

CUDA

Kernel execution	<code>kernelName << <num_blocks, num_threads, stream>> (args);</code>	Copy to constant	<code>cudaMemcpy(dst, src, sizeInBytes, cudaMemcpyKind);</code>
Kernel definition	<code>__global__ kernelName(args){...}</code>	Documentation	https://docs.nvidia.com/cuda/cuda-runtime-api/index.html
Create memory on device	<code>cudaMalloc(memPtr, sizeInBytes);</code>	Device-only function	<code>__device__ auto functionName(args) -> returnType</code>
Free device memory	<code>cudaFree(memPtr);</code>	Host-only function	<code>__host__ auto functionName(args) -> returnType</code>
Create pinned memory on host	<code>cudaMallocHost(memPtr, size_in_bytes);</code>	Host and device function	<code>__device__ __host__ auto functionName(args) -> returnType{...}</code>
Free pinner memory	<code>cudaFreeHost(memPtr);</code>	Thrust	
Create stream	<code>cudaStreamCreate(&stream);</code>	Universal Vector	<code>thrust::universal_vector<type> name(size);</code>
Destroy stream	<code>cudaStreamDestroy(stream);</code>		
Wait for Device	<code>cudaDeviceSynchronize();</code>		
Wait for Stream	<code>cudaStreamSynchronize(stream);</code>		
Copy data	<code>cudaMemcpy(dst, src, numElements, cudaMemcpyDefault);</code>		
Async mem copy	<code>cudaMemcpyAsync(dst, src, numElements, cudaMemcpyDefault, stream);</code>		
Static device variable	<code>__device__ typeName = value;</code>		
Static shared mem	<code>__shared__ typeName[size];</code>		
Dynamic shared mem	<code>extern __shared__ typeName;</code>		
Constant device mem	<code>__constant__ typeName;</code>		

