

# Concurrent CUDA Streams



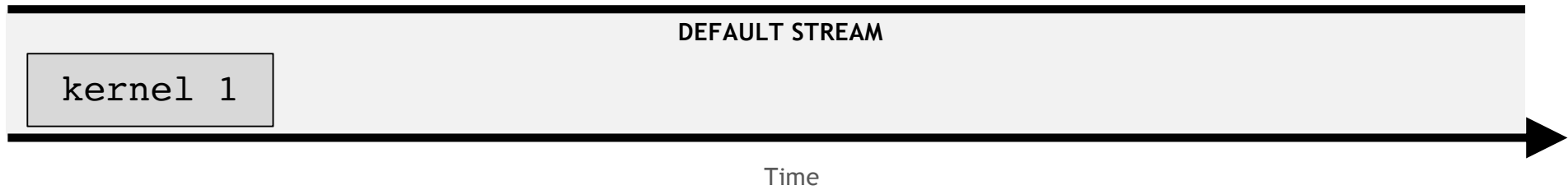
A **stream** is a series of instructions,  
and CUDA has a **default stream**



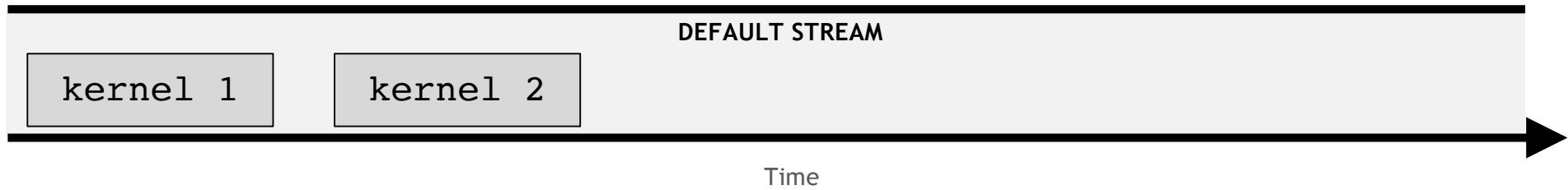
**DEFAULT STREAM**

Time

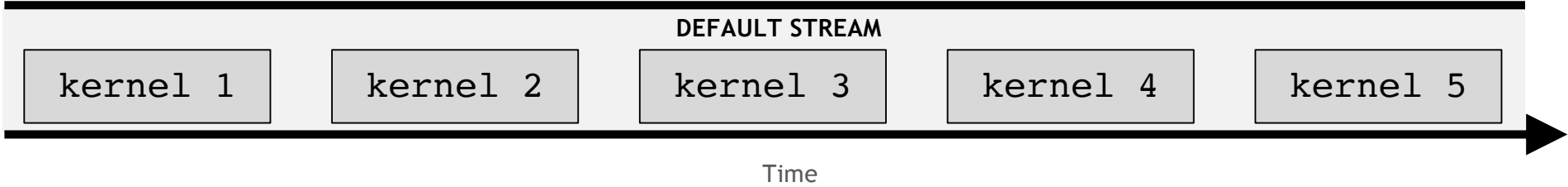
By default, CUDA kernels run in the **default stream**



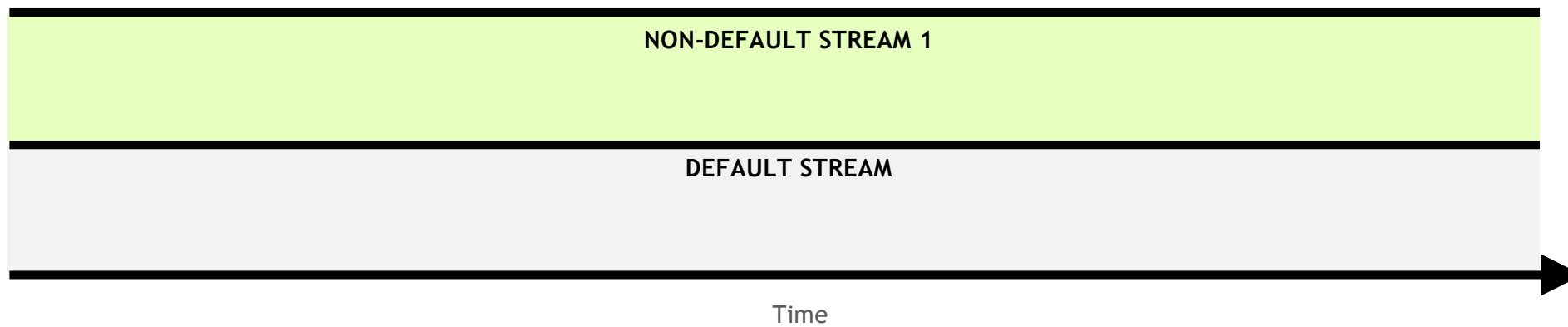
In any stream, including the default, an instruction in it (here a kernel launch) must complete before the next can begin



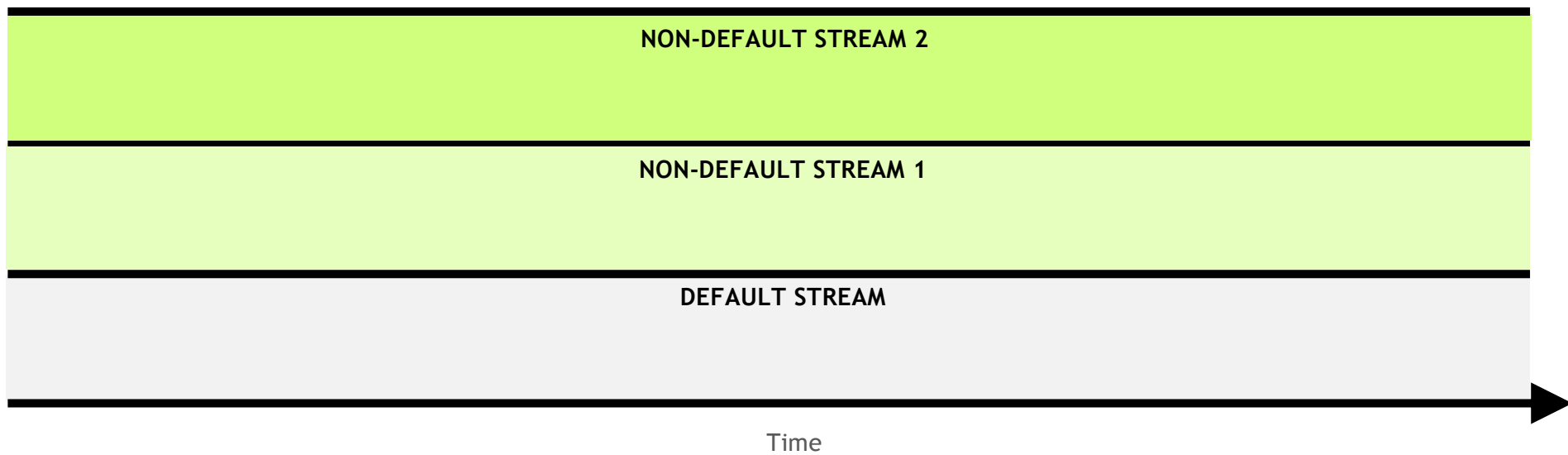
In any stream, including the default, an instruction in it (here a kernel launch) must complete before the next can begin



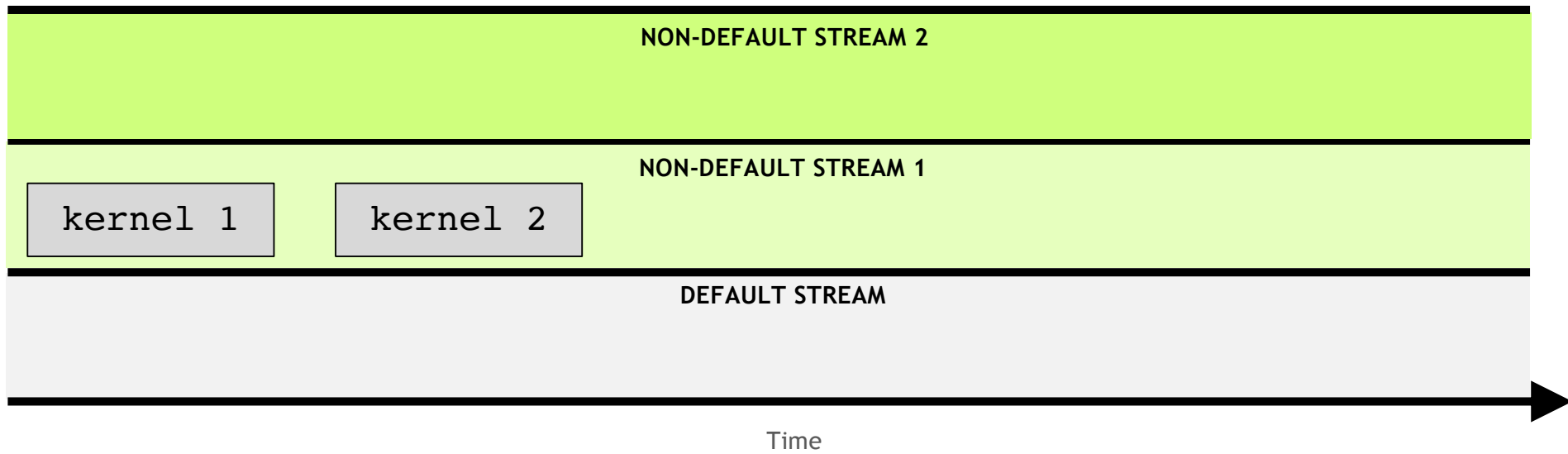
**Non-default streams** can also be created for kernel execution



