HIP and ROCm

Presenter: Sam Antao
LUMI Comprehensive Training
Oct 29th, 2024

together we advance_

Thanks to all the AMD staff for their contributions to this presentation

- Suyash Tandon
- Justin Chang
- Julio Maia
- Noel Chalmers
- Paul T. Bauman
- Nicholas Curtis
- Nicholas Malaya
- Alessandro Fanfarillo
- Jose Noudohouenou
- Chip Freitag
- Damon McDougall
- Noah Wolfe
- Jakub Kurzak
- Samuel Antao
- George Markomanolis
- Bob Robey
- Gina Sitaraman
- and many more DC GPU colleagues…

Agenda

1. AMD GPU programming concepts
2. HIP API calls and GPU kernel code
3. ROCm and ROCm libraries
4. Error checking, device management, and asynchronous computing
5. Shared memory and thread synchronization

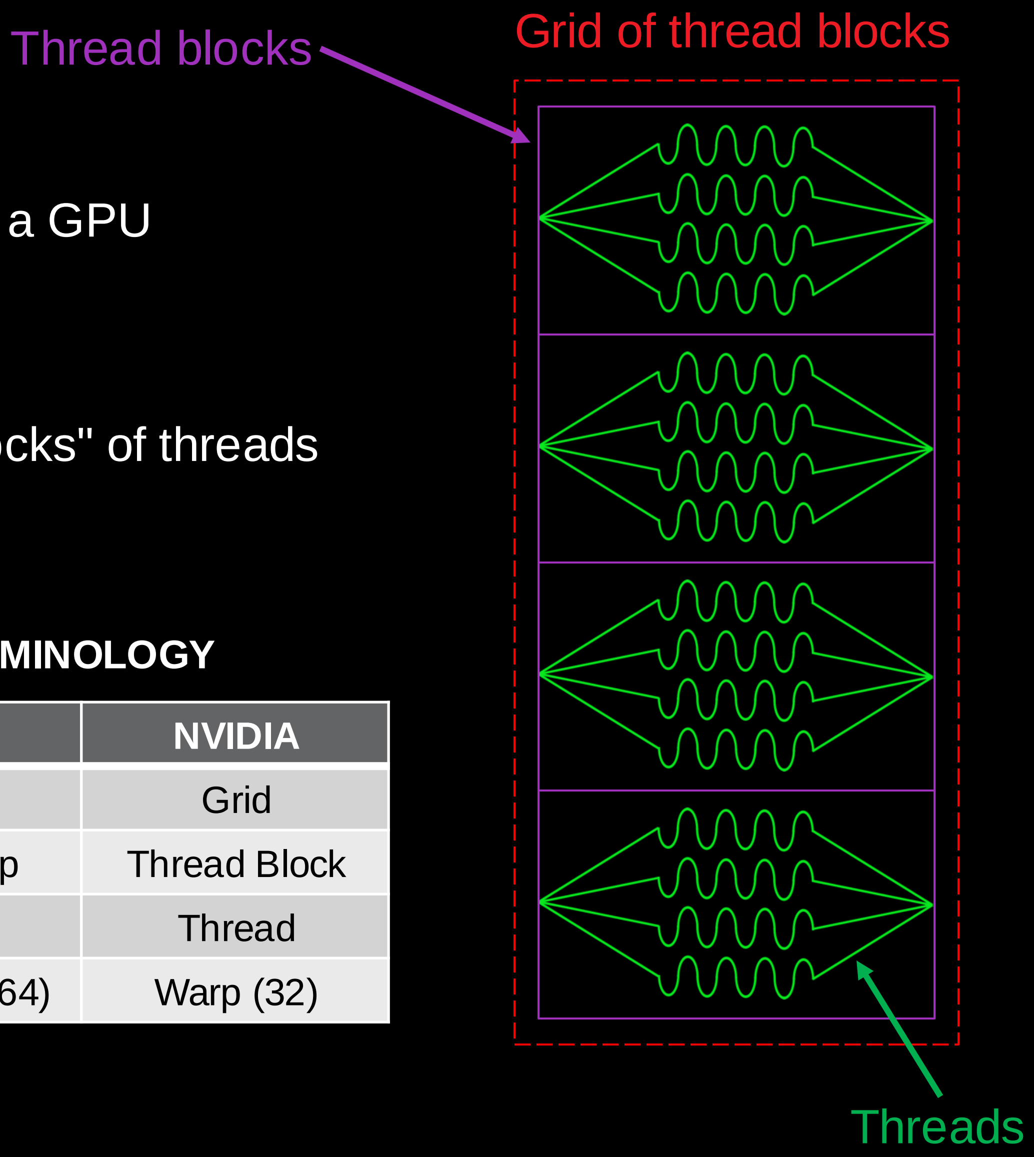
1. AMD GPU programming concepts

Device Kernels: Grid Hierarchy

- In HIP, kernels are executed on a "grid" of threads that run on a GPU
  - 1D, 2D, and 3D grids are supported, but most work maps well to 1D
  - The grid is what you map your problem to
- Each dimension of the grid is partitioned into **equal sized** "blocks" of threads
- Each block is made up of multiple "threads"
- The grid and its associated blocks are just organizational constructs, the threads are the things that do the work
- If you're familiar with CUDA already, the grid+block structure is very similar in HIP

TERMINOLOGY

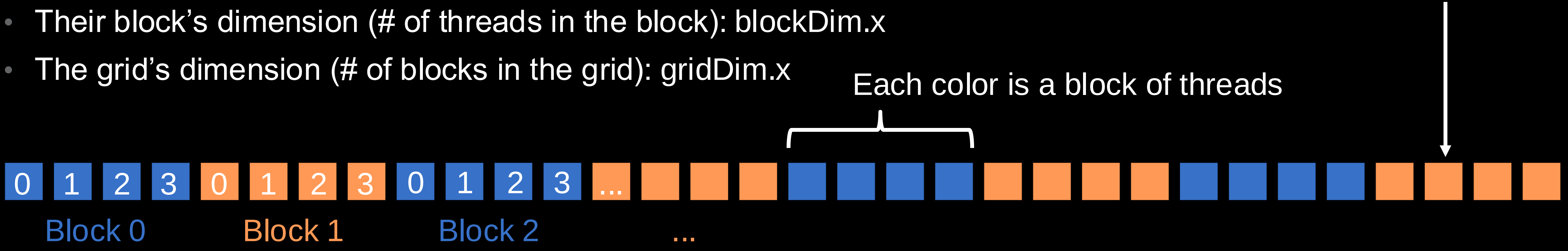
AMD	NVIDIA
Grid	Grid
Workgroup	Thread Block
Thread	Thread
Wavefront (64)	Warp (32)



The Grid: blocks of threads in 1D

Threads in grid have access to:

- Their respective block (workgroup): blockIdx.x
- Their respective thread ID **in a block**: threadIdx.x
- Their block's dimension (# of threads in the block): blockDim.x
- The grid's dimension (# of blocks in the grid): gridDim.x



