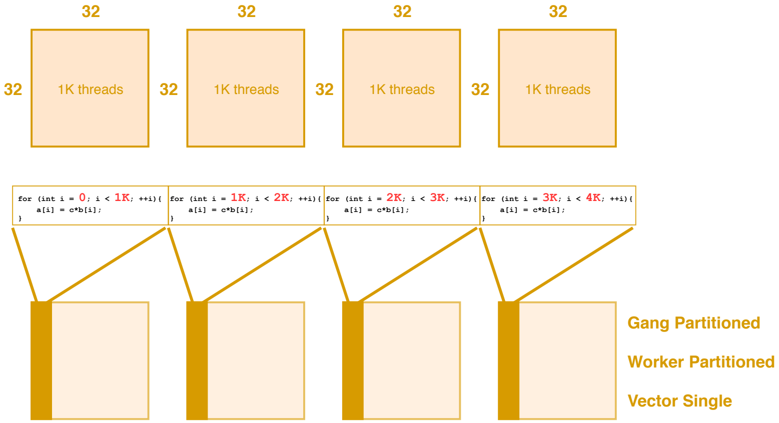
```
for (int i = 2K; i < 3K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 3K; i < 4K; ++i){
    a[i] = c*b[i];
}
```

Execution model

Modes of execution

```
for (int i = 0; i < 4096; ++i) {
    a[i] = c*b[i];
}
```



```
for (int i = 0; i < 1K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 1K; i < 2K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 2K; i < 3K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 3K; i < 4K; ++i){
    a[i] = c*b[i];
}
```

Gang Partitioned

Worker Partitioned

Vector Single

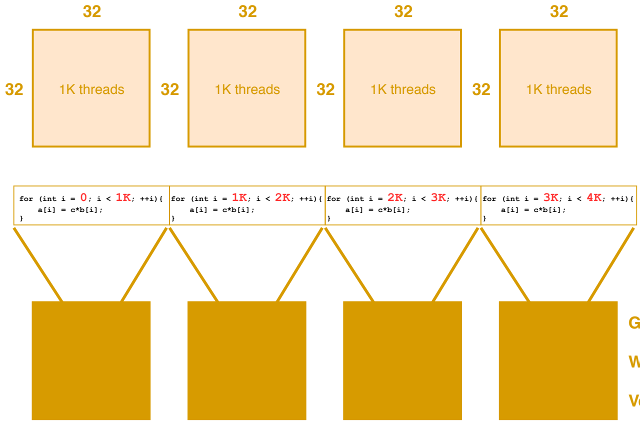
CSCS

ETH zürich

Execution model

Modes of execution

```
for (int i = 0; i < 4096; ++i) {
    a[i] = c*b[i];
}
```



```
for (int i = 0; i < 1K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 1K; i < 2K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 2K; i < 3K; ++i){
    a[i] = c*b[i];
}
```

```
for (int i = 3K; i < 4K; ++i){
    a[i] = c*b[i];
}
```

Gang Partitioned

Worker Partitioned

Vector Partitioned

Execution model

The `kernels` construct

Multiple loops inside kernels construct

```
!$acc kernels
    !GR mode
    do i = 1, N
        !compiler decides on the partitioning (GP/WP/VP modes)
        y(i) = y(i) + a*x(i)
    enddo
    do i = 1, N
        !compiler decides on the partitioning (GP/WP/VP modes)
        y(i) = b*y(i) + a*x(i)
    enddo
!$acc end kernels
```

- Compiler will try to deduce parallelism
- Loops are launched as different GPU kernels

Execution model

The `parallel` construct

Parallel construct

```
!$acc parallel
    do i = 1, N
        ! loop executed in GR mode
        y(i) = y(i) + a*x(i)
    enddo
    !$acc loop
    do i = 1, N
        !compiler decides on the partitioning (GP/WP/VP modes)
        y(i) = b*y(i) + a*x(i)
    enddo
!$acc end parallel
```

- No automatic parallelism deduction → parallel loops must be specified explicitly
- Implicit gang barrier at the end of `parallel`

Execution model

Work-sharing loops

- C/C++: `#pragma acc loop`
  - Applies to the immediately following `for` loop
- Fortran: `!$acc loop`
  - Applies to the immediately following `do` loop
- Loop will be automatically striped and assigned to different threads
  - Use the `independent` clause to force striping
- Convenience syntax combines `parallel`/`kernels` and `loop` constructs
  - `#pragma acc parallel loop`
  - `#pragma acc kernels loop`
  - `!$acc parallel loop`
  - `!$acc kernels loop`

Execution model

Work-sharing loops – the `collapse` clause

Collapse loops

```
!$acc loop collapse(2)
do i = 1,N
    do j = 1,N
        A(i,j) = coeff*B(i,j)
    enddo
enddo
```

- OpenACC vs. OpenMP
  - OpenACC: apply the `loop` directive to the following N loops and possibly collapse their iteration spaces if independent
  - OpenMP: Collapse the iteration spaces of the following N loops