**Non-default streams** can also be created for kernel execution

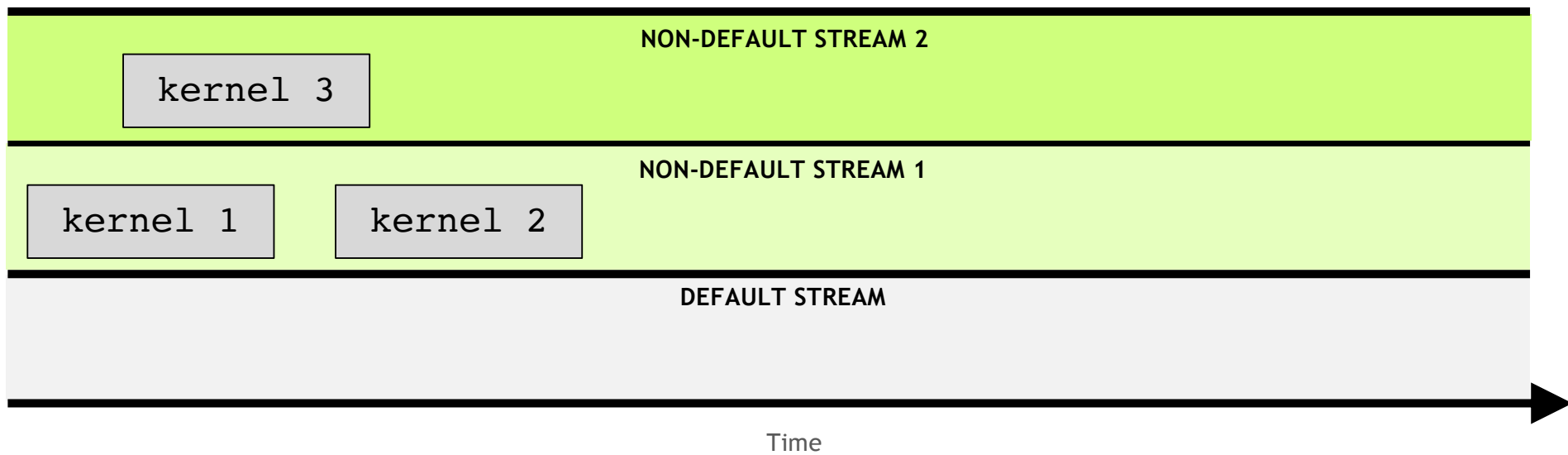


Kernels within any single stream must execute in order

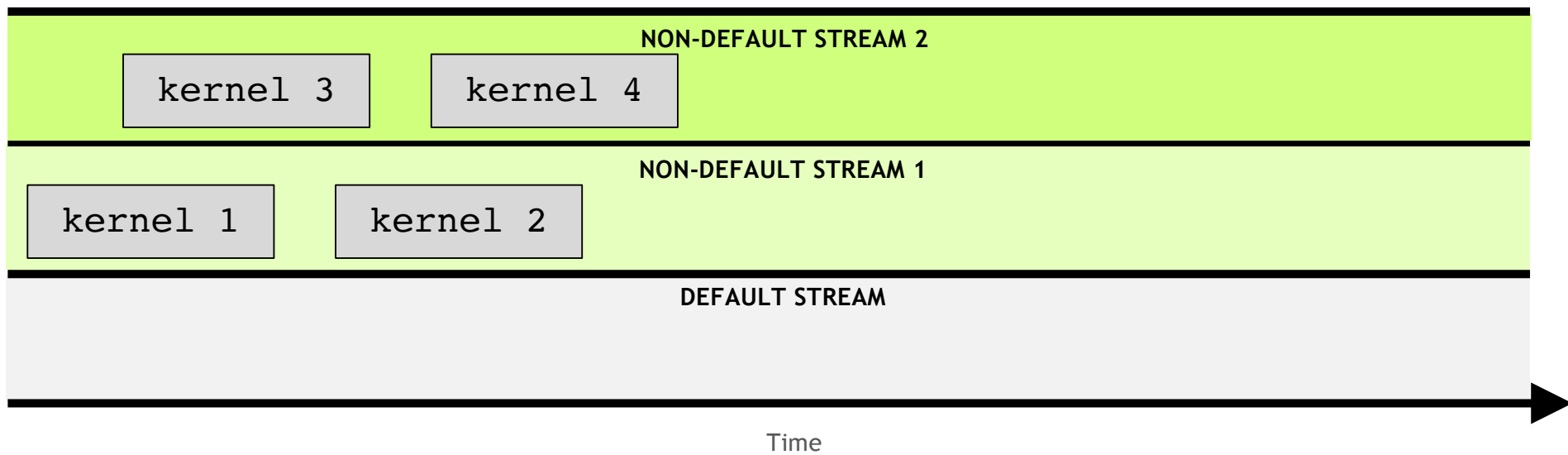




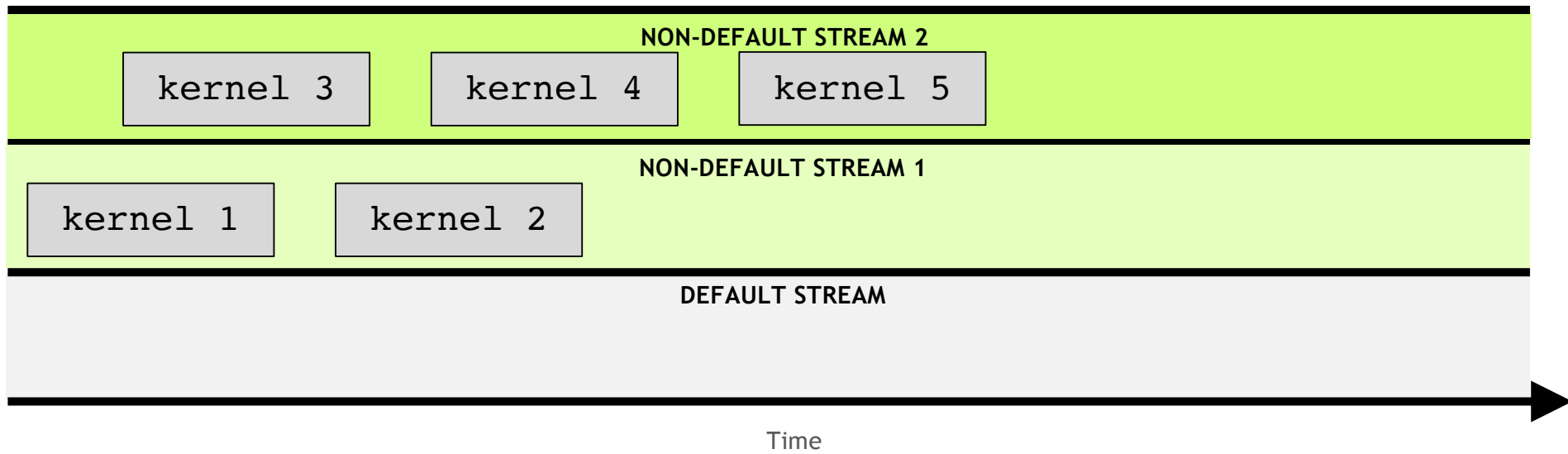
However, kernels in **different, non-default streams**, can interact concurrently



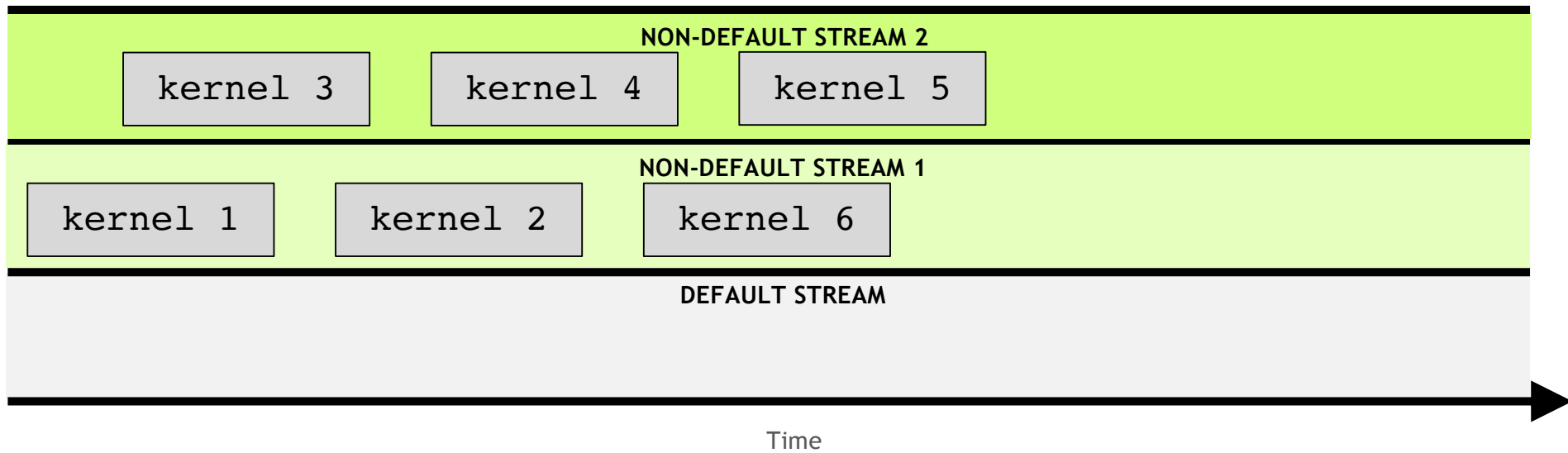
However, kernels in **different, non-default streams**, can interact concurrently