Each small square is a thread

Each color is a block of threads

Global thread ID

```
int id = blockDim.x * blockIdx.x + threadIdx.x;
```

For example, thread 3 of block 2
would have a global thread ID of 11

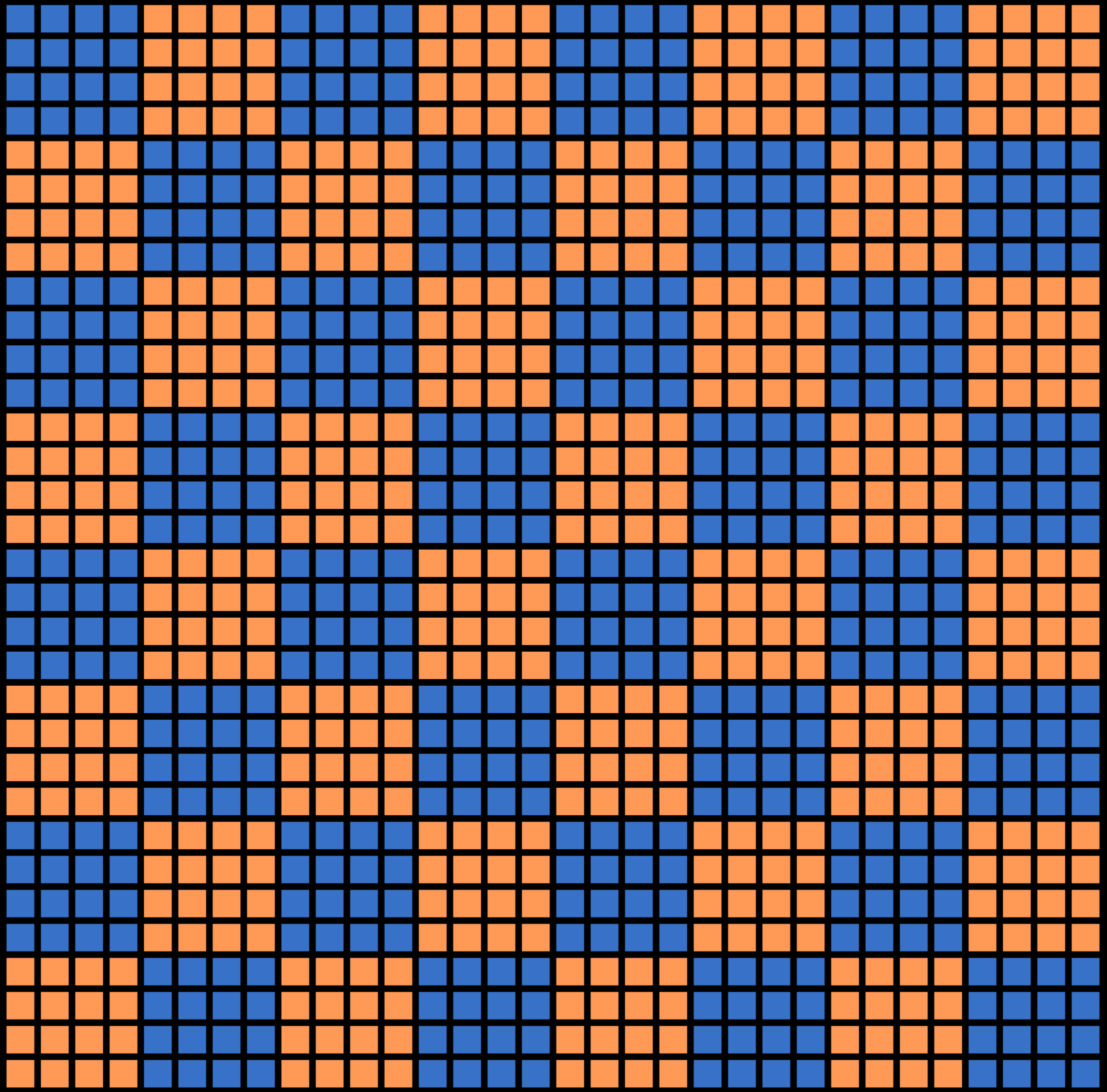
```
       = 4          * 2          + 3
       = 11
```

The Grid: blocks of threads in 2D

- The concept is the same in 1D and 2D
- In 2D each block and thread now has a two-dimensional index

Threads in grid have access to:

- Their respective block IDs: blockIdx.x, blockIdx.y
- Their respective thread IDs in a block: threadIdx.x, threadIdx.y
- Etc.

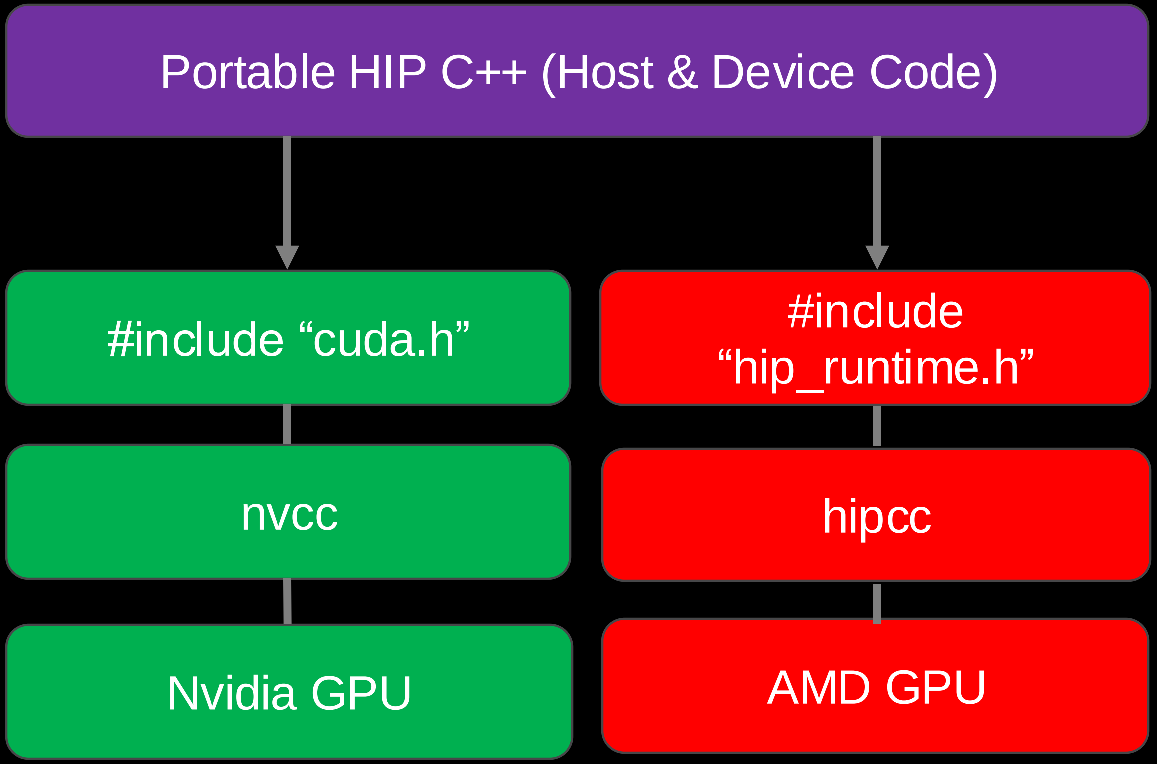


2. HIP API calls and GPU kernel code

What is HIP?

AMD's **H**eterogeneous-compute **I**nterface for **P**ortability, or **HIP**, is a C++ runtime API and kernel language that allows developers to create portable applications that can run on AMD's accelerators as well as CUDA devices

- **Open-source**
- Syntactically similar to CUDA. Most CUDA API calls can be converted in place: cuda -> hip
- Supports a strong subset of CUDA runtime functionality



https://rocm.docs.amd.com/projects/HIP/en/latest/.doxygen/docBin/html/index.html

HIP API

Device Management:

- `hipSetDevice(), hipGetDevice(), hipGetDeviceProperties()`

Memory Management

- `hipMalloc(), hipMemcpy(), hipMemcpyAsync(), hipFree()`

Streams

- `hipStreamCreate(), hipDeviceSynchronize(), hipStreamSynchronize(), hipStreamDestroy()`

Events

- `hipEventCreate(), hipEventRecord(), hipStreamWaitEvent(), hipEventElapsedTime()`

Device Kernels

- `__global__, __device__`

Device code

- `threadIdx, blockIdx, blockDim, __shared__`, 200+ math functions covering entire CUDA math library.