Execution model

Controlling parallelism

- Amount of parallelism at the `kernels` and `parallel` level
  - `num_gangs(...)`, `num_workers(...)`, `vector_length(...)`
- At the `loop` level
  - `gang`, `worker`, `vector`

100 thread blocks with 128 threads each

```
!$acc parallel num_gangs(100), vector_length(128)
    !$acc loop gang, vector
    do i = 1, n
        y(i) = y(i) + a*x(i)
    enddo
!$acc end parallel
```

Execution model

Variable scoping

- Allowed in the `parallel` directive only
- By default, if outside of a code block, variables are shared in global memory
- `private`: A copy of the variable is placed in each *gang* (CUDA block)
- `firstprivate`: Same as `private` but initialized from the host value

Execution model

Variable scoping

- Allowed in the `parallel` directive only
- By default, if outside of a code block, variables are shared in global memory
- `private`: A copy of the variable is placed in each *gang* (CUDA block)
- `firstprivate`: Same as `private` but initialized from the host value

Implicit scoping:

- (C/C++/Fortran) Loop variables are private to the *thread* that executes the loop
- (C/C++ only) Scope of variables declared inside a parallel block depends on the current execution mode:
  - *Vector-partitioned* mode → private to the thread
  - *Worker-partitioned, Vector-single* mode → private to the worker
  - *Worker-single* mode → private to the gang

Execution model

Reduction operations

- `#pragma acc parallel reduction(<op>:<var>)`
  - e.g., `#pragma acc parallel reduction(+:sum)`
- `#pragma acc loop reduction(<op>:<var>)`
- `var` must be scalar
- `var` is copied and default initialized within each gang
- Intermediate results from each gang are combined and made available outside the parallel region
- Complex numbers are also supported
- Operators: `+`, `*`, `max`, `min`, `&`, `|`, `%`, `&&`, `||`

Execution model

Calling functions from parallel regions

- `#pragma acc routine {gang | worker | vector | seq}`
  - Just before the function declaration or definition
- `!$acc routine {gang | worker | vector | seq}`
  - In the specification part of the subroutine
- Parallelism level of the routine
  - `gang`: must be called from GR context
  - `worker`: must be called from WS context
  - `vector`: must be called from VS context
  - `seq`: must be called from sequential context

Memory model

Where is my data?

- The host and the device have separate address spaces
  - Data management between the host and the device is the programmer’s responsibility
  - You must make sure that all the necessary data for a computation is available on the accelerator before entering the compute region
  - You must make sure to transfer the processed data back to the host if needed

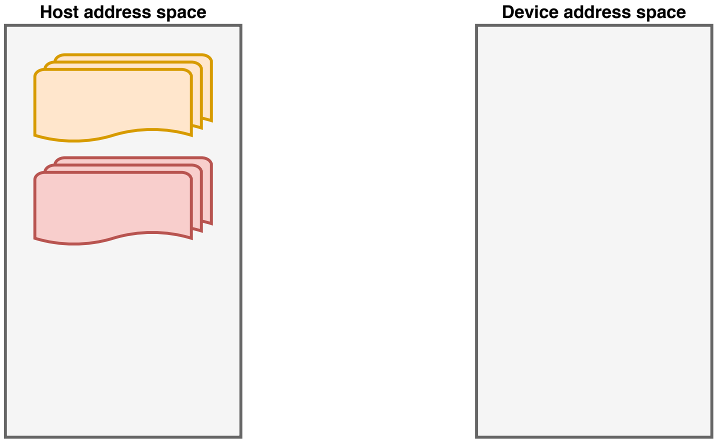
Memory model

Where is my data?

- The host and the device have separate address spaces
  - Data management between the host and the device is the programmer's responsibility
  - You must make sure that all the necessary data for a computation is available on the accelerator before entering the compute region
  - You must make sure to transfer the processed data back to the host if needed
- But there can be some exceptions:
  - The "device" might be the multicore → no need for data management
  - Some compilers may infer automatically the necessary data transfers
  - Nvidia Pascal GPUs provide efficient support for a unified memory view between the host and the accelerator