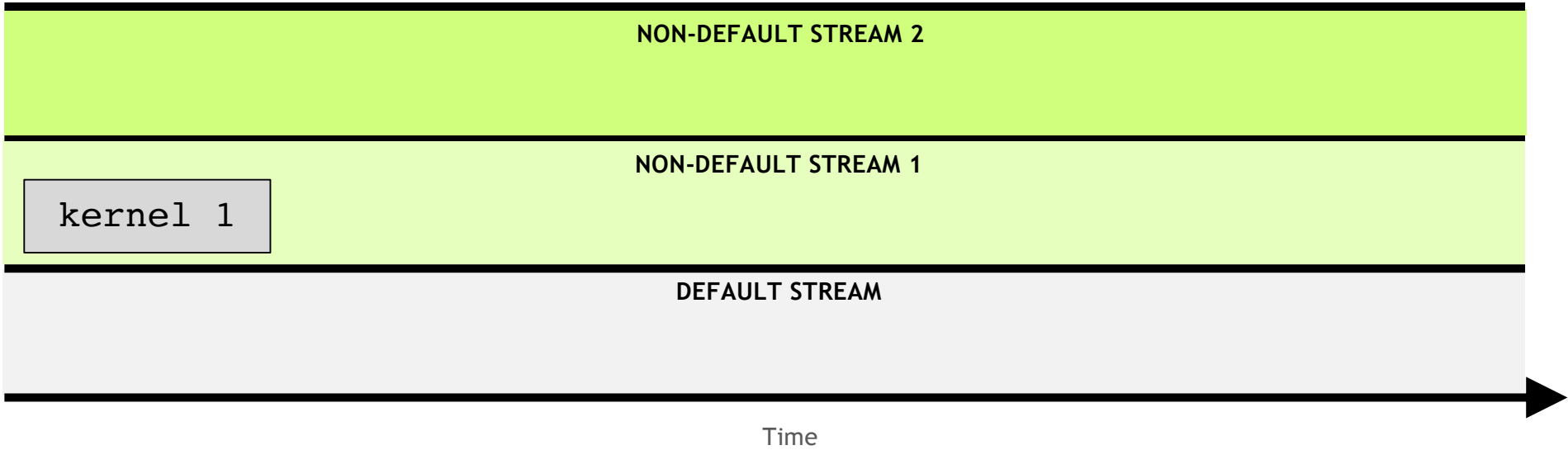
However, kernels in **different, non-default streams**, can interact concurrently



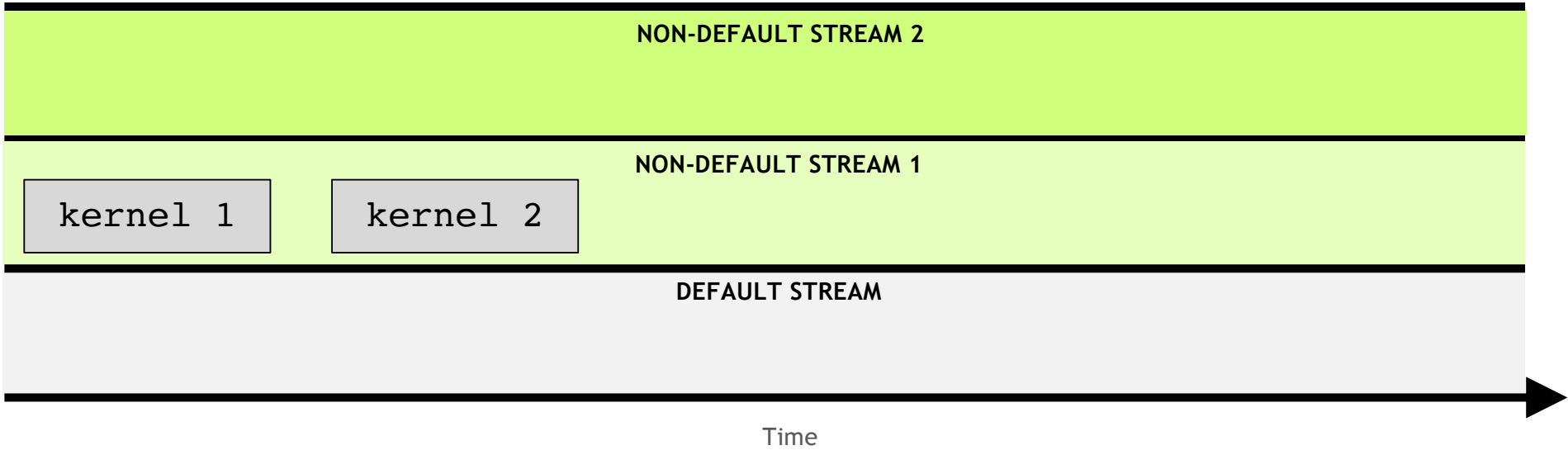
However, kernels in **different, non-default streams**, can interact concurrently



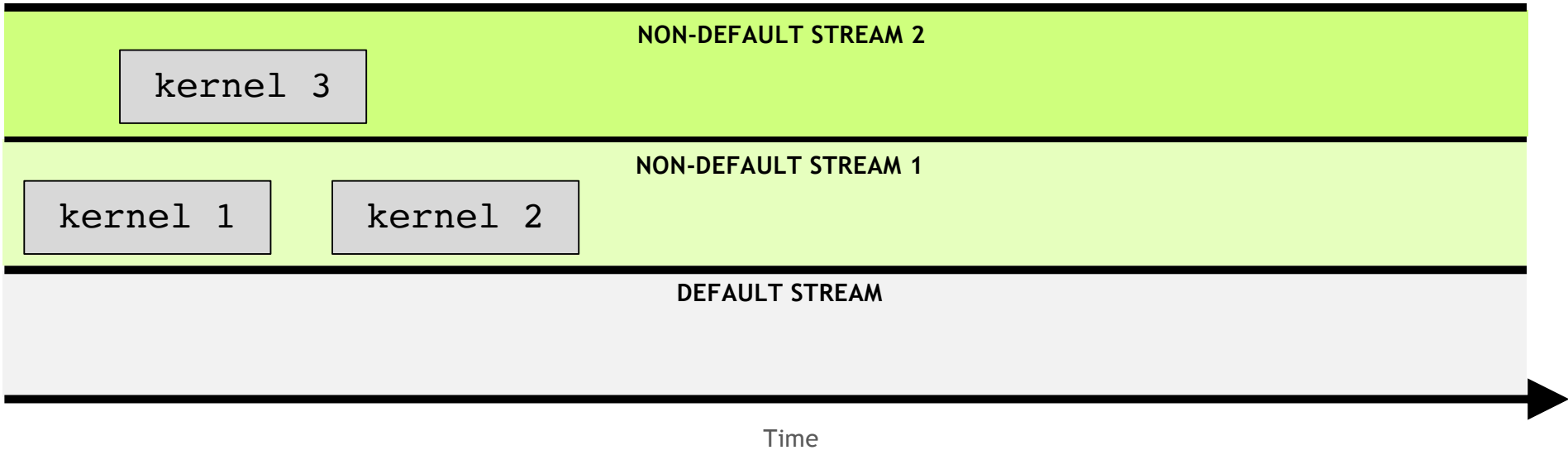
The default stream is special: **it blocks all kernels in all other streams**



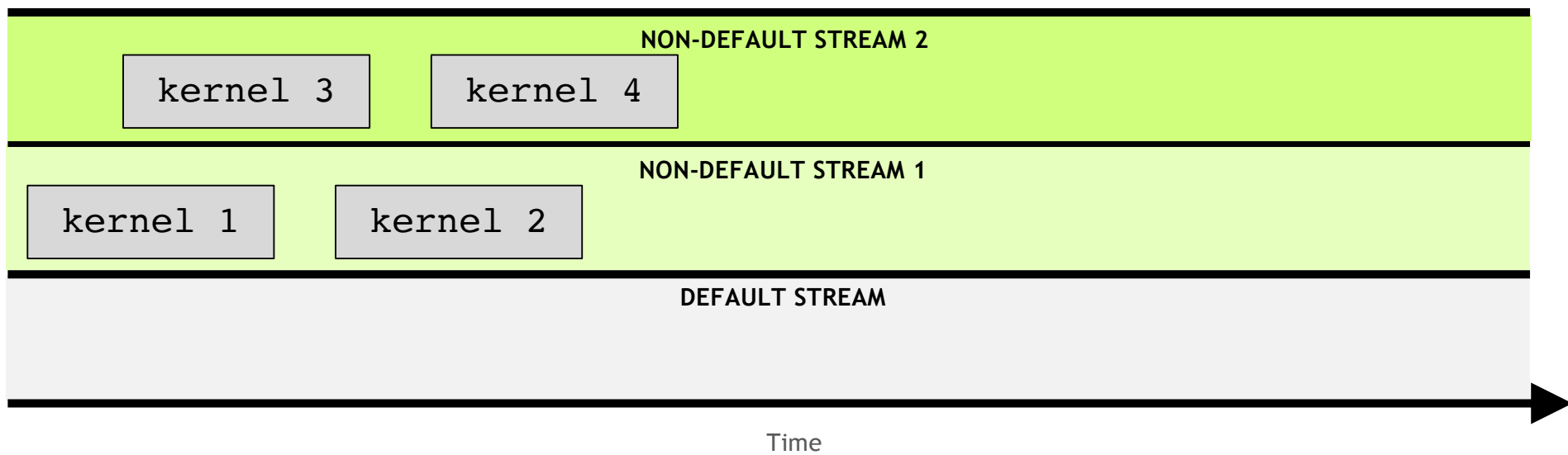
The default stream is special: **it blocks all kernels in all other streams**