Error handling

- `hipGetLastError(), hipGetErrorString()`

Example: simple discrete GPU multiply

```
#include <stdio.h>
#include <math.h>
#include "hip/hip_runtime.h"

__global__ void multiply(double *A, int n)
{
  int id = blockDim.x * blockIdx.x + threadIdx.x;
  if (id < n) A[id] = 2.0 * A[id];
}

int main(int argc, char *argv[]){

  int N = 1024 * 1024;
  size_t bytes = N * sizeof(double);

  double *h_A = (double*)malloc(bytes);

  for(int i=0; i<N; i++){
    h_A[i] = (double)rand()/(double)RAND_MAX;
  }
```

```
  double *d_A;
  hipMalloc(&d_A, bytes);

  hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);

  int thr_per_blk = 256;
  int blk_in_grid = ceil( float(N) / thr_per_blk );

  multiply<<<blk_in_grid,thr_per_blk>>>(d_A, N);

  hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

  free(h_A);
  hipFree(d_A);

  printf("__SUCCESS__\n");

  return 0;
}
```

Example: simple discrete GPU multiply

Include header for HIP runtime

```
#include <stdio.h>
#include <math.h>
#include "hip/hip_runtime.h"

__global__ void multiply(double *A, int n)
{
  int id = blockDim.x * blockIdx.x + threadIdx.x;
  if (id < n) A[id] = 2.0 * A[id];
}

int main(int argc, char *argv[]){

  int N = 1024 * 1024;
  size_t bytes = N * sizeof(double);

  double *h_A = (double*)malloc(bytes);

  for(int i=0; i<N; i++){
    h_A[i] = (double)rand()/(double)RAND_MAX;
  }
```

```
  double *d_A;
  hipMalloc(&d_A, bytes);

  hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);

  int thr_per_blk = 256;
  int blk_in_grid = ceil( float(N) / thr_per_blk );

  multiply<<<blk_in_grid,thr_per_blk>>>(d_A, N);

  hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

  free(h_A);
  hipFree(d_A);

  printf("__SUCCESS__\n");

  return 0;
}
```

Example: simple discrete GPU multiply

```
#include <stdio.h>
#include <math.h>
#include "hip/hip_runtime.h"
```

GPU kernel

```
__global__ void multiply(double *A, int n)
{
  int id = blockDim.x * blockIdx.x + threadIdx.x;
  if (id < n) A[id] = 2.0 * A[id];
}
```

```
int main(int argc, char *argv[]){

  int N = 1024 * 1024;
  size_t bytes = N * sizeof(double);

  double *h_A = (double*)malloc(bytes);

  for(int i=0; i<N; i++){
    h_A[i] = (double)rand()/(double)RAND_MAX;
  }
```

```
  double *d_A;
  hipMalloc(&d_A, bytes);

  hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);

  int thr_per_blk = 256;
  int blk_in_grid = ceil( float(N) / thr_per_blk );

  multiply<<<blk_in_grid,thr_per_blk>>>(d_A, N);

  hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

  free(h_A);
  hipFree(d_A);

  printf("__SUCCESS__\n");

  return 0;
}
```

Example: simple discrete GPU multiply

```
#include <stdio.h>
#include <math.h>
#include "hip/hip_runtime.h"

__global__ void multiply(double *A, int n)
{
  int id = blockDim.x * blockIdx.x + threadIdx.x;
  if (id < n) A[id] = 2.0 * A[id];
}
```

Allocate and initialize host memory buffer

```
int main(int argc, char *argv[]){

  int N = 1024 * 1024;
  size_t bytes = N * sizeof(double);

  double *h_A = (double*)malloc(bytes);

  for(int i=0; i<N; i++){
    h_A[i] = (double)rand()/(double)RAND_MAX;
  }
```

```
  double *d_A;
  hipMalloc(&d_A, bytes);

  hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);

  int thr_per_blk = 256;
  int blk_in_grid = ceil( float(N) / thr_per_blk );

  multiply<<<blk_in_grid,thr_per_blk>>>(d_A, N);

  hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

  free(h_A);
  hipFree(d_A);

  printf("__SUCCESS__\n");

  return 0;
}
```

Example: simple discrete GPU multiply

```
#include <stdio.h>
#include <math.h>
#include "hip/hip_runtime.h"

__global__ void multiply(double *A, int n)
{
  int id = blockDim.x * blockIdx.x + threadIdx.x;
  if (id < n) A[id] = 2.0 * A[id];
}

int main(int argc, char *argv[]){

  int N = 1024 * 1024;
  size_t bytes = N * sizeof(double);

  double *h_A = (double*)malloc(bytes);

  for(int i=0; i<N; i++){
    h_A[i] = (double)rand()/(double)RAND_MAX;
  }
```

Allocate GPU buffer and copy values from CPU buffer to GPU buffer

Not needed for unified memory

```
  double *d_A;
  hipMalloc(&d_A, bytes);

  hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);
```

```
  int thr_per_blk = 256;
  int blk_in_grid = ceil( float(N) / thr_per_blk );

  multiply<<<blk_in_grid,thr_per_blk>>>(d_A, N);

  hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

  free(h_A);
  hipFree(d_A);

  printf("__SUCCESS__\n");

  return 0;
}
```

Example: simple discrete GPU multiply

```
#include <stdio.h>
#include <math.h>
#include "hip/hip_runtime.h"

__global__ void multiply(double *A, int n)
{
  int id = blockDim.x * blockIdx.x + threadIdx.x;
  if (id < n) A[id] = 2.0 * A[id];
}

int main(int argc, char *argv[]){

  int N = 1024 * 1024;
  size_t bytes = N * sizeof(double);

  double *h_A = (double*)malloc(bytes);

  for(int i=0; i<N; i++){
    h_A[i] = (double)rand()/(double)RAND_MAX;
  }
```

```
  double *d_A;
  hipMalloc(&d_A, bytes);

  hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);

  int thr_per_blk = 256;
  int blk_in_grid = ceil( float(N) / thr_per_blk );

  multiply<<<blk_in_grid,thr_per_blk>>>(d_A, N);

  hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

  free(h_A);
  hipFree(d_A);

  printf("__SUCCESS__\n");

  return 0;
}
```

Launch GPU kernel

Example: simple discrete GPU multiply

```
#include <stdio.h>
#include <math.h>
#include "hip/hip_runtime.h"

__global__ void multiply(double *A, int n)
{
  int id = blockDim.x * blockIdx.x + threadIdx.x;
  if (id < n) A[id] = 2.0 * A[id];
}

int main(int argc, char *argv[]){

  int N = 1024 * 1024;
  size_t bytes = N * sizeof(double);

  double *h_A = (double*)malloc(bytes);

  for(int i=0; i<N; i++){
    h_A[i] = (double)rand()/(double)RAND_MAX;
  }
```

```
  double *d_A;
  hipMalloc(&d_A, bytes);

  hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);

  int thr_per_blk = 256;
  int blk_in_grid = ceil( float(N) / thr_per_blk );

  multiply<<<blk_in_grid,thr_per_blk>>>(d_A, N);

  hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

  hipFree(d_A);

  free(h_A);
  printf("__SUCCESS__\n");

  return 0;
}
```

Not needed for unified memory

Copy data from GPU buffer to CPU buffer and free memory

Example: simple discrete GPU multiply

- Device memory management

```
// Allocate memory on the device
 double *d_A;
 hipMalloc(&d_A, bytes);

 // Copy values of host array (h_A) to device array (d_A)
 hipMemcpy(d_A, h_A, bytes, hipMemcpyHostToDevice);

 ...

 // Copy values of device array (d_A) to host array (h_A)
 hipMemcpy(h_A, d_A, bytes, hipMemcpyDeviceToHost);

 ...

 // Free device memory
 hipFree(d_A);
```

Example: simple discrete GPU multiply

- Kernel

```
for (int id=0; id<n; id++){
  a[id] = 2.0 * a[id];
}
```

CPU Implementation

Indicates this is a HIP kernel function launched from host

GPU kernels do not return anything

Kernel arguments

```
__global__ void multiply(double *A, int n)
{
    int id = blockDim.x * blockIdx.x + threadIdx.x;
    if (id < n) A[id] = 2.0 * A[id];
}
```

