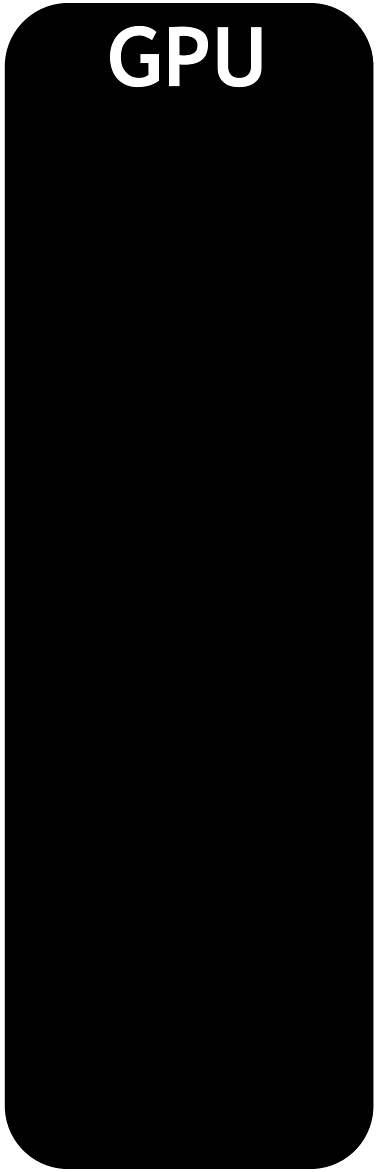
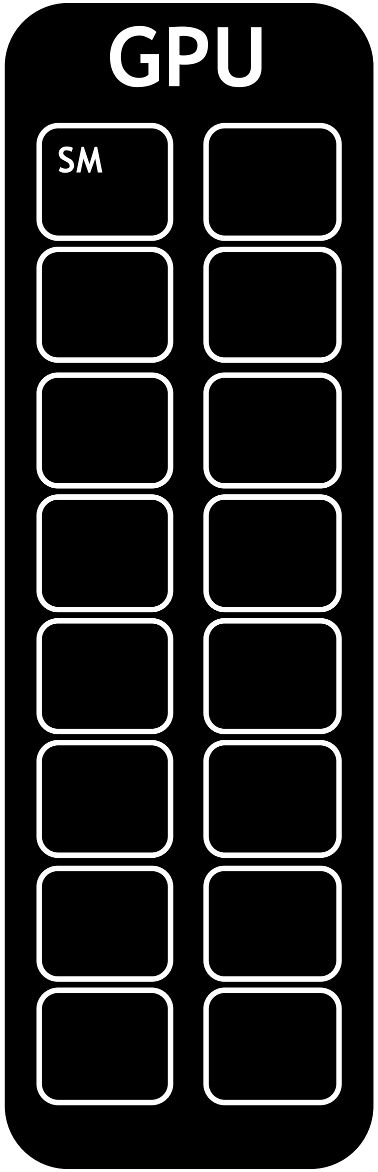
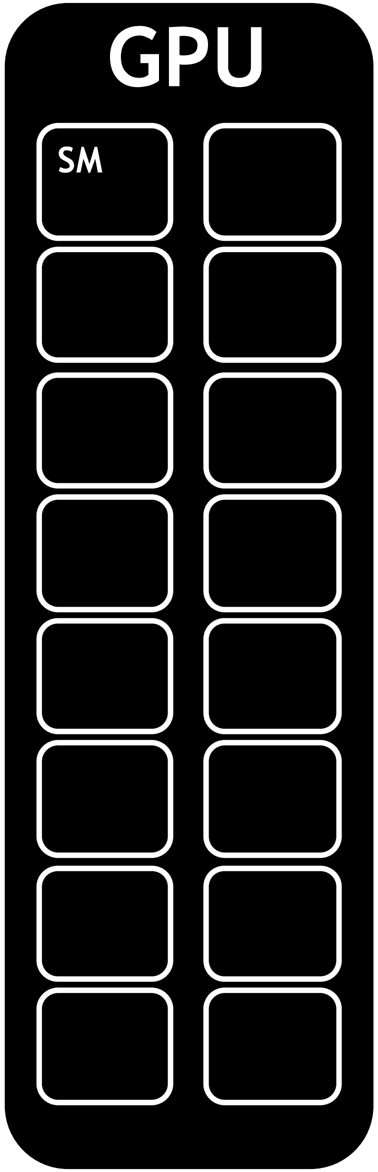
# Streaming Multiprocessors