The default stream is special: **it blocks all kernels in all other streams**

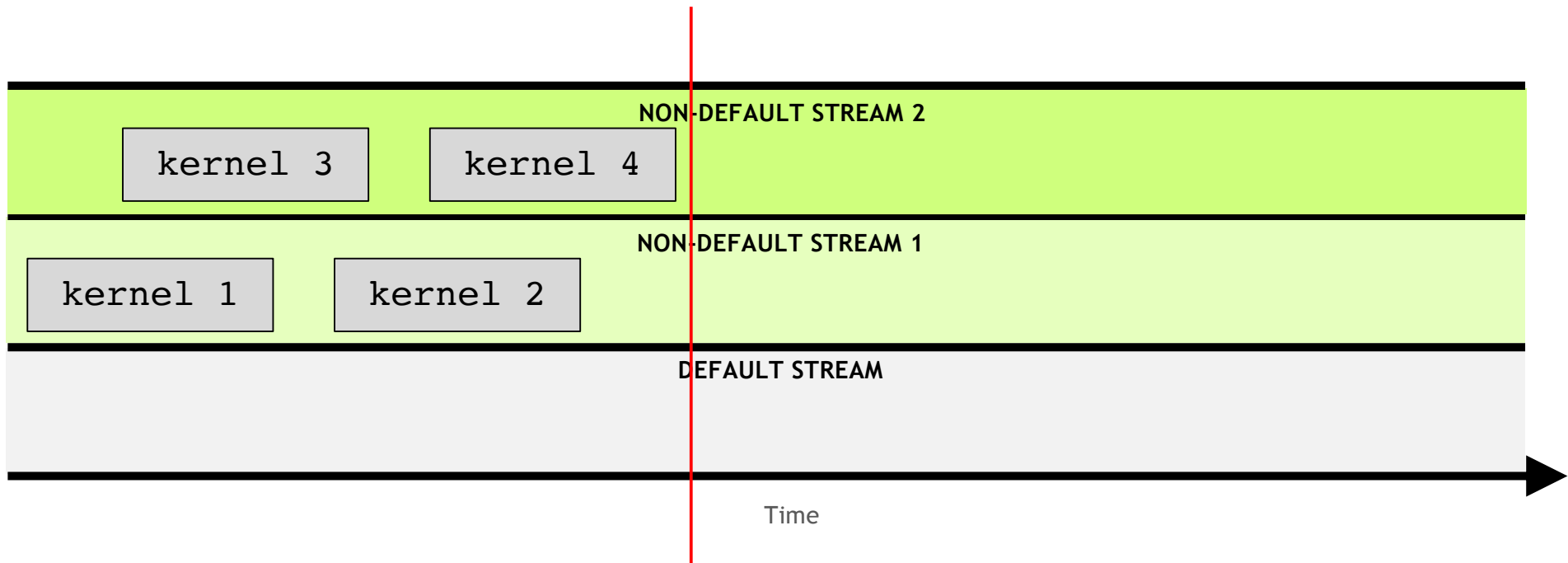


The default stream is special: **it blocks all kernels in all other streams**

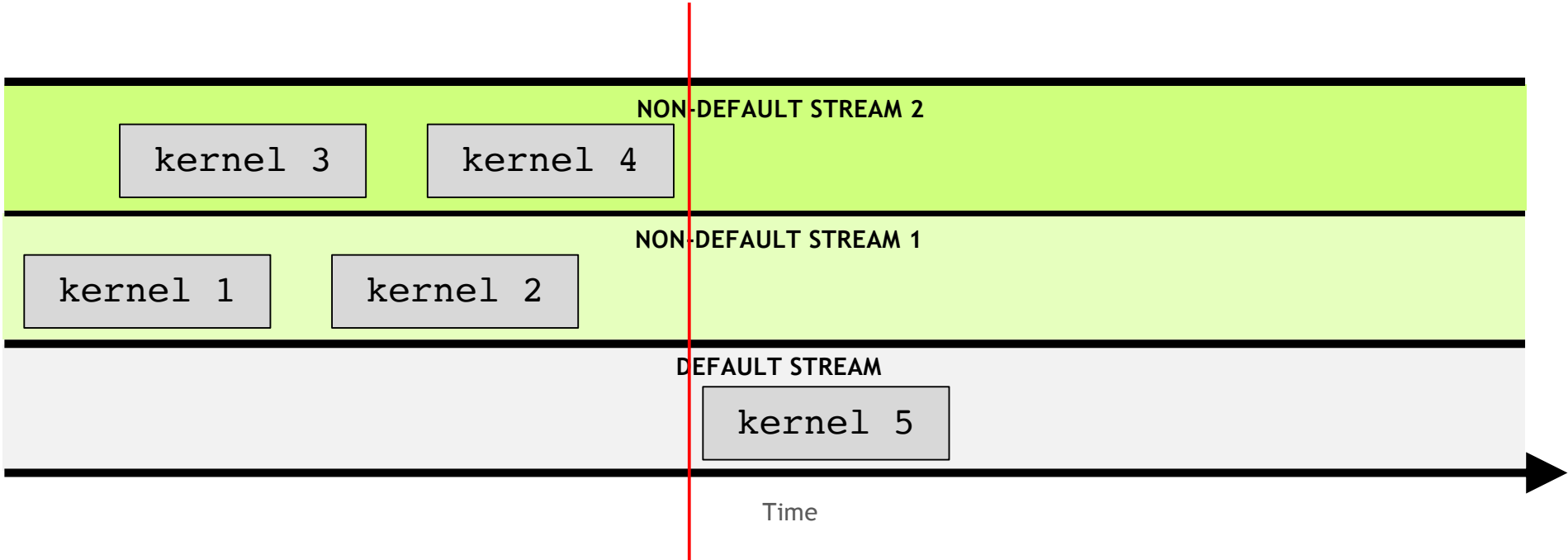




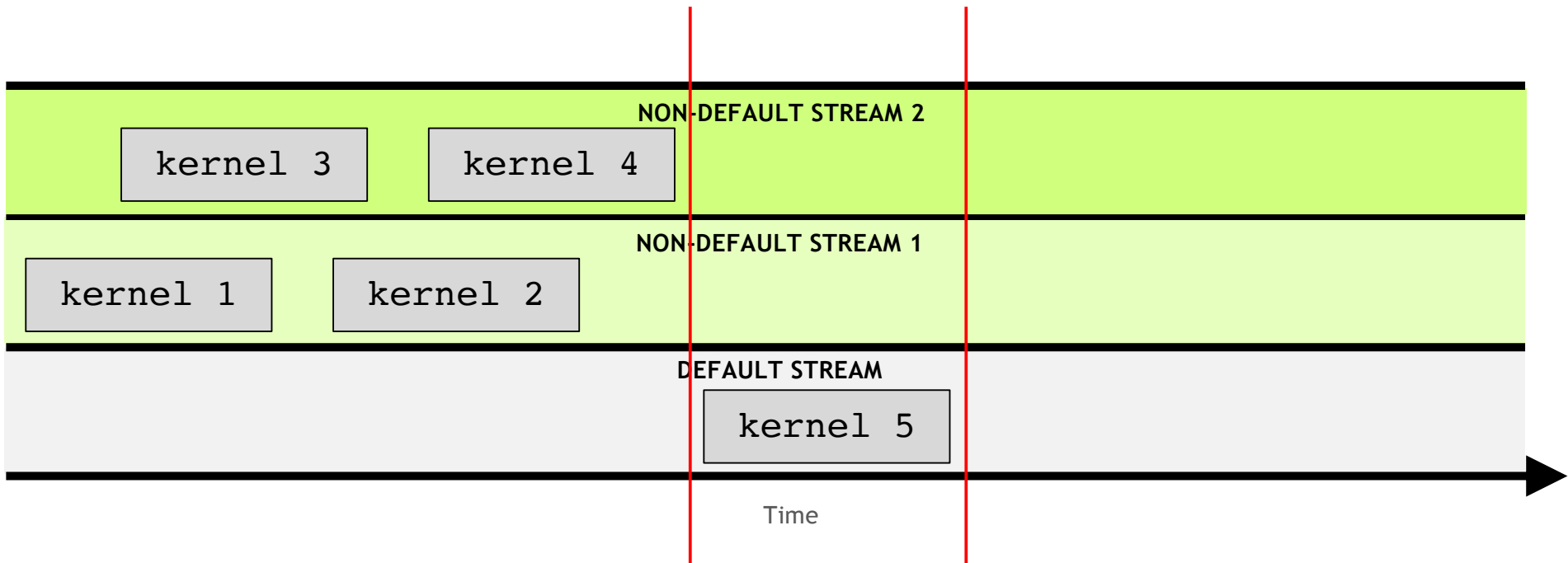
The default stream is special: **it blocks all kernels in all other streams**