Define global thread ID

Ensure we do not access memory that does not belong to us

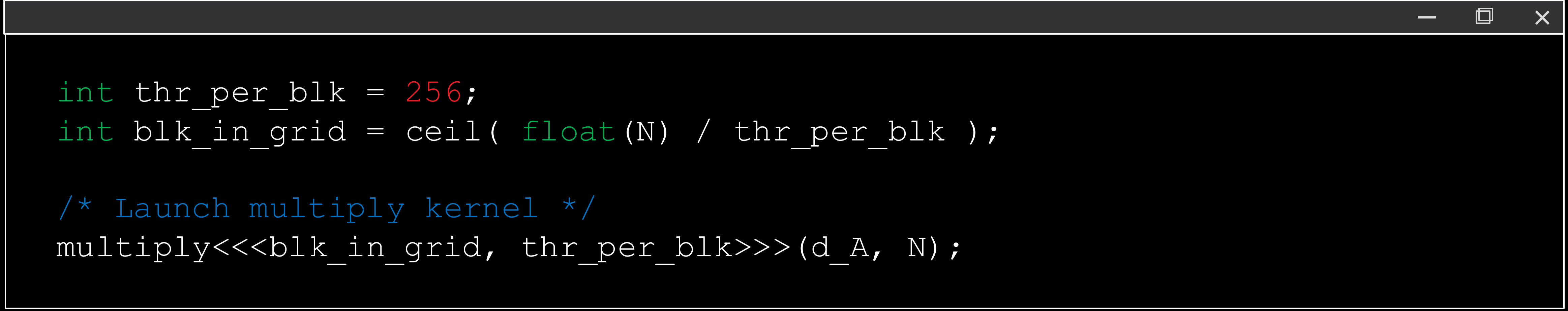
Example: simple discrete GPU multiply

- Launching the kernel

Type `dim3`

Ex: `BLOCKS_IN_GRID(<nblocksx>,`
`<nblocksy>,`
`<nblocksz>)`

```
kernel_name<<< BLOCKS_IN_GRID, THREADS_PER_BLOCK,
               [OPTIONAL] BYTES_OF_SHARED_MEMORY, [OPTIONAL] STREAM_ID >>>
               (ARG1, ARG2, ...);
```

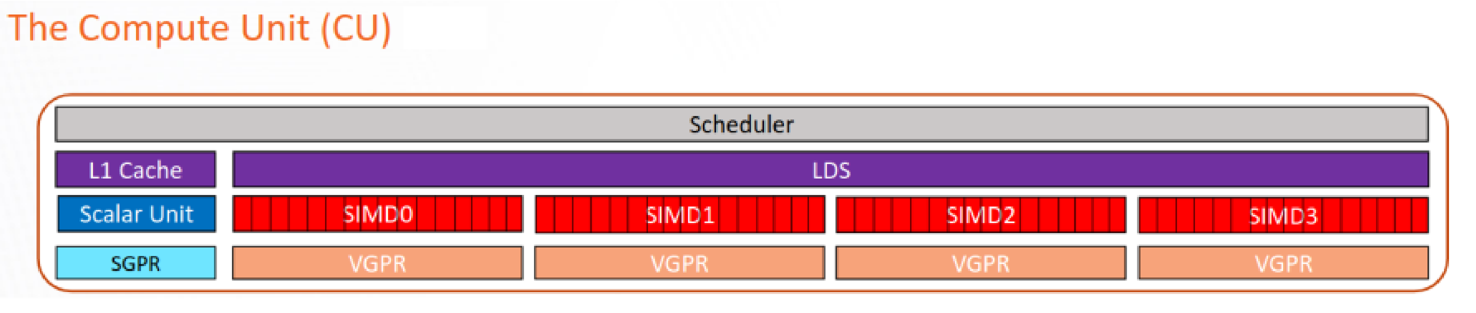


```
int thr_per_blk = 256;
int blk_in_grid = ceil( float(N) / thr_per_blk );

/* Launch multiply kernel */
multiply<<<blk_in_grid, thr_per_blk>>>(d_A, N);
```

NOTE: GPU kernel launches are asynchronous with respect to the host.

Software to hardware mapping



Blocks and threads allow a natural mapping of kernels to hardware:

- Upon kernel launch, a grid of thread blocks is launched to compute the kernel on the compute units (CUs)

Threads within a thread block (workgroup):

- **Execute on the same CU in chunks of 64 threads** called **wavefronts** (or waves).
- Share Local Data Share (LDS) memory and L1 cache
- Can synchronize

About wavefronts:

- Wavefronts execute on SIMD units (located inside the CU)
- If a wavefront stalls (e.g., data dependency) CUs can quickly **context switch** to another wavefront

A good practice is to make the **block size** a multiple of 64 and have several wavefronts (e.g., 256 threads)

3. ROCm and ROCm libraries

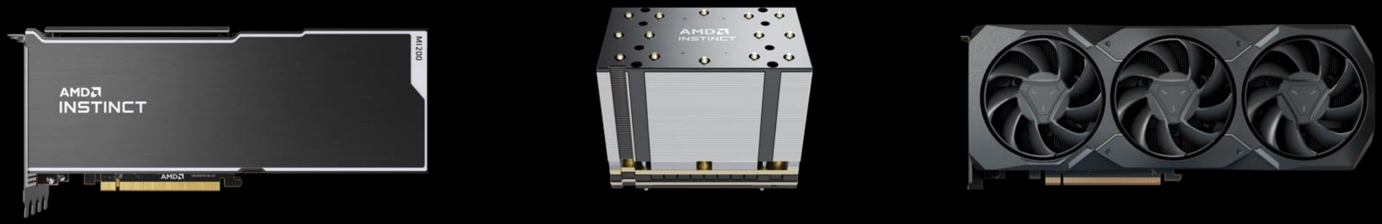
ROCm

ROCm is an open-source platform for GPU computing (including drivers, development tools, APIs, and libraries) on AMD GPUs.

- ROCm drivers allow the OS to communicate with the GPU hardware.
- ROCm libraries provide optimized routines for scientific computing and machine learning tasks, such as BLAS, FFT, etc.
- **ROCm is powered by AMD's HIP programming environment and runtime.**

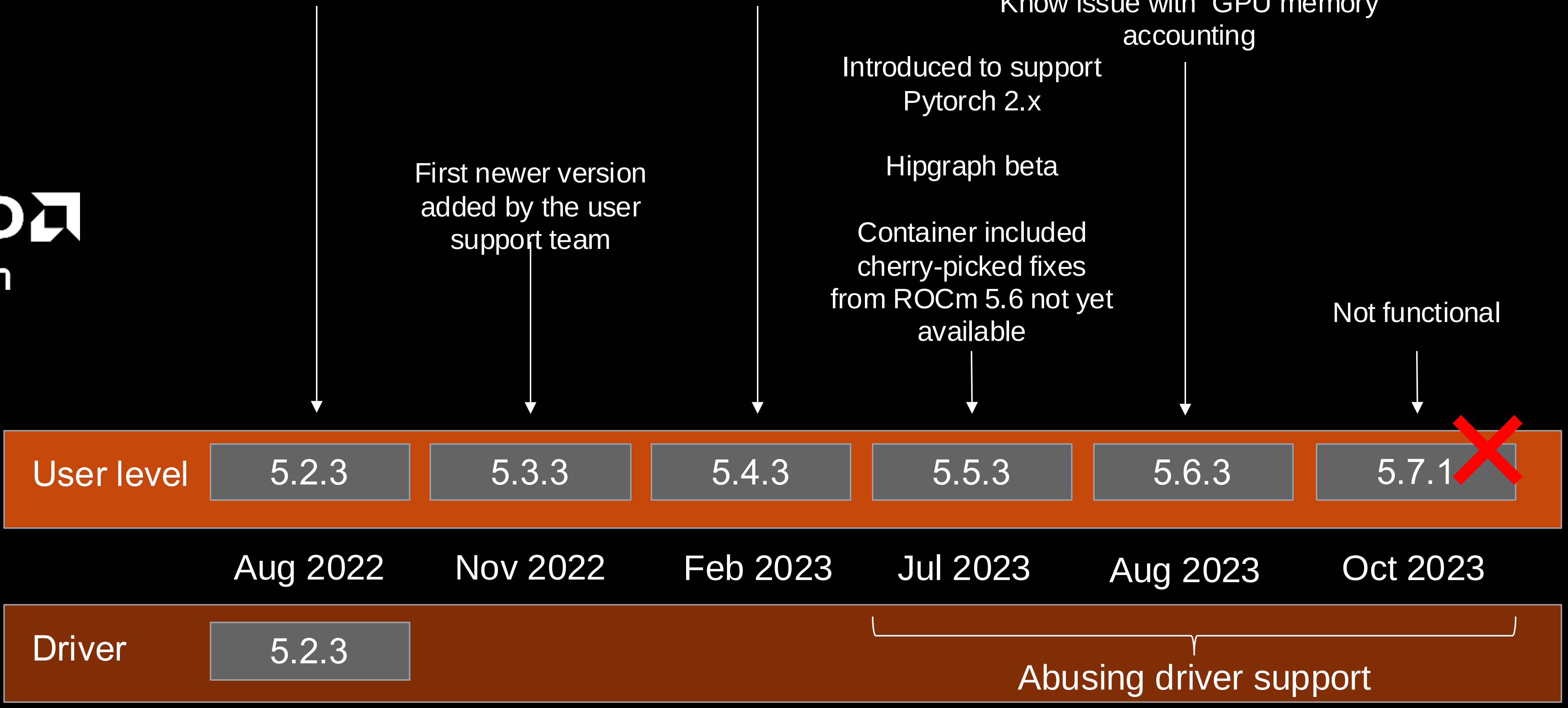
ROCm is supported on AMD **INSTINCT** & certain **RADEON** GPUs.