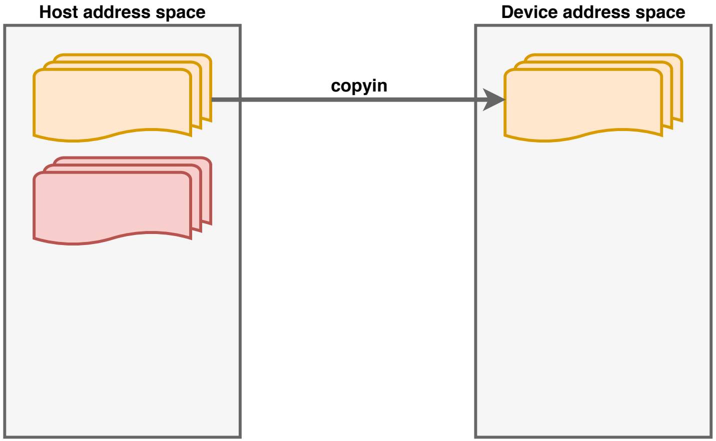
Memory model

Separate address spaces



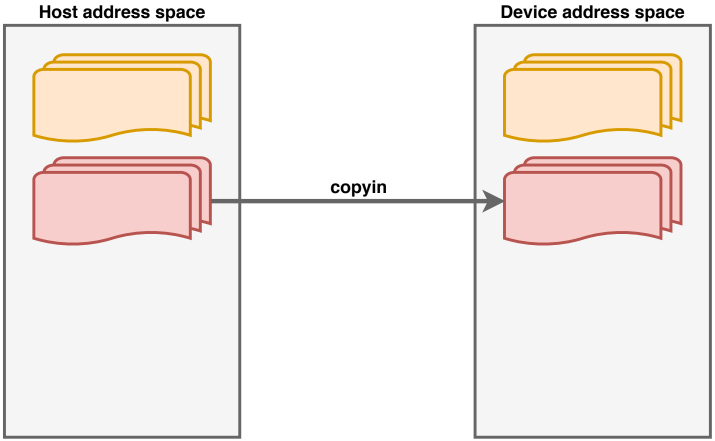
Memory model

Separate address spaces



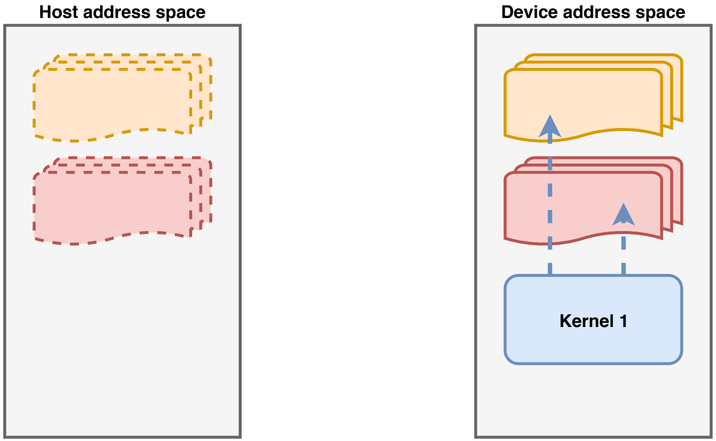
Memory model

Separate address spaces



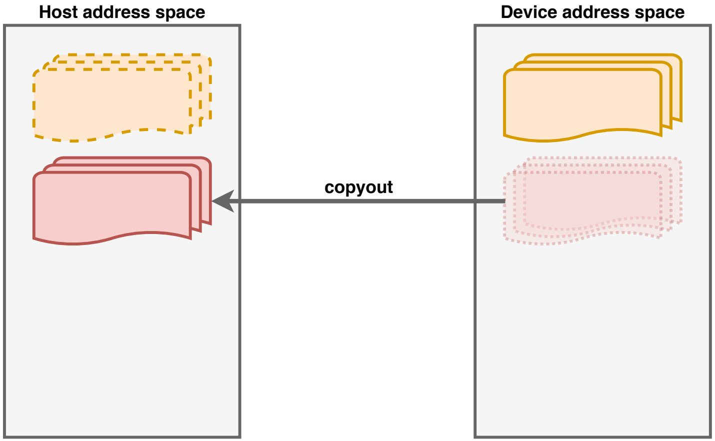
Memory model

Separate address spaces



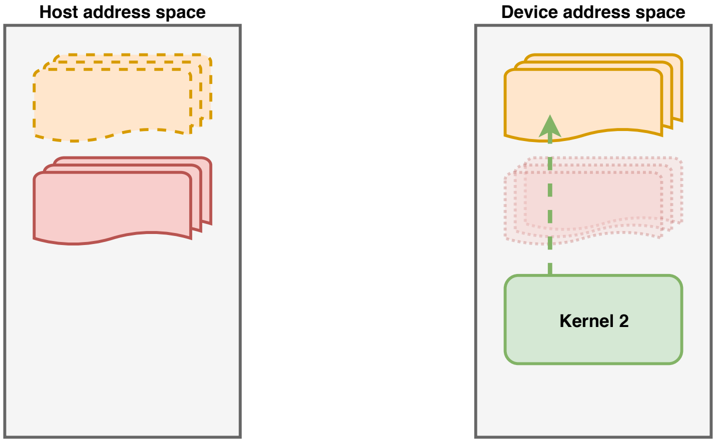
Memory model

Separate address spaces



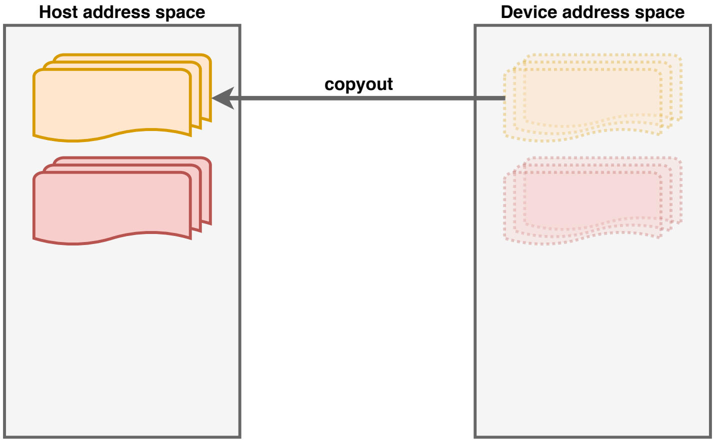
Memory model

Separate address spaces



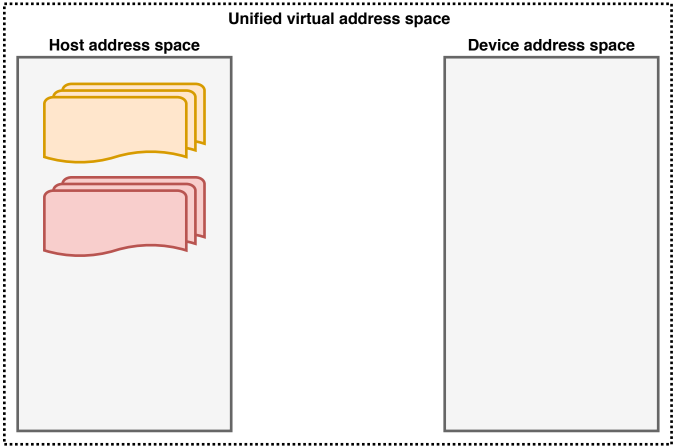
Memory model

Separate address spaces



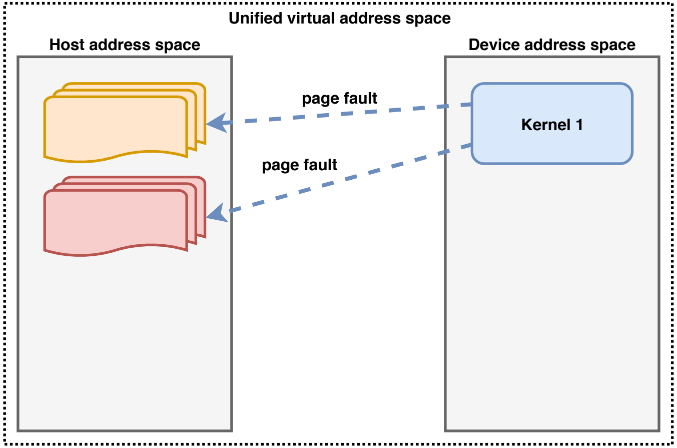
Memory model

Unified memory address space



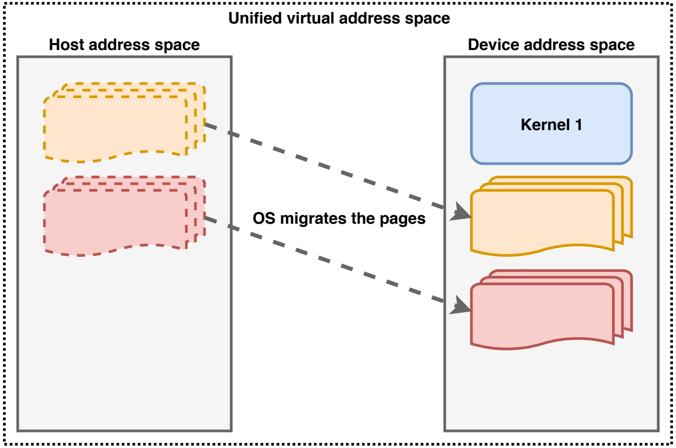
Memory model

Unified memory address space



Memory model

Unified memory address space



Memory model

Directives accepting data clauses

Data clauses may appear in the following directives:

- Compute directives:
  - `#pragma acc kernels`
  - `#pragma acc parallel`
- Data directives:
  - `#pragma acc data`
  - `#pragma acc enter data`
  - `#pragma acc exit data`
  - `#pragma acc declare`
  - `#pragma acc update`

Memory model

Data clauses

- `create(a[0:n])`: Allocate array `a` on device
- `copyin(a[0:n])`: Copy array `a` to device
- `copyout(a[0:n])`: Copy array `a` from device
- `copy(a[0:n])`: Copy array `a` to and from device
- `present(a)`: Inform OpenACC runtime that array `a` is on device
- `delete(a)`: Deallocate array `a` from device (`exit data` only)