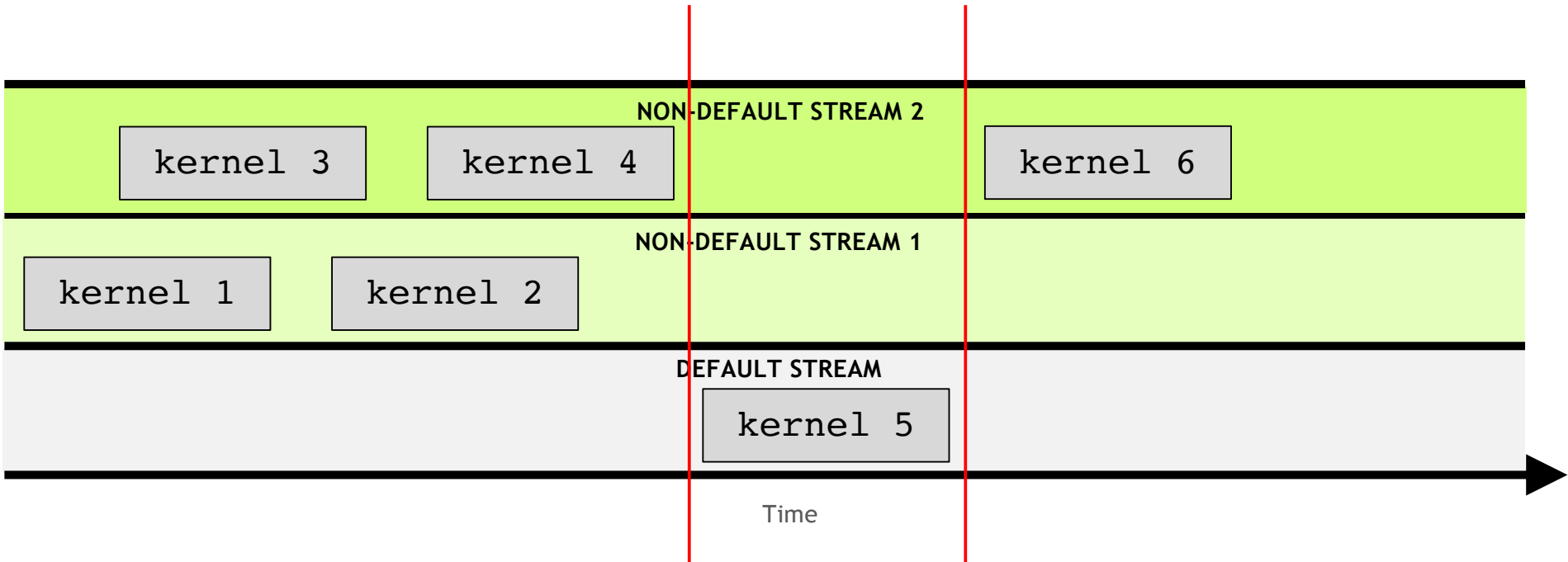
The default stream is special: **it blocks all kernels in all other streams**



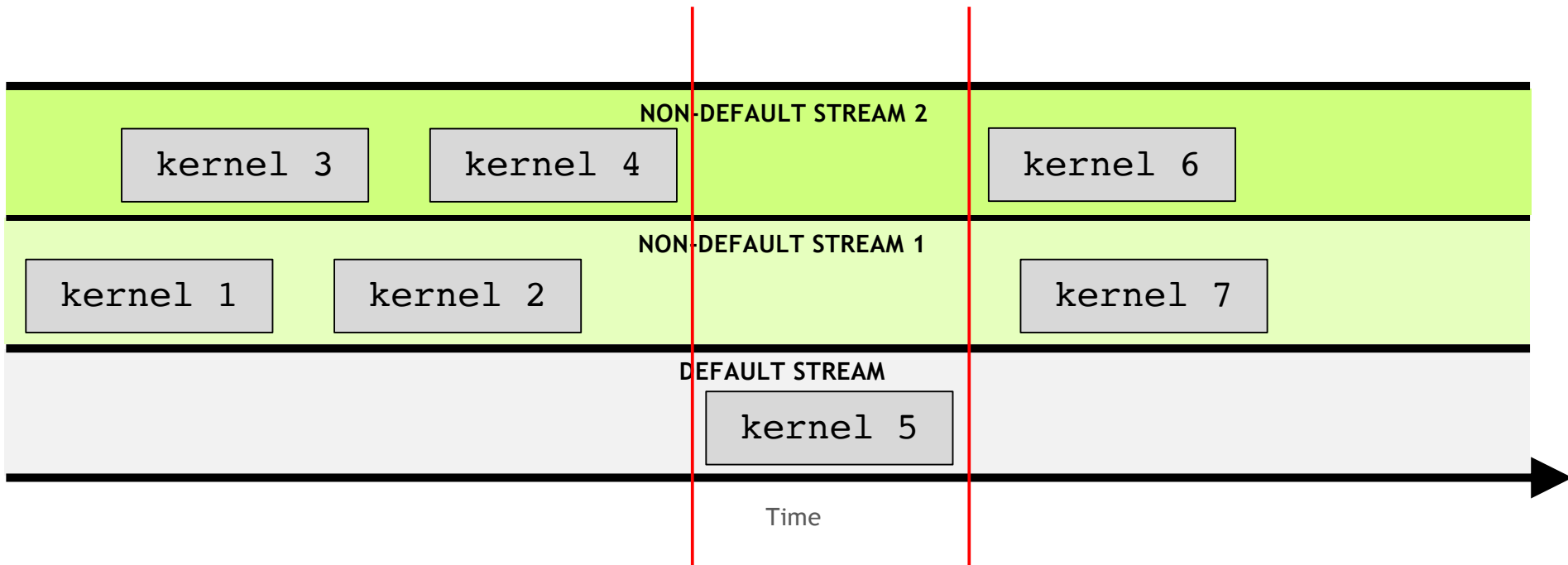
The default stream is special: **it blocks all kernels in all other streams**



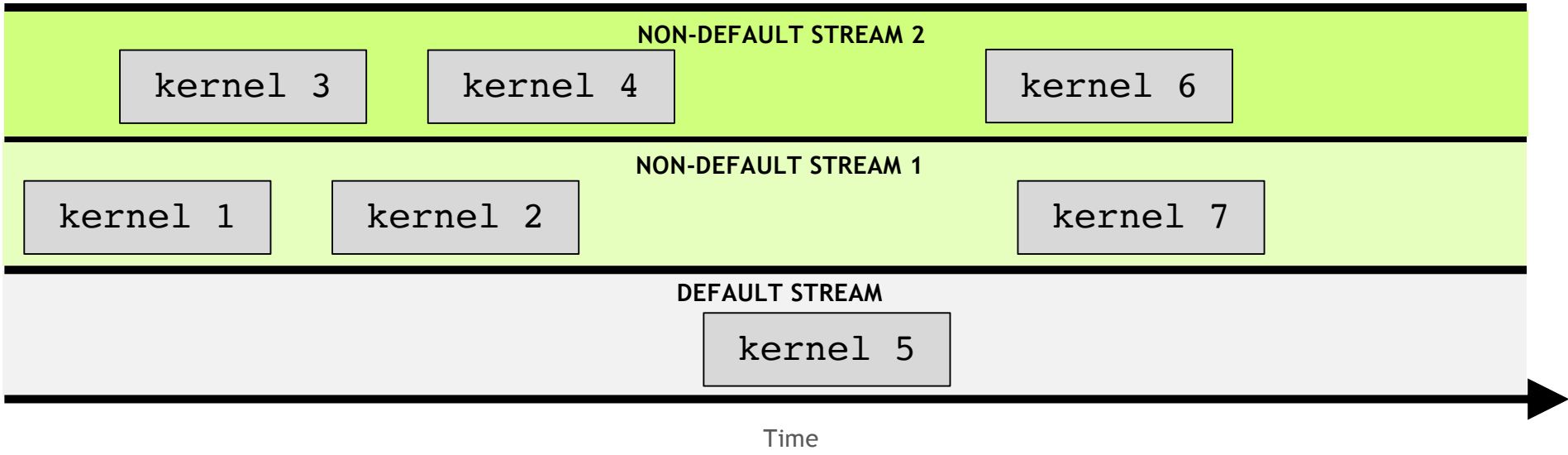
The default stream is special: **it blocks all kernels in all other streams**