For the full list, please visit https://rocm.docs.amd.com/en/latest/release/gpu_os_support.html#linux-supported-gpus



[Public]

ROCm on LUMI (past)



Default version officially support

First newer version added by the user support team

Several users were working at this level

Most training courses used this level

Introduced to support Pytorch 2.x

Hipgraph beta

Container included cherry-picked fixes from ROCm 5.6 not yet available

Most AI users were at this level as it provided better support for AI frameworks (Pytorch, Triton, etc) and half precision workloads

Know issue with GPU memory accounting

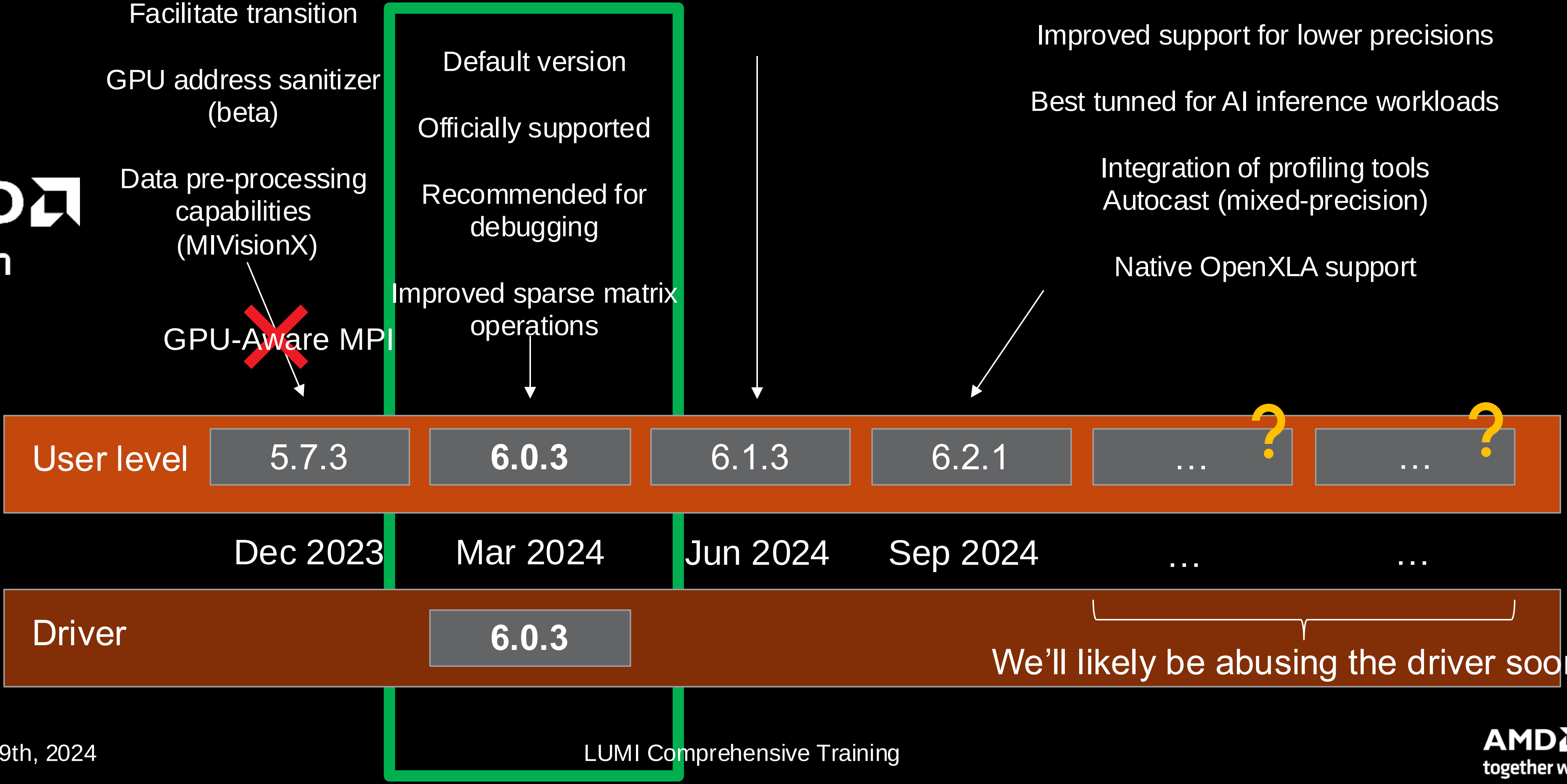
Not functional

User level	5.2.3	5.3.3	5.4.3	5.5.3	5.6.3	5.7.1 ✗
	Aug 2022	Nov 2022	Feb 2023	Jul 2023	Aug 2023	Oct 2023
Driver	5.2.3					

Abusing driver support

ROCm on LUMI

AMD ROCm



Meant to support older version of apps and frameworks

Facilitate transition

GPU address sanitizer (beta)

Data pre-processing capabilities (MIVisionX)

~~GPU-Aware MPI~~

Default version

Officially supported

Recommended for debugging

Improved sparse matrix operations

Latest Pytorch and other AI frameworks require this version

Introduced many performance improvements

Many stability and performance improvements for performance libraries

Improved support for lower precisions

Best tunned for AI inference workloads

Integration of profiling tools
Autocast (mixed-precision)

Native OpenXLA support

User level	5.7.3	**6.0.3**	6.1.3	6.2.1	… ?	… ?
	Dec 2023	Mar 2024	Jun 2024	Sep 2024	…	…
Driver		**6.0.3**				

We'll likely be abusing the driver soon

AMD together we advance_

ROCm 6.2 release specific modifications

With the release of ROCm 6.2 (https://github.com/ROCm/ROCm/releases) Omnitrace and Omniperf are **included in the ROCm stack**, but they still need to be installed.

