"Brief Summary of CUDA"

Function Qualifers

__global__	called from host, executed on device
__device__	called from device, executed on device

Variable Qualifiers (Device)

__device__	variable on device (global memory)

Built-in Variables (Device)

dim3 gridDim	dimensions of the current grid (gridDim.x, . . .)
dim3 blockDim	dimensions of the current block (composed of threads)
uint3 blockIdx	block location in the grid (blockIdx.x, . . .)
uint3 threadIdx	thread location in the block (threadIdx.x, . . .)

Thread ID

1-D int tid = threadIdx.x + blockDim.x * blockIdx.x;

2-D
```
int col = threadIdx.x + blockIdx.x*blockDim.x;
int row = threadIdx.y + blockIdx.y*blockDim.y;
int index = col + row * N;
```

Host / Device Memory

cudaMalloc(&devptr, size)	Allocate Device Memory
cudaFree(devptr)	Free Device Memory
cudaMemcpy(dst, src, size, cudaMemcpyKind kind)	Transfer Memory kind=cudaMemcpyHostToDevice, cudaMemcpyDeviceToHost,...

Synchronizing

syncthreads()	Synchronizing one Block (device call)
cudaDeviceSynchronize()	Synchronizing all Blocks (host call)

Kernel

kernel<<<dim3 blocks, dim3 threads[, ...]>>>(arguments)	Kernel launch