The default stream is special: **it blocks all kernels in all other streams**



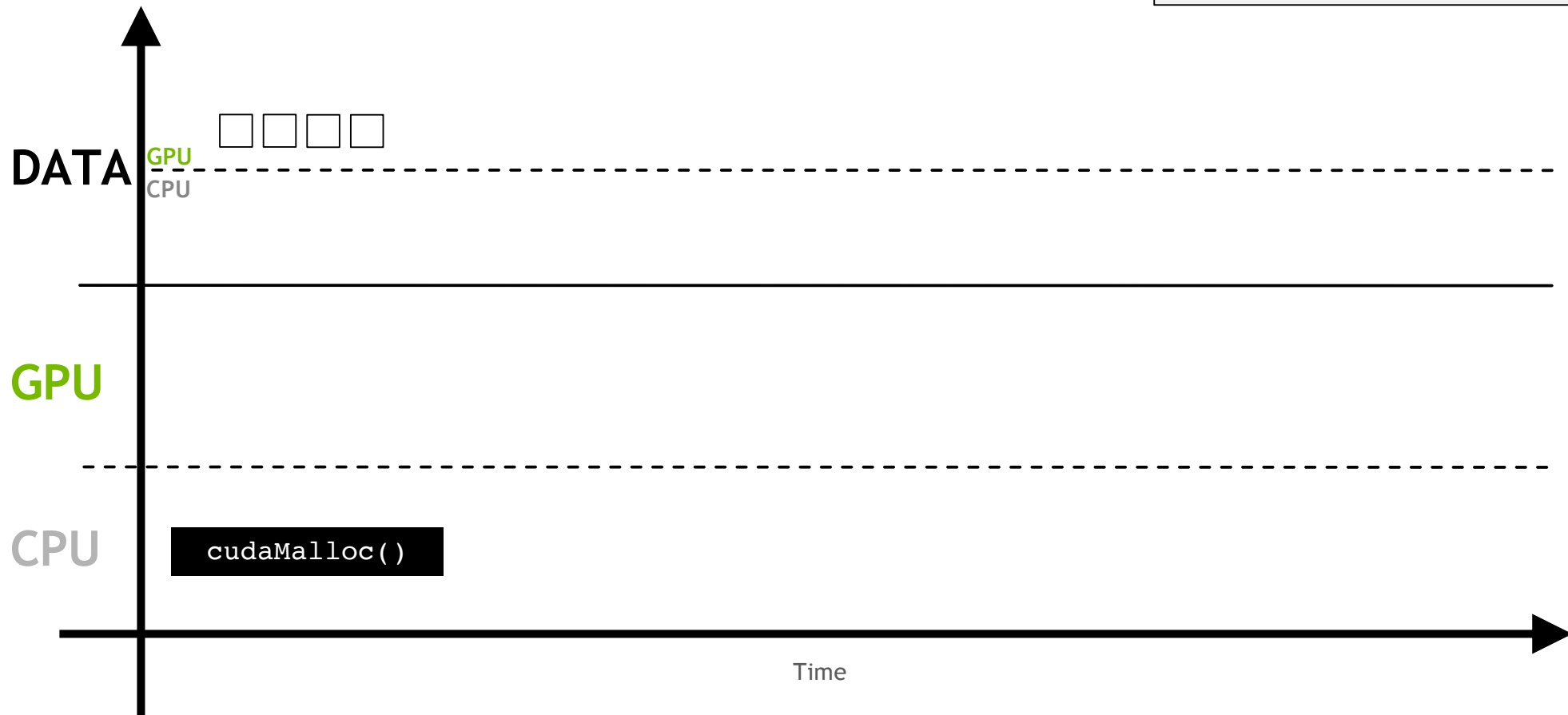
The default stream is special: **it blocks all kernels in all other streams**



# Non-Unified Memory

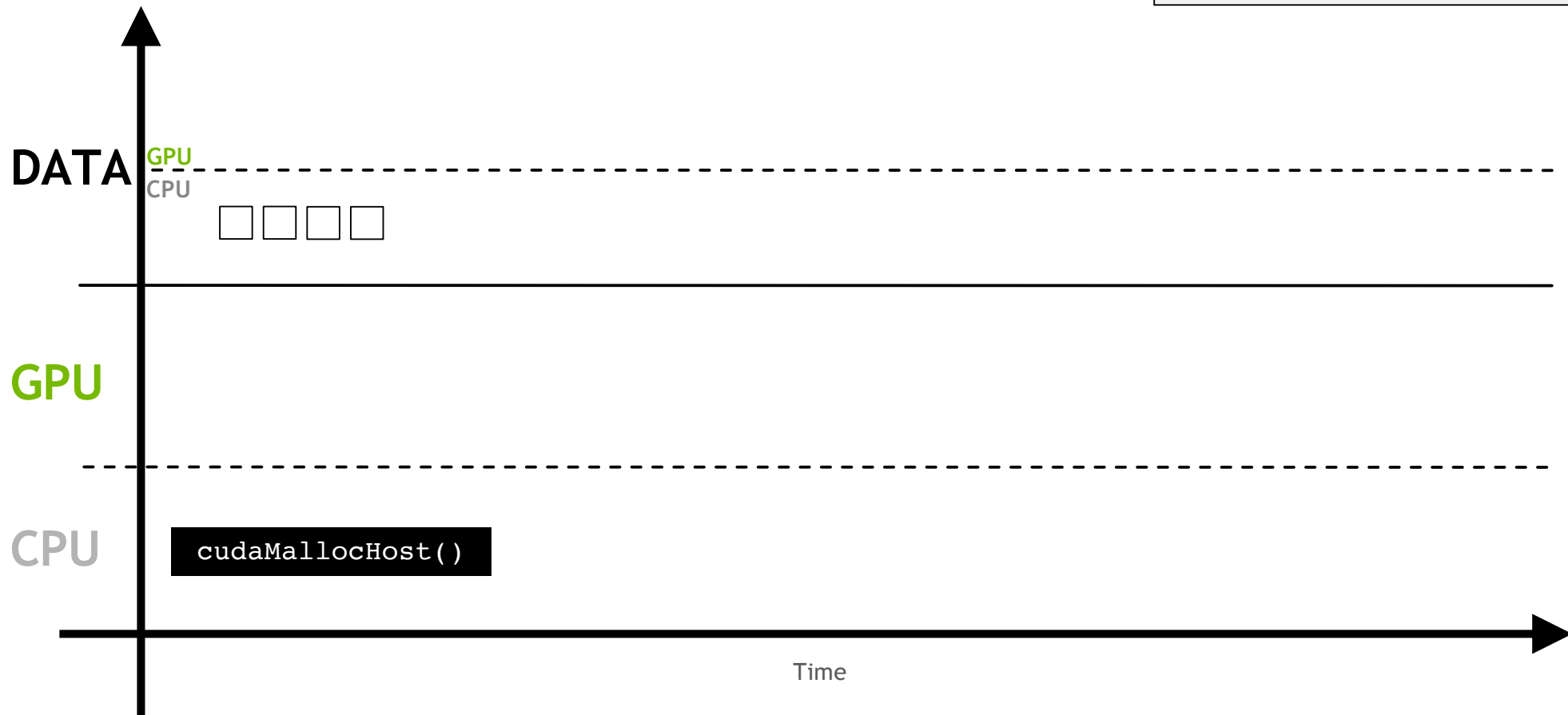
The background of the slide is a solid green color. Overlaid on this background is a complex, white network pattern. This pattern consists of numerous small, interconnected nodes and lines, resembling a mesh or a data network. The nodes are arranged in a somewhat regular grid-like fashion, but the connections between them are irregular, creating a dense, web-like structure. The overall effect is a technical and digital aesthetic.

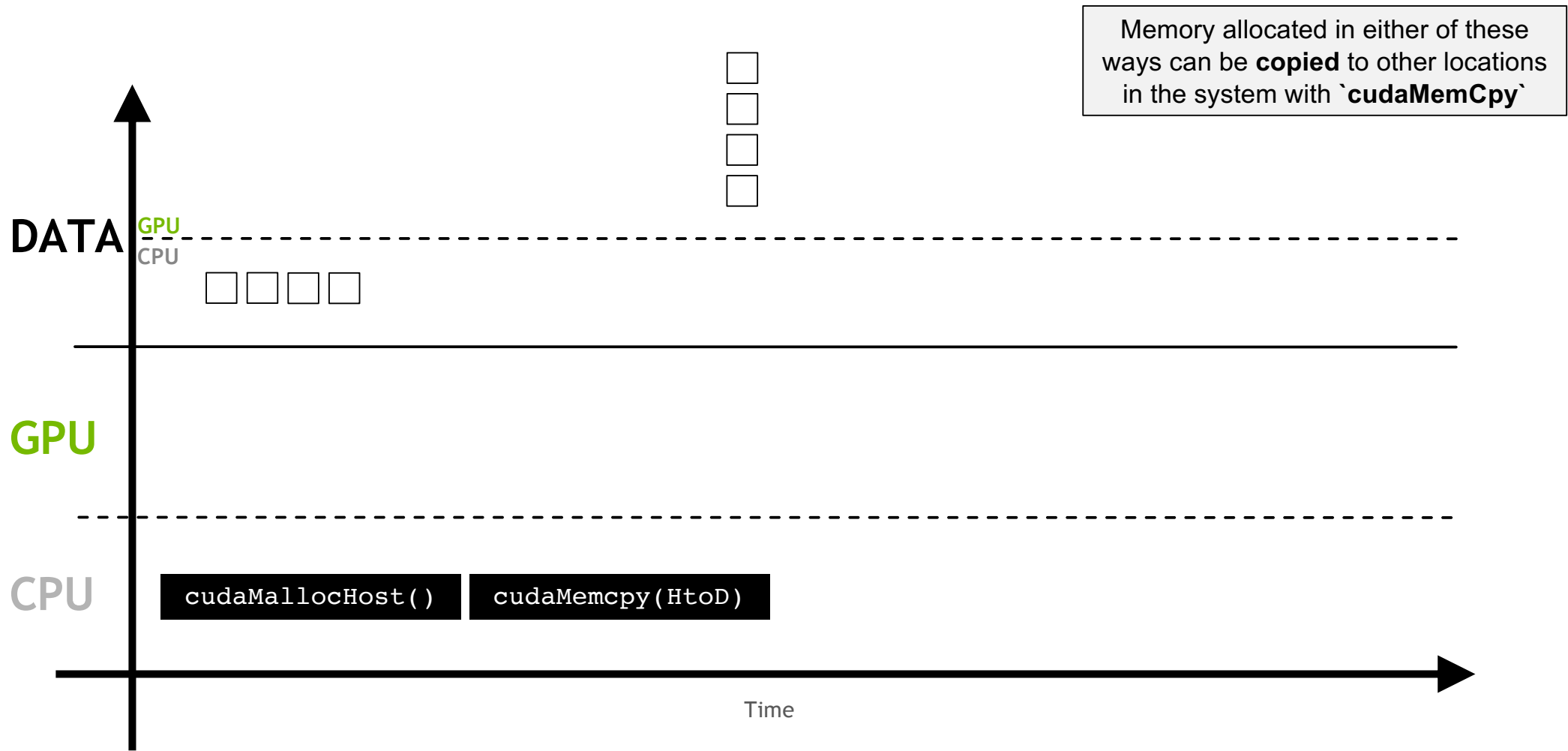
Memory can be allocated directly to the GPU with `cudaMalloc`