One LUMI, we are including both version of Omnitrace and Omniperf:

- The built-in versions included in the ROCm 6.2.2 software stack (installed with `sudo apt-get` as above)
  - These can be used loading the modules:
    ```
    module use /appl/local/containers/test-modules
    module load rocm/6.2.2 omnitrace/1.12.0-rocm6.2.x omniperf/2.1.0
    ```

- The latest versions from AMD Research that would be used for ROCm releases < 6.2 (install from source)
  - These can be used by loading their dedicated modules:
    ```
    module use /appl/local/containers/test-modules
    module load rocm/6.0.3 omnitrace/1.12.0-rocm6.0.x
    module load omniperf/2.1.0
    ```

ROCm GPU libraries

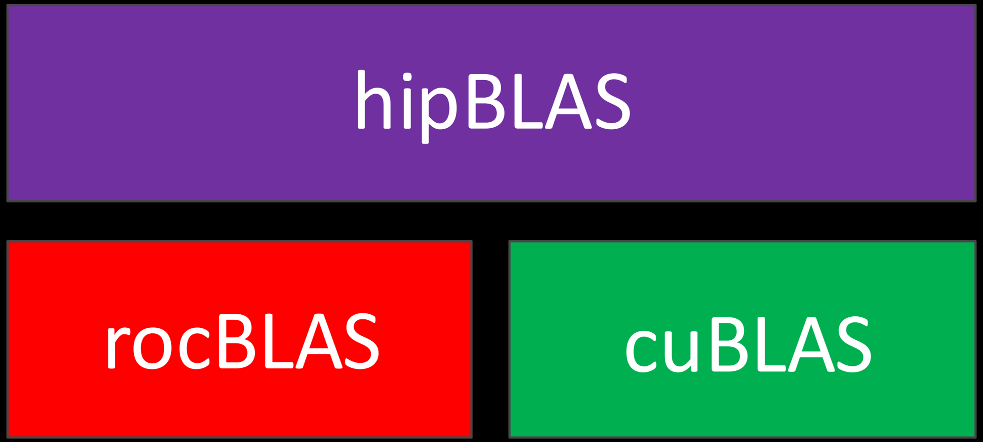
ROCm provides several GPU math libraries

- Typically, two versions:
  - roc* -> AMD GPU library, usually written in HIP
  - hip* -> Thin interface between roc* and Nvidia cu* library

When developing an application meant to target both CUDA and AMD devices, use the hip* libraries (portability)

When developing an application meant to target only AMD devices, may prefer the roc* library API (performance).

- Some roc* libraries perform **better** by using addition APIs not available in the cu* equivalents



AMD math library equivalents: “decoder ring”

CUBLAS	ROCBLAS	Basic Linear Algebra Subroutines
CUFFT	ROCFFT	Fast Fourier Transforms
CURAND	ROCRAND	Random Number Generation
THRUST	ROCTHRUST	C++ Parallel Algorithms
CUB	ROCPRIM	Optimized Parallel Primitives

AMD math library equivalents: “decoder ring”

CUSPARSE	**ROCSPARSE**	Sparse BLAS, SpMV, etc.
CUSOLVER	**ROCSOLVER**	Linear Solvers
AMGX	**ROCALUTION**	Solvers and preconditioners for sparse linear systems

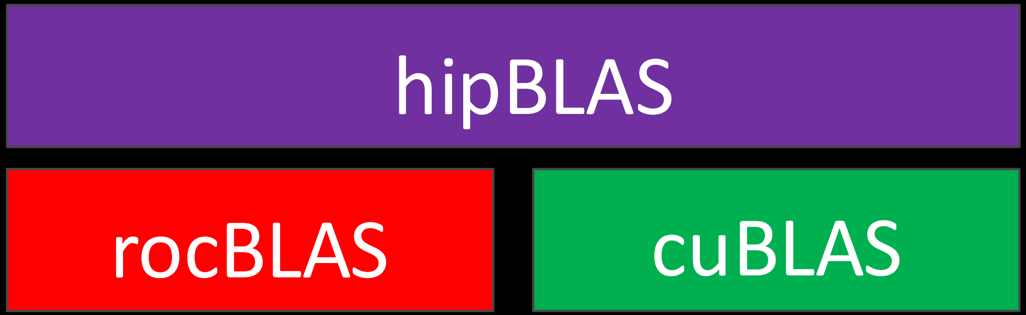
See the link below for the full list:

HTTPS://GITHUB.COM/ROCM/HIP/BLOB/AMD-STAGING/DOCS/HOW-TO/HIP_PORTING_GUIDE.MD

Example: BLAS

- rocBLAS – `sudo apt-get install rocblas`
  - Source code: https://github.com/ROCm/rocBLAS
  - Documentation: https://rocm.docs.amd.com/projects/rocBLAS/en/latest/index.html
  - Basic linear algebra functionality
    - axpy, gemv, trsm, etc
  - Use this if you need ad-hoc performance on AMD devices

- hipBLAS -
  - Source code: https://github.com/ROCm/hipBLAS
  - Documentation: https://rocm.docs.amd.com/projects/hipBLAS/en/latest/
  - Use this if you need portability between AMD and NVIDIA
  - It is just a thin wrapper:
    - It can dispatch calls to rocBLAS for AMD devices
    - It can dispatch calls to cuBLAS for NVIDIA devices



Querying system

- `rocminfo`: Queries and displays information on the system’s hardware
    - More info at: https://github.com/ROCm/rocminfo

Querying ROCm version:

- If you install ROCm in the standard location (`/opt/rocm`) version info is at: `/opt/rocm/.info/version-dev`

- `rocm-smi`: Queries and sets AMD GPU frequencies, power usage, and fan speeds
    - sudo privileges are needed to set frequencies and power limits
    - sudo privileges are not needed to query information
    - Get more info by running `rocm-smi -h` or looking at: https://github.com/ROCm/rocm_smi_lib/tree/master/python_smi_tools

```
$ /opt/rocm/bin/rocm-smi
=======================ROCm System Management Interface=======================
==============================================================================
GPU  Temp   AvgPwr  SCLK    MCLK    Fan   Perf    PwrCap  VRAM%  GPU%
1    38.0c  18.0W   1440Mhz 945Mhz  0.0%  manual  220.0W    0%   0%
==============================================================================
=============================End of ROCm SMI Log =============================
```

4. Error checking, device management, and asynchronous computing

Error Checking

There are two main types of HIP errors to check for:

- Errors returned from HIP **API calls**
  → HIP API calls return a `hipError_t` status

- Errors from HIP **kernels**
  → Synchronous errors: related to kernel launch
  → Asynchronous errors: related to kernel execution

Let’s look at how to check for these errors…

Error checking – API errors

The `hipError_t` value should be checked for all HIP API calls!

The easiest method is wrapping the API calls in a **macro**, which can be reused in all your HIP codes.

```
/* Macro for checking GPU API return values */
#define gpuCheck(call)                                                                           \
do{                                                                                              \
    hipError_t gpuErr = call;                                                                    \
    if(hipSuccess != gpuErr){                                                                    \
        printf("GPU API Error - %s:%d: '%s'\n", __FILE__, __LINE__, hipGetErrorString(gpuErr)); \
        exit(1);                                                                                 \
    }                                                                                            \
}while(0)


int main(int argc, char *argv[]){
    ...

    gpuCheck( hipMalloc(&d_A, bytes) );

    ...
}
```

Error checking – kernel errors

```
...

/* Launch multiply kernel */
multiply<<<blk_in_grid, thr_per_blk>>>(d_A, N);

/* Check for kernel launch errors */
gpuCheck( hipGetLastError() );

/* Check for kernel execution errors */
gpuCheck ( hipDeviceSynchronize() );

...
```

Why are kernel errors handled differently?

- HIP kernels do not have a return value.
- When a kernel is launched, execution is immediately given back to the host process.

So how do we handle kernel errors?

- Errors related to the kernel launch (e.g., invalid execution parameters)
  - → Manually check for the last error that occurred using `hipGetLastError()`
  - → These are known as synchronous errors
- Errors related to kernel execution (e.g., invalid memory access) can happen at any time while the kernel is running
  - → Must synchronize the device to make sure we catch these errors (`hipDeviceSychronize()`).
  - → These are known as asynchronous errors

NOTE: Device synchronization can cause reduced performance so should be reserved for debugging.

Blocking vs Nonblocking API functions

- Launching a kernel is **non-blocking for the host**
  - After sending instructions/data, the host continues to do more work while the device executes the kernel

- However, `hipMemcpy` is **blocking for the host**
  - The data pointed to in the arguments can be safely accessed/modified after the function returns

- To make asynchronous copies, we need to allocate non-pageable (pinned) host memory using hipHostMalloc and copy using hipMemcpyAsync

```
hipHostMalloc(h_a, Nbytes, hipHostMallocDefault);
hipMemcpyAsync(d_a, h_a, Nbytes, hipMemcpyHostToDevice, stream);
```

- It is not safe to access/modify the arguments of `hipMemcpyAsync` without some sort of synchronization.

Side Note: H2D/D2H bandwidth **increases significantly when host memory is pinned**

- It is good practice to use pinned host memory where data is frequently transferred to/from the device

Streams

- A stream in HIP is a **queue of tasks** (e.g. kernels, memcpys, events).
  - Tasks enqueued in a stream **complete in order on that stream**.
  - Tasks being executed in different streams are allowed to overlap and share device resources.