Not for the `acc update` directive

Memory model

The `acc data` directive

- Defines a scoped data region
  - Data will be copied in at entry of the region and copied out at exit
  - A *structural reference count* is associated with each memory region that appears in the data clauses
- C/C++: `#pragma acc data` *[data clauses]*
  - The next block of code is a data region
- Fortran: `!$acc data` *[data clauses]*
  - Defines a data region until `!$acc end data` is encountered

Memory model

The `acc enter/exit data` directives

- Defines an unscoped data region
  - Data will be resident on the device until a corresponding `exit data` directive is found
  - Useful for managing data on the device across compilation units
  - A *dynamic reference count* is associated with each memory region that appears in the data clauses
- C/C++:
  - `#pragma acc enter data` *[data clauses]*
  - `#pragma acc exit data` *[data clauses]*
- Fortran:
  - `!$acc enter data` *[data clauses]*
  - `!$acc exit data` *[data clauses]*

Memory model

The `acc declare` directive

- Functions, subroutines and programs define *implicit data regions*
- The `acc declare` directive is used in variable declarations for making them available on the device during the lifetime of the implicit data region
- Useful for copying global variables to the device

- C/C++: `#pragma acc declare` *[data clauses]*
- Fortran: `!$acc declare` *[data clauses]*

Memory model

The `acc update` directive

- May be used during the lifetime of device data for updating the copies on either host or the device

- `#pragma acc update host(<var-list>)`
  - Update host copy with corresponding data from the device

- `#pragma acc update device(<var-list>)`
  - Update device copy with corresponding data from the host

Memory model

Array ranges

Data clauses may accept as arguments

- Whole arrays
  - C/C++: You *must* specify bounds for dynamically allocated arrays
    - `#pragma acc data copyin(a[0:n])`
    - But `#pragma acc data present(a)` is acceptable: `a`'s bounds can be inferred by the runtime
  - Fortran: array shape information is already embedded in the data type
    - `!$acc data copyin(a)`
- Array subranges
  - `#pragma acc data copyin(a[2:n-2])`

Synchronization directives

- Atomic operations
  - `#pragma acc atomic` [atomic-clause]
  - `!$acc atomic` [atomic-clause]
  - Atomic clauses: `read`, `write`, `update` and `capture`
  - Example of “capturing” a value:
    - `v = x++;`
- No global barriers → cannot be implemented due to hardware restrictions
- No equivalent of `__syncthreads()`

Leverage the unified memory

- Virtual address space shared between CPU and GPU
- The CUDA driver and the hardware take care of the page migration
- Introduced with the Kepler architecture and CUDA 6, but is significantly improved with Pascal

Leverage the unified memory

- Virtual address space shared between CPU and GPU
- The CUDA driver and the hardware take care of the page migration
- Introduced with the Kepler architecture and CUDA 6, but is significantly improved with Pascal

- You could completely omit the data management in OpenACC !
- Supported by the PGI compiler using the `-ta=tesla:managed` option

Hands-on exercises

General information

The initial course material is available on Github and it will be update during the course:

- `git clone https://github.com/vkarak/openacc-training.git`
- `git pull origin master` to get the latest version

Directory structure:

- `practicals/`: The hands-on exercises
- `scripts/`: Set up scripts to make your life easier
- `slides/`: Slides of the course
- `ci/`: Continuous integration tests for the exercises (ask me offline if interested)

Hands-on exercises

General information

- `grep TODO *.{cpp,f90,f03}`
- Both Cray/PGI compilers are supported, unless otherwise stated
  - Suggest using PGI for the advanced examples
- `source <repodir>/scripts/setup_pgi.sh` → will make available PGI 18.4
- `module load craype-accel-nvidia60` for loading CUDA and set the target architecture to the GPU
- `make`
  - For Cray compiler you may use `make VERBOSE=1` to get diagnostics information

Hands-on

Exercise 1 – AXPY

- `practicals/axpy/axpy_openacc.{cpp,f90}`
- Run as:

```
srun --reserv=openacc -Cgpu ./axpy.openacc [ARRAY_SIZE]
```

  - `ARRAY_SIZE` is power of 2, default is 16
- Try with different sizes. Does the GPU outperform the CPU version?

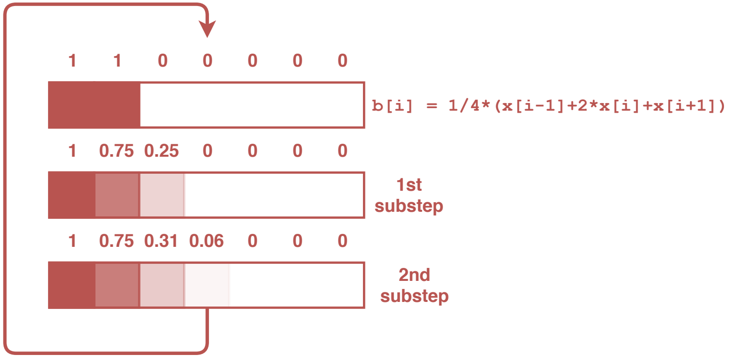
Hands-on

Exercise 2 – Dot product

- `practicals/basics/dot_openacc.{cpp,f90}`
- Run as: `srun --reserv=openacc -Cgpu ./dot.openacc [ARRAY_SIZE]`
  - `ARRAY_SIZE` is power of 2, default is 2
- Try with different sizes. Does the GPU outperform the CPU version?