Memory can be allocated directly to the host with `cudaMallocHost`



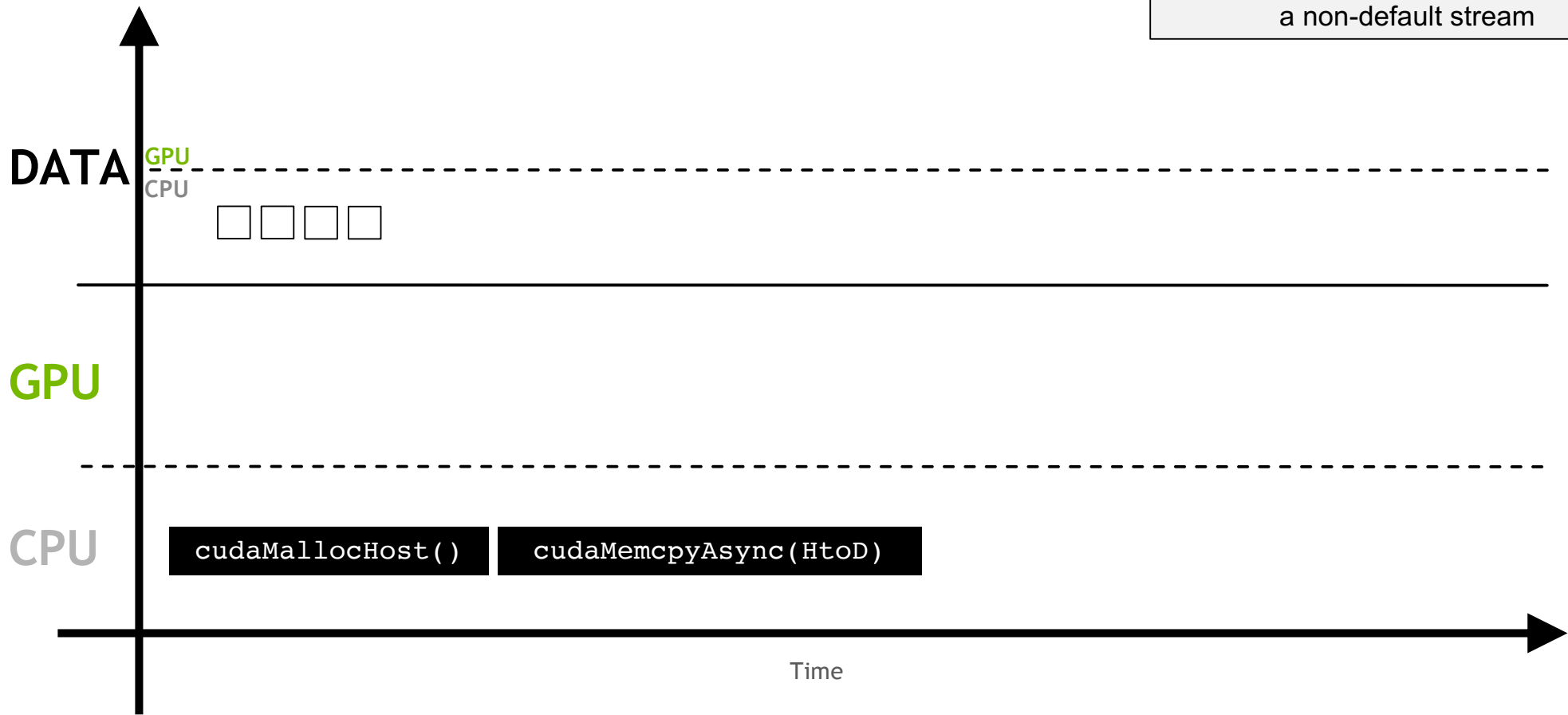






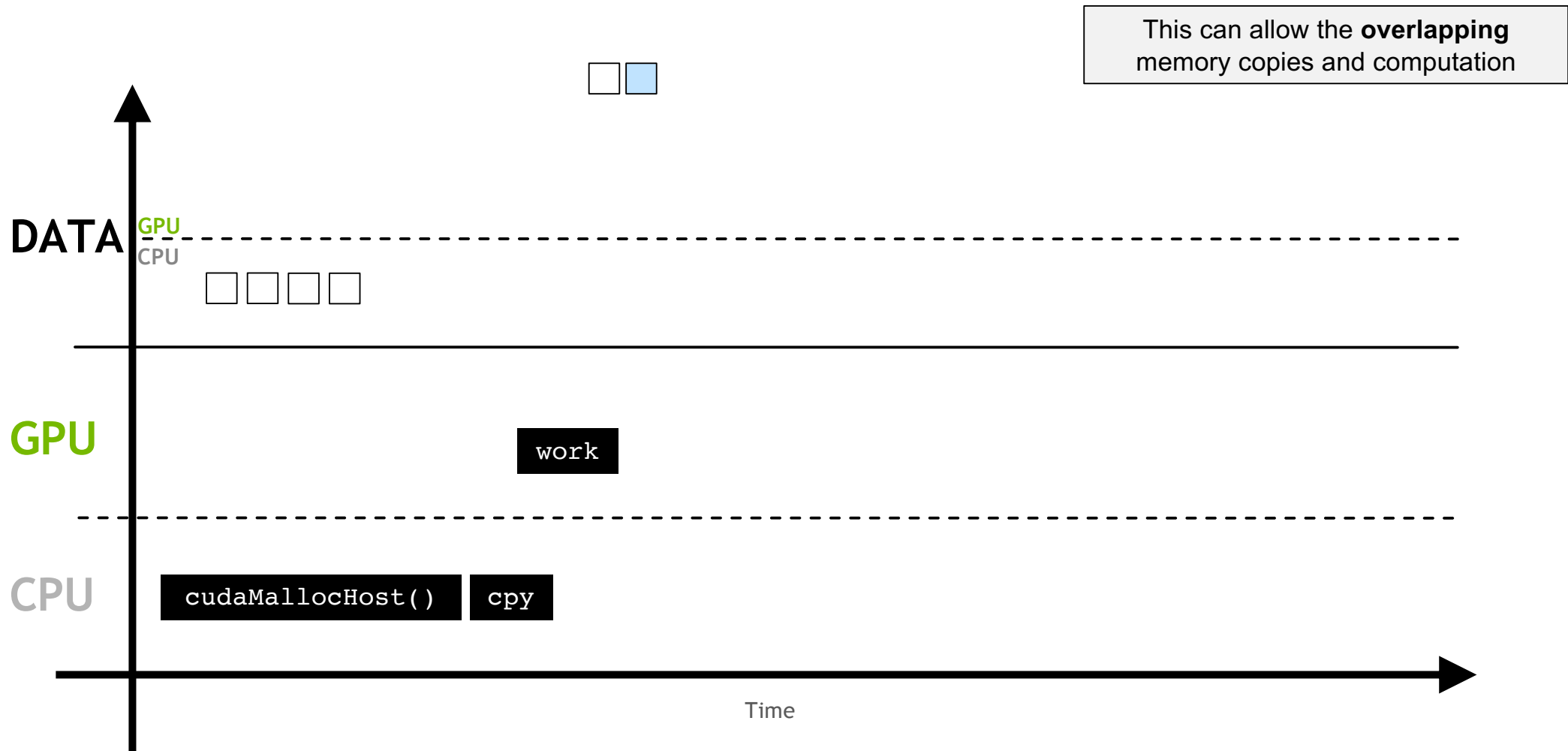
**cudaMemcpyAsync**

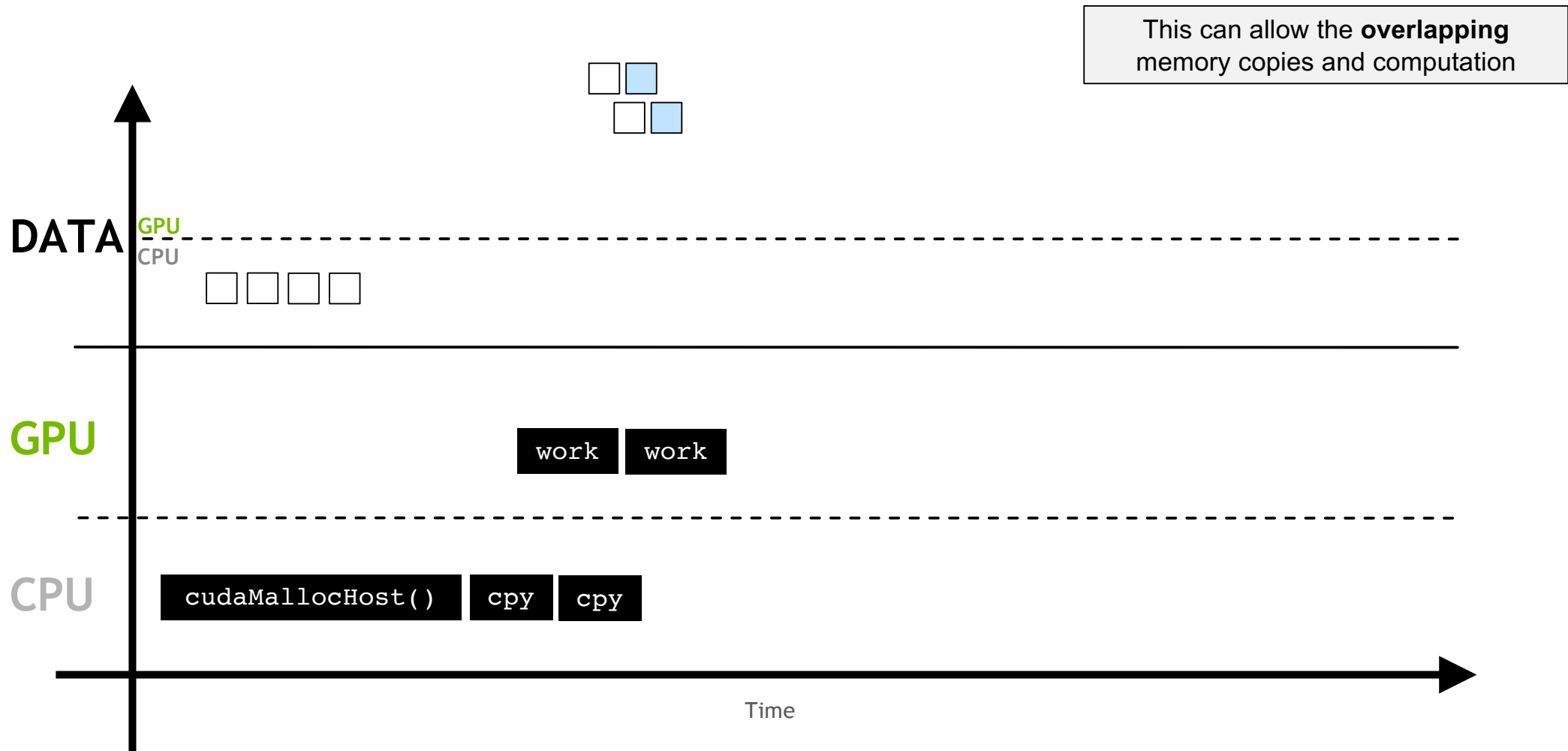
`cudaMemcpyAsync` can asynchronously transfer memory over a non-default stream



