CUDA Terminology

- CUDA Terminology
  - **Kernel** = Function which is executed on the GPU
  - **Host** = CPU
  - **Device** = GPU
- Qualifiers
  - *__global__*
  - *__device__*
- Built-in variables
  - *threadIdx.{x,y,z}*
  - *blockIdx.{x,y,z}*
  - *blockDim.{x,y,z}*
  - *gridDim.{x,y,z}*
  - *dim3*

Allocate device memory

```
cudaError_t  cudaMalloc  (  void**  devPtr,
                            size_t  size,
                         )
```

- Allocates device memory
- Implicit synchronization

```
cudaError_t  cudaFree  (  void*  devPtr )
```

- Frees device memory

Device ↔ Host communication

```
cudaError_t   cudaMemcpy   (   void*                  dst,
                               const void*            src,
                               size_t                 count,
                               enum cudaMemcpyKind    kind
                           )
```

- Synchronous copy data between host and device
- Kind is of form:
  - *cudaMemcpyHostToHost*
  - *cudaMemcpyHostToDevice*
  - *cudaMemcpyDeviceToHost*
  - *cudaMemcpyDeviceToDevice*

Error handling

RWTH AACHEN UNIVERSITY

```
cudaError_t cudaGetLastError ( void )
```

- Returns the last error and removes it from the error queue
- Returns *cudaSuccess* if no error has occurred

```
const char* cudaGetErrorString ( cudaError_t error )
```

- Returns a meaningful error message based on the input error

```
cudaError_t   cudaEventCreate   (   cudaEvent_t*   event )
```

```
cudaError_t   cudaEventDestroy   (   cudaEvent_t   event )
```

```
cudaError_t   cudaEventElapsedTime   (   float*         ms,
                                         cudaEvent_t   start,
                                         cudaEvent_t   stop
                                     )
```

- Measures the time between two events and stores it into *ms*

```
cudaError_t  cudaEventRecord  (  cudaEvent_t  event,
                                  cudaStrea_t  stream = 0
                               )
```

- Inserts an event into a given stream

```
cudaError_t  cudaEventSynchronize  (  cudaEvent_t  event )
```

- Waits for completion of the given event

Async Data Transfer

```
cudaError_t   cudaMallocHost   (   void**   ptr,
                                   size_t   size,
                               )
```

- Allocates page-locked host memory

```
cudaError_t   cudaFreeHost   (   void*   ptr )
```

```
cudaError_t   cudaMemcpyAsync   (   void*                   dst,
                                    const void*             src,
                                    size_t                  count,
                                    enum cudaMemcpyKind     kind
                                    cudaStream_t            stream =
                                )
```

- Asynchronous data transfer
- Requires page-locked memory