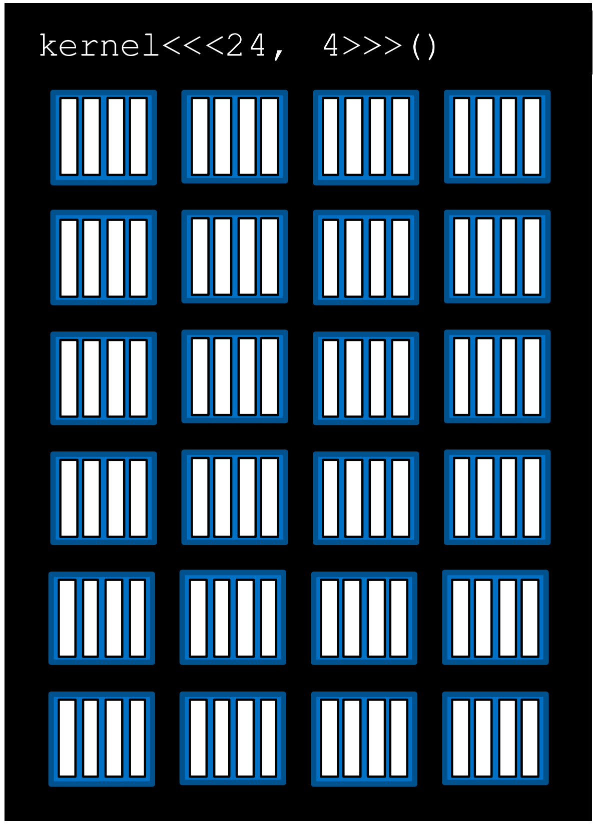


NVIDIA GPUs contain functional units called **Streaming Multiprocessors**, or **SMs**

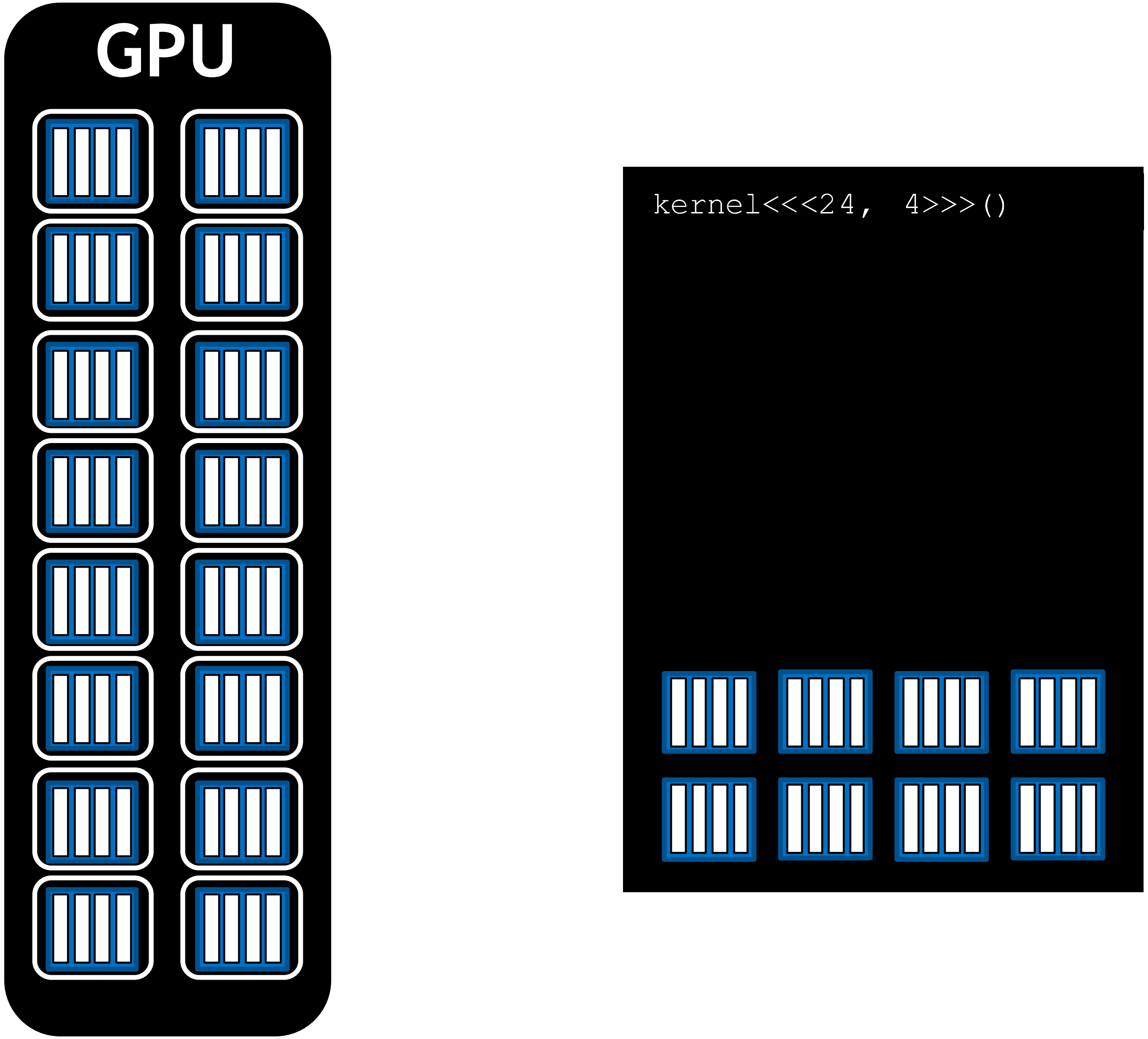


NVIDIA GPUs contain functional units called **Streaming Multiprocessors**, or