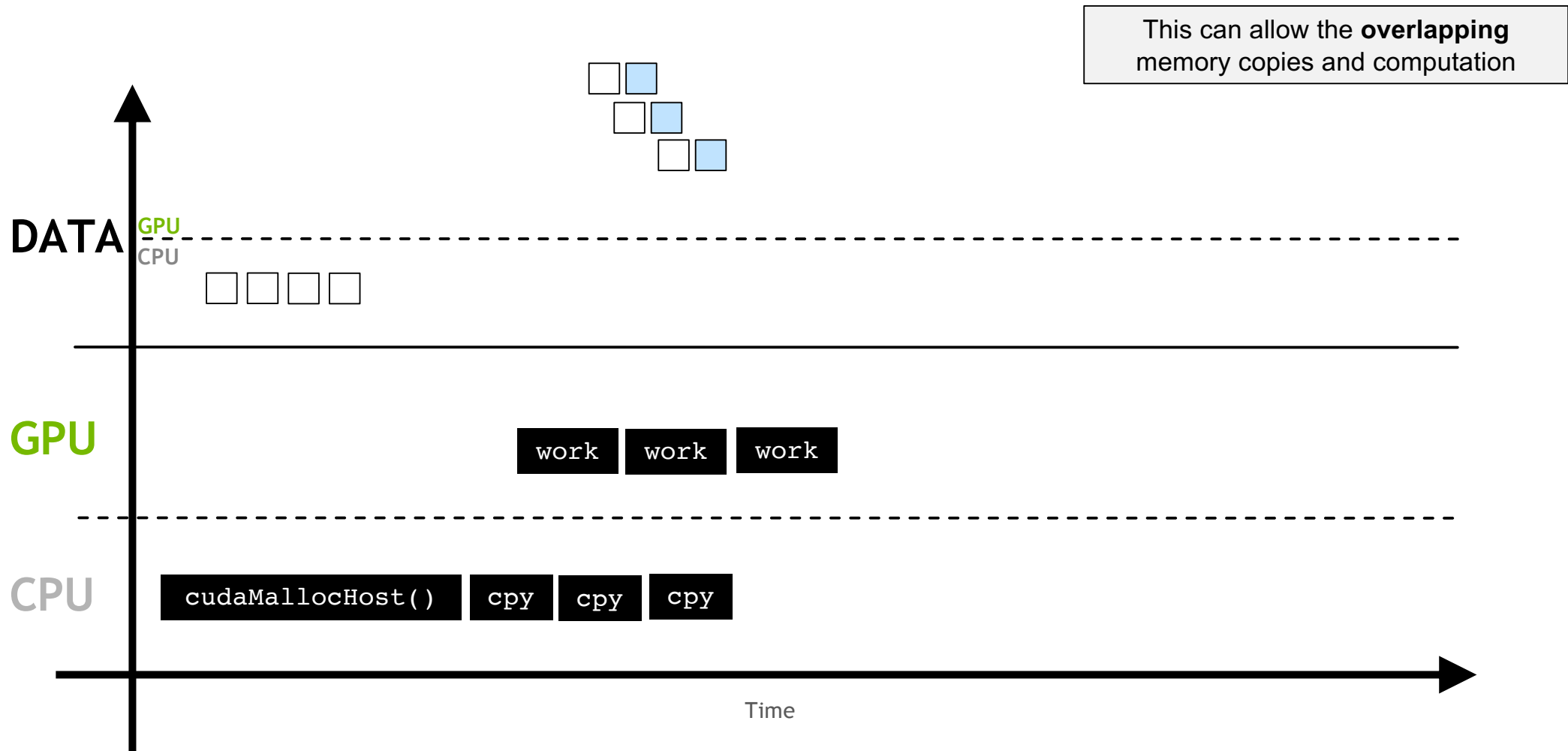
This can allow the **overlapping** memory copies and computation

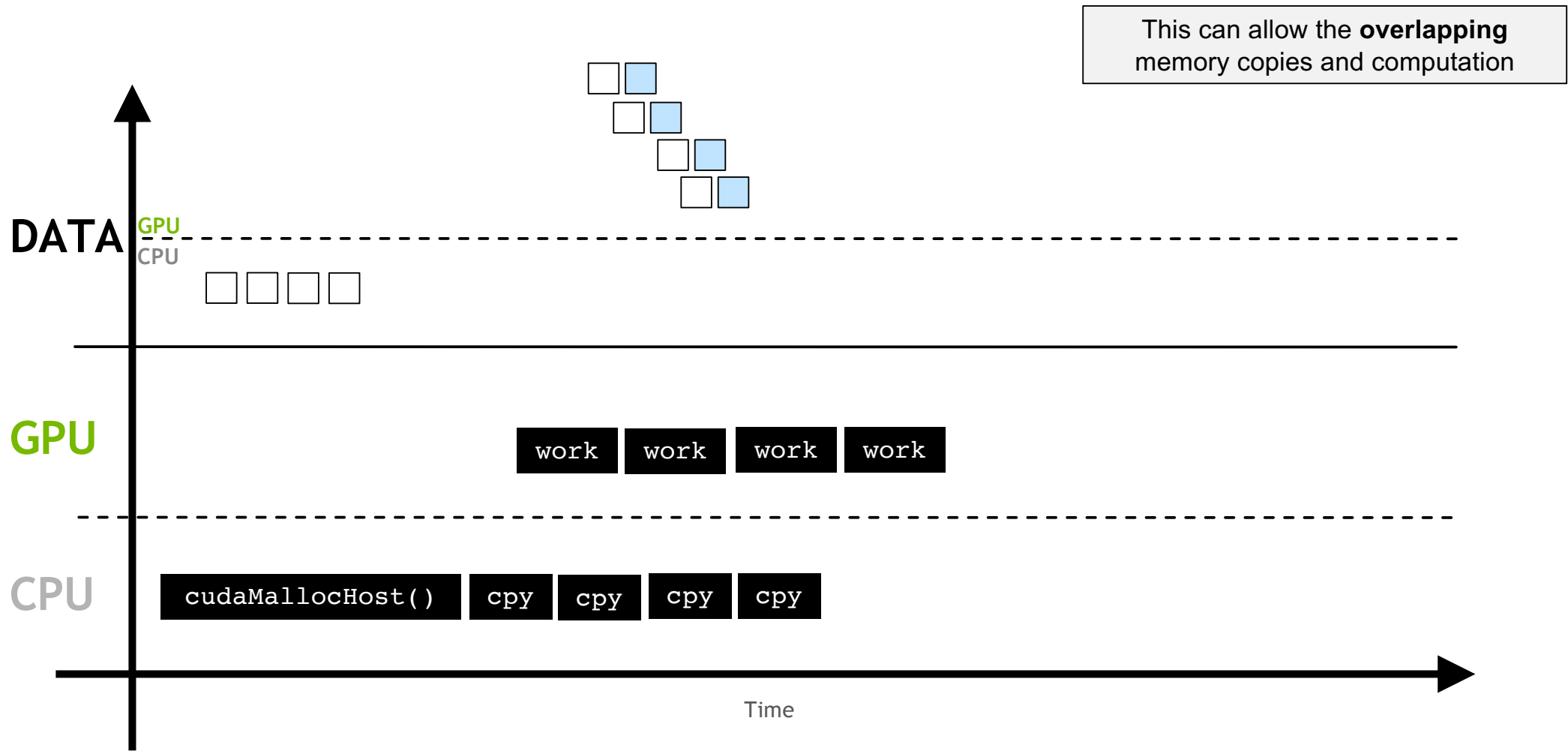














DEEP  
LEARNING  
INSTITUTE

[www.nvidia.com/dli](http://www.nvidia.com/dli)