Hands-on

Exercise 3 – 1D blur kernel



Hands-on

Exercise 3 – 1D blur kernel

- `practicals/basics/blur_openacc.{cpp,f90}`
- Run as:
  ```
  srun --reserv=openacc -Cgpu ./blur.openacc [ARRAY_SIZE]
  ```
  - `ARRAY_SIZE` is power of 2, default is 20
- Offload to GPU the loops of the naive kernel; why is it so slow?

Hands-on

Exercise 3 – 1D blur kernel

- `practicals/basics/blur_openacc.{cpp,f90}`
- Run as:
  `srun --reserv=openacc -Cgpu ./blur.openacc [ARRAY_SIZE]`
  - `ARRAY_SIZE` is power of 2, default is 20
- Offload to GPU the loops of the naive kernel; why is it so slow?

- Moving data to and from the device is slow (≈7–8 GB/s per direction)
- Avoid unnecessary data movement in the `nocopies` kernel
  - Move the necessary data to GPU early enough and keep it there as long as possible
  - Update host copies using `#pragma acc update` directive if needed

Hands-on

Exercise 4 – Experiment with the unified memory

- Remove all the data directives and data clauses
- Compile the `blur_twice_naive` kernel with `-Mcuda=managed`
- How does it compare to the manual data management in terms of performance?
- Can you explain the performance difference?

Interoperability with CUDA

- Can I use a CUDA pointer inside OpenACC context?
- Can I call a CUDA function from OpenACC context?

Short answer is *yes*.

Interoperability with CUDA

Use CUDA pointers inside OpenACC context

A scenario:

- Have a CUDA code that needs to call a function that uses OpenACC.
- This function may accept an array that has been allocated already on the GPU by CUDA.

The problem?

- OpenACC only knows of pointers that it is managing itself; the `present` clause won't work. No idea what this pointer is; never seen it before!

Interoperability with CUDA

Use CUDA pointers inside OpenACC context

Solution:

- We need to instruct the OpenACC runtime to trust this pointer and that it is a valid device pointer.
- OpenACC runtime will just treat that pointer as known, but it won't check its shape.
- Use the `deviceptr(<ptrlist>)` clause with `parallel`, `kernels` and `data` directives

Interoperability with CUDA

Use CUDA pointers inside OpenACC context – Example

```
void copy(double *dst, const double *src, size_t n) {
  #pragma acc parallel loop deviceptr(dst, src)
  for (size_t i = 0; i < n; ++i) {
    dst[i] = src[i];
  }
}

int main() {
  double *a, *b;
  cudaMalloc(&a, 1024);
  cudaMalloc(&b, 1024);
  ...
  copy(b, a, 1024);
  return 0;
}
```

Interoperability with CUDA

Call a CUDA function from OpenACC context

Scenario:

- My code is in OpenACC, but I need to call an optimized library written in CUDA, which accepts device pointers, e.g., cuBLAS.

Interoperability with CUDA

Call a CUDA function from OpenACC context

Scenario:

- My code is in OpenACC, but I need to call an optimized library written in CUDA, which accepts device pointers, e.g., cuBLAS.

Problem:

- I only “see” device pointers while in a parallel region, but I want to get a device pointer, while executing on the host.

Interoperability with CUDA

Call a CUDA function from OpenACC context

Scenario:

- My code is in OpenACC, but I need to call an optimized library written in CUDA, which accepts device pointers, e.g., cuBLAS.

Problem:

- I only “see” device pointers while in a parallel region, but I want to get a device pointer, while executing on the host.

Solution:

- Use a `host_data` region
  - `#pragma acc host_data use_device(<varlist>)`

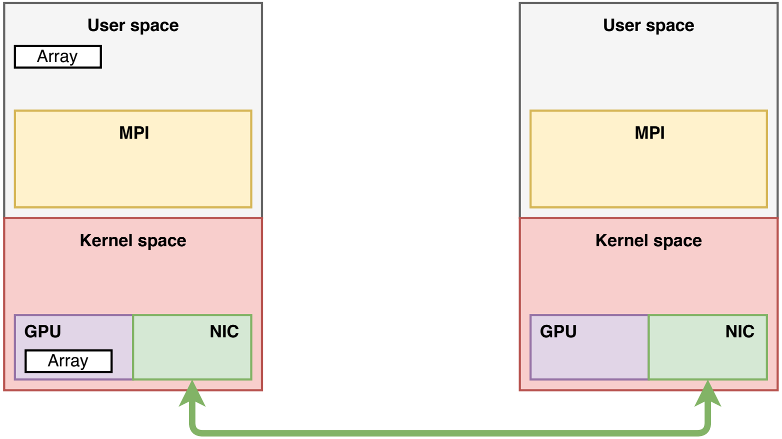
Interoperability with CUDA

The `host_data` directive

- C/C++: `#pragma acc host_data use_device(<varlist>)`
  - In the next block of code the compiler will make available the device address of any variable in `<varlist>`.
- Fortran: `!$acc host_data use_device(<varlist>)`
  - The compiler will make available the device address of any variable in `<varlist>` until a matching `!$acc end host_data` is found.
- Optional clauses:
  - `if(`*`condition`*`)`: Use the device pointer if *`condition`* is true.
  - `if_present`: Use the device pointer if variables in `<varlist>` are present on the device.

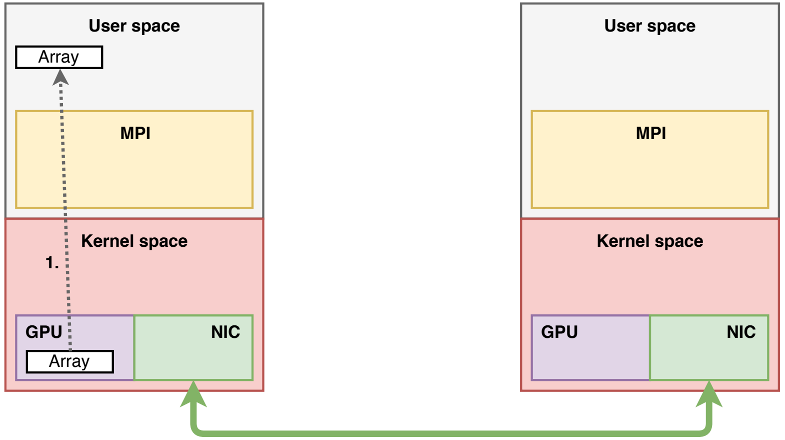
Interoperability with MPI

The communication data path



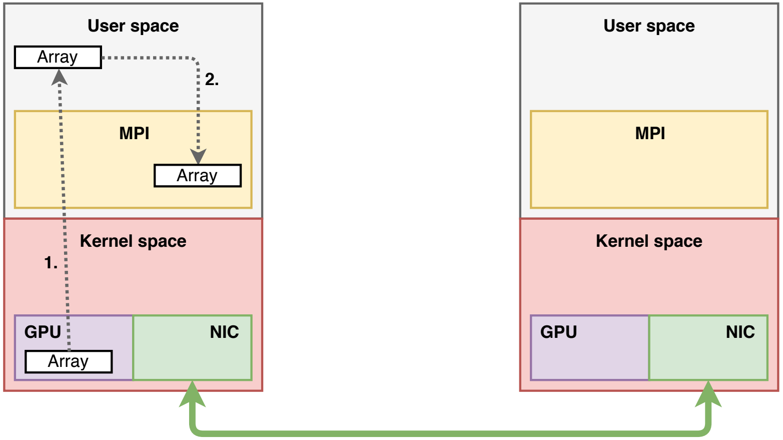
Interoperability with MPI

The communication data path



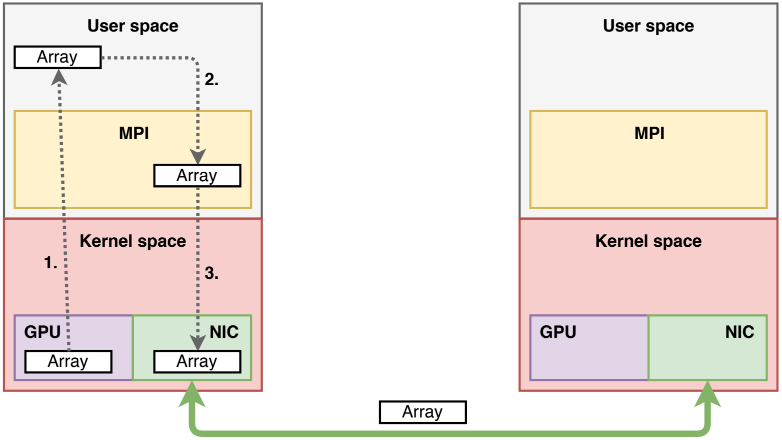
Interoperability with MPI

The communication data path



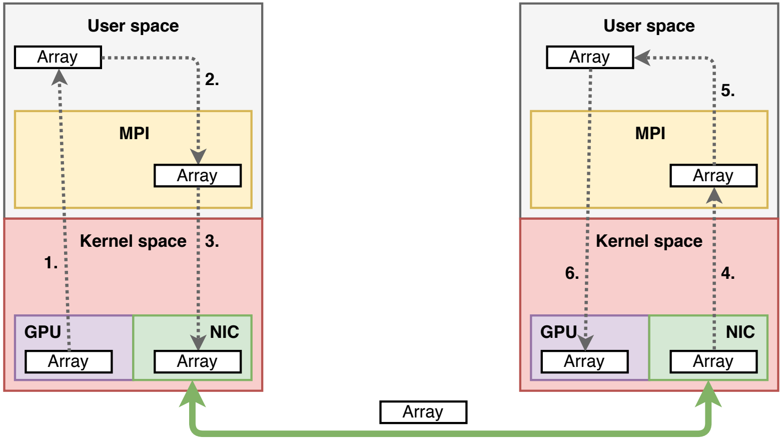
Interoperability with MPI

The communication data path



Interoperability with MPI

The communication data path



Interoperability with MPI

The communication data path

- Aren't there too many copies?
- This path is not fully unoptimized though! It still bypasses a copy by enabling RDMA between the NICs.

Interoperability with MPI

The communication data path

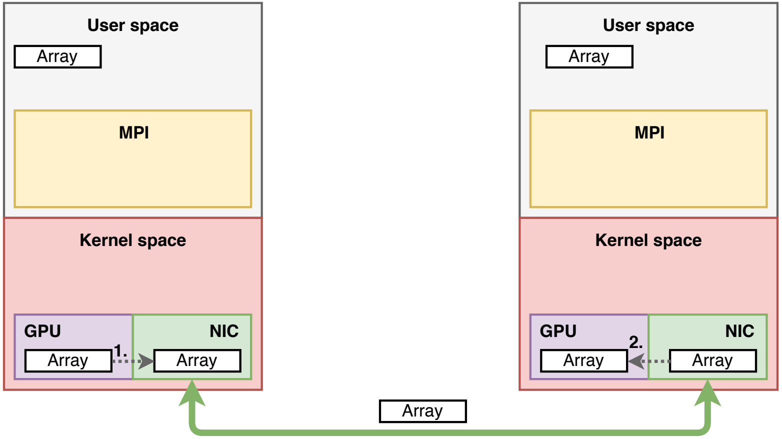
- Aren't there too many copies?
- This path is not fully unoptimized though! It still bypasses a copy by enabling RDMA between the NICs.

Ideally, we would like to avoid all these copies and have the GPU talk directly to the NIC.

- Answer: GPUDirect

Interoperability with MPI

The GPUDirect optimized RDMA path



Interoperability with MPI

How to enable this optimized path with OpenACC

- The MPI implementation *must* support it! Cray MPICH implementation does.
  - Enabled by setting `MPICH_RDMA_ENABLED_CUDA=1`.

- If a GPU pointer passed in Send/Recv call, the MPI implementation enables the optimized GPUDirect data path.

Interoperability with MPI

How to enable this optimized path with OpenACC

- The MPI implementation *must* support it! Cray MPICH implementation does.
  - Enabled by setting `MPICH_RDMA_ENABLED_CUDA=1`.
- If a GPU pointer passed in Send/Recv call, the MPI implementation enables the optimized GPUDirect data path.
- Wrap the Send/Recv calls using the `host_data` directive to get the OpenACC device pointers.

Hands-on

Exercise 5 – Calling cuBLAS methods

Source code:

- `practicals/gemm/gemm.cpp`
- Run as:
  `srun --reserv=openacc -Cgpu ./axpy.openacc [ARRAY_SIZE]`
  - `ARRAY_SIZE` is power of 2, default is 16

Steps:

1. Compile with ‘`make CPPFLAGS=`’ to get also the naive implementation → too slow!
2. Offload the GEMM method to the GPU using OpenACC
3. Make use of cuBLAS GEMM through OpenACC
4. When does it start to pay of using cuBLAS?

Hands-on

Exercise 6.1 – 2D diffusion example

Source code:

- `diffusion2d_omp.{cpp,f90}`: our baseline code
  - Single node OpenMP version for the CPU
- `diffusion2d_openacc.{cpp,f90}`
  - Single node OpenACC version
  - Run as:

    ```
    srun --reserv=openacc -Cgpu ./diffusion2d.openacc [ARRAY_SIZE]
    ```

    - `ARRAY_SIZE` is power of 2, default is 16
  - Fill in the parts where `OPENACC_DATA` is defined.

Hands-on

Exercise 6.2 – 2D diffusion example using CUDA data management

Source code:

- `diffusion2d_openacc.{cpp,f90}`
  - Single node OpenACC version
  - Run as:

```
srun --reserv=openacc -Cgpu ./diffusion2d.openacc.cuda [ARRAY_SIZE
```

    - `ARRAY_SIZE` is power of 2, default is 16
  - Fill in the parts where `OPENACC_DATA` is defined.

Hands-on

Exercise 6.3 – 2D diffusion example using MPI

Source code:

- `diffusion2d_openacc_mpi.{cpp,f90}`
  - Run as:
    ```
    srun --reserv=openacc -Cgpu ./diffusion2d.openacc.cuda [ARRAY_SIZE
    ```
    - `ARRAY_SIZE` is power of 2, default is 16
  - Fill in the parts where `OPENACC_DATA` is defined.