- Streams are created via:

```
hipStream_t stream;
hipStreamCreate(&stream);
```

- And destroyed via:

```
hipStreamDestroy(stream);
```

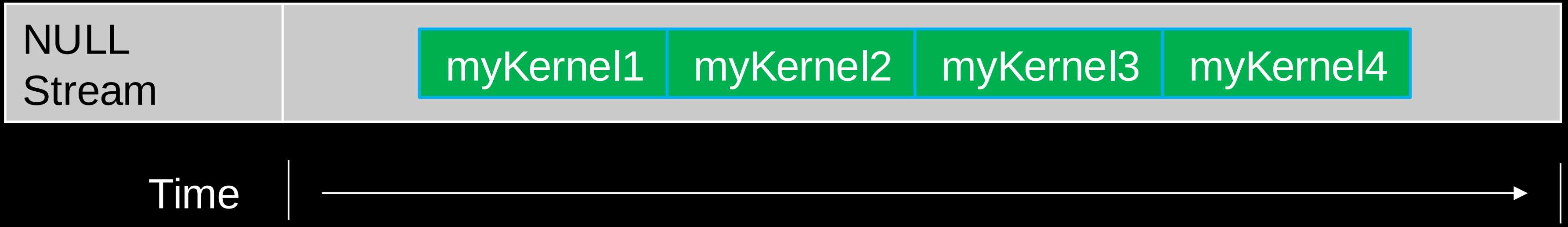
- Passing `0` or `NULL` as the `hipStream_t` argument to a function instructs the function to execute on a stream called the '**NULL Stream**':
  - No task on the NULL stream will begin until **all previously enqueued tasks in all other streams have completed**.
  - Blocking calls like `hipMemcpy` run on the NULL stream.

Streams

- Suppose we have 4 small kernels to execute:

```
myKernel1<<<dim3(1), dim3(256), 0, 0>>>(256, d_a1);
myKernel2<<<dim3(1), dim3(256), 0, 0>>>(256, d_a2);
myKernel3<<<dim3(1), dim3(256), 0, 0>>>(256, d_a3);
myKernel4<<<dim3(1), dim3(256), 0, 0>>>(256, d_a4);
```

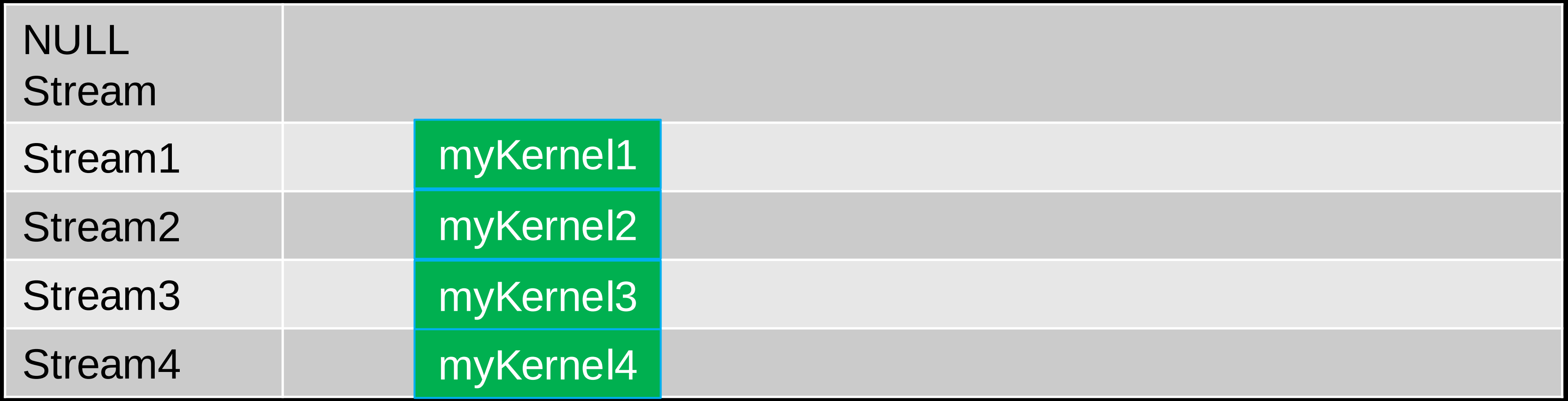
- Even though these kernels use only one block each, they’ll execute in serial on the NULL stream:



Streams

- With streams we can effectively share the GPU’s compute resources:

```
myKernel1<<<dim3(1), dim3(256), 0, stream1>>>(256, d_a1);
myKernel2<<<dim3(1), dim3(256), 0, stream2>>>(256, d_a2);
myKernel3<<<dim3(1), dim3(256), 0, stream3>>>(256, d_a3);
myKernel4<<<dim3(1), dim3(256), 0, stream4>>>(256, d_a4);
```



NULL Stream	
Stream1	myKernel1
Stream2	myKernel2
Stream3	myKernel3
Stream4	myKernel4

Note 1: Kernels must modify different parts of memory to avoid data races.
Note 2: With large kernels, overlapping computations may not help performance.

Streams

- There is another use for streams besides concurrent kernels:
  - Overlapping kernels with data movement.

- AMD GPUs have **separate engines** for:
  - Host->Device memcpys
  - Device->Host memcpys
  - Compute kernels.

- These three different operations can overlap without dividing the GPU’s resources.
  - The overlapping operations should be in separate, non-NULL, streams.
  - The host memory should be **pinned.**

Streams

Suppose we have 3 kernels which require moving data to and from the device:

```
hipMemcpy(d_a1, h_a1, Nbytes, hipMemcpyHostToDevice));
hipMemcpy(d_a2, h_a2, Nbytes, hipMemcpyHostToDevice));
hipMemcpy(d_a3, h_a3, Nbytes, hipMemcpyHostToDevice));

myKernel1<<<blocks, threads, 0, 0>>>(N, d_a1);
myKernel2<<<blocks, threads, 0, 0>>>(N, d_a2);
myKernel3<<<blocks, threads, 0, 0>>>(N, d_a3);

hipMemcpy(h_a1, d_a1, Nbytes, hipMemcpyDeviceToHost);
hipMemcpy(h_a2, d_a2, Nbytes, hipMemcpyDeviceToHost);
hipMemcpy(h_a3, d_a3, Nbytes, hipMemcpyDeviceToHost);
```

NULL Stream

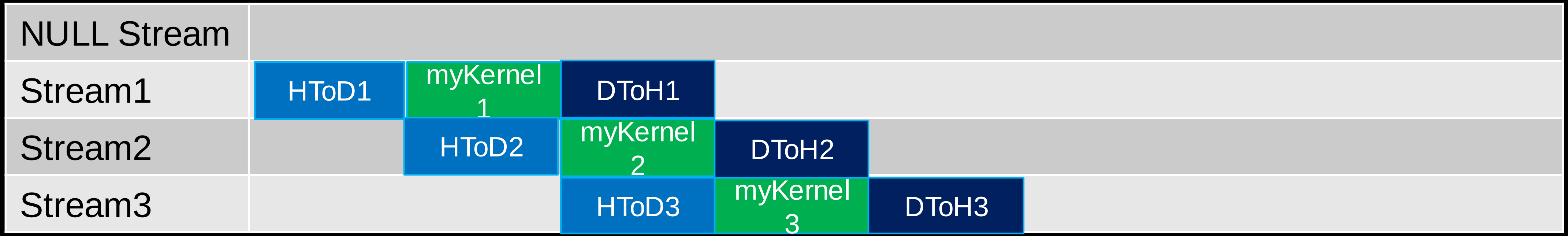
Streams

Changing to asynchronous memcpys and using streams:

```
hipMemcpyAsync(d_a1, h_a1, Nbytes, hipMemcpyHostToDevice, stream1);
hipMemcpyAsync(d_a2, h_a2, Nbytes, hipMemcpyHostToDevice, stream2);
hipMemcpyAsync(d_a3, h_a3, Nbytes, hipMemcpyHostToDevice, stream3);

myKernel1<<<blocks, threads, 0, stream1>>>(N, d_a1);
myKernel2<<<blocks, threads, 0, stream2>>>(N, d_a2);
myKernel3<<<blocks, threads, 0, stream3>>>(N, d_a3);

hipMemcpyAsync(h_a1, d_a1, Nbytes, hipMemcpyDeviceToHost, stream1);
hipMemcpyAsync(h_a2, d_a2, Nbytes, hipMemcpyDeviceToHost, stream2);
hipMemcpyAsync(h_a3, d_a3, Nbytes, hipMemcpyDeviceToHost, stream3);
```



5. Shared memory and thread syncronization

Synchronization

How do we coordinate execution on device streams with host execution? Need some synchronization points.

- `hipDeviceSynchronize();`
  - Heavy-duty sync point.
  - Blocks host until **all work** in **all device streams** has reported complete.

  - `hipStreamSynchronize(stream);`
  - Blocks host until **all work in stream** has reported complete.

Can a stream synchronize with another stream? For that we need 'Events':

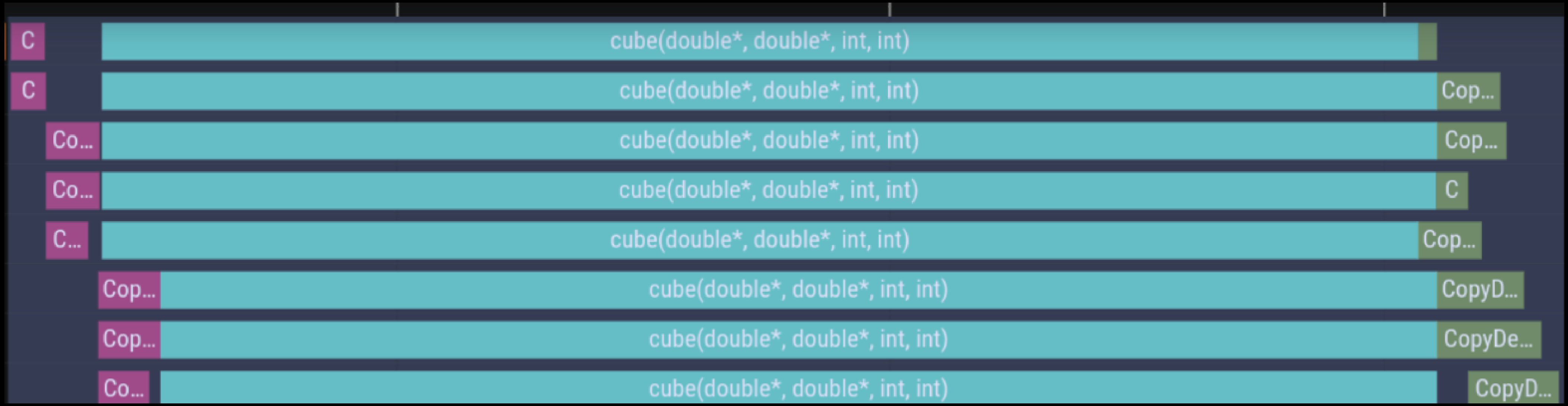
https://rocm.docs.amd.com/projects/HIP/en/latest/.doxygen/docBin/html/group___event.html

HIP stream example

In real stream overlapping, the communication and computation time will not be the same

For a real example of overlapping compute and communication in HIP

```
git clone https://github.com/AMD/HPCTrainingExamples
cd HPCTrainingExamples/HIP/Stream_Overlap
```



Device management

Multiple GPUs in system? Multiple host threads/MPI ranks? What device are we running on?

- Host can query number of devices visible to system:

```
int numDevices = 0;
hipGetDeviceCount(&numDevices);
```

- Host tells the runtime to issue instructions to a particular device:

```
int deviceID = 0;
hipSetDevice(deviceID);
```

- Host can query what device is currently selected and device properties:

```
hipGetDevice(&deviceID);
hipDeviceProp_t props;
hipGetDeviceProperties(&props, deviceID);
```

The host can manage several devices by swapping the currently selected device during runtime.
Different processes can use different devices or over-subscribe (share) the same device.

Function qualifiers

hipcc makes two compilation passes through source code. One to compile host code, and one to compile device code.

- __global__ functions:
  - These are entry points to device code, called from the host
  - Code in these regions will execute on SIMD units
- __device__ functions:
  - Can be called from __global__ and other __device__ functions.
  - Cannot be called from host code.
  - Not compiled into host code – essentially ignored during host compilation pass
- __host__ __device__ functions:
  - Can be called from __global__, __device__, and host functions.
  - Will execute on SIMD units when called from device code!

Memory declarations in device code

- Malloc/free not supported in device code.

- Variables/arrays can be declared on the stack.

- Stack variables declared in device code are allocated in registers and are private to each thread.

- Threads can all access common memory via device pointers, but otherwise do not share memory.
  - Important exception: __shared__ memory

- Stack variables declared as __shared__:
  - Allocated once per block in LDS memory
  - **Shared and accessible by all threads in the same block**
  - Access is faster than device global memory (but slower than register)
  - Must have size known at compile time

Shared memory

```
__global__ void reverse(double *d_a) {
  __shared__ double s_a[256]; //array of doubles, shared in this block

  int tid = threadIdx.x;
  s_a[tid] = d_a[tid];    //each thread fills one entry

  //all wavefronts must reach this point before any wavefront is allowed to continue.
  __syncthreads();

  d_a[tid] = s_a[255-tid]; //write out array in reverse order
}

int main() {
  …
  reverse<<<dim3(1), dim3(256), 0, 0>>>(d_a); //Launch kernel
  …
}
```

Thread synchronization

_syncthreads():

- Blocks a wavefront from continuing execution until all wavefronts have reached __syncthreads()
- Memory transactions made by a thread before __syncthreads() are visible to all other threads in the block after __syncthreads()
- Can have a noticeable overhead if called repeatedly

Best practice: Avoid deadlocks by checking that **all** threads in a block execute **the same** __syncthreads() instruction.

- *Note 1*: So long as at least one thread in the wavefront encounters __syncthreads(), the whole wavefront is considered to have encountered __syncthreads().

- *Note 2*: Wavefronts can synchronize at different __syncthreads() instructions, and if a wavefront exits a kernel completely, other wavefronts waiting at a __syncthreads() may be allowed to continue.

Hands-on exercises

https://hackmd.io/@sfantao/lumi-training-ams-2024#HIP-Exercises

https://hackmd.io/@sfantao/lumi-training-ams-2024#Hipify

We welcome you to explore our HPC Training Examples repo:

https://github.com/amd/HPCTrainingExamples

A table of contents for the READMEs if available at the top-level **README** in the repo

Relevant exercises for this presentation located in **HIP** directory.

Link to instructions on how to run the tests: **HIP/README.md and subdirectories**

[Public]

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED 'AS IS." AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Third-party content is licensed to you directly by the third party that owns the content and is not licensed to you by AMD. ALL LINKED THIRD-PARTY CONTENT IS PROVIDED "AS IS" WITHOUT A WARRANTY OF ANY KIND. USE OF SUCH THIRD-PARTY CONTENT IS DONE AT YOUR SOLE DISCRETION AND UNDER NO CIRCUMSTANCES WILL AMD BE LIABLE TO YOU FOR ANY THIRD-PARTY CONTENT. YOU ASSUME ALL RISK AND ARE SOLELY RESPONSIBLE FOR ANY DAMAGES THAT MAY ARISE FROM YOUR USE OF THIRD-PARTY CONTENT.

© 2024 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD CDNA, AMD ROCm, AMD Instinct, and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names are for informational purposes only and may be trademarks of their respective owners.

OpenCL™ is a trademark of Apple Inc. used by permission by Khronos Group, Inc.
The OpenMP® name and the OpenMP logo are registered trademarks of the OpenMP Architecture Review Board
Git and the Git logo are either registered trademarks or trademarks of Software Freedom Conservancy, Inc., corporate home of the Git Project, in the United States and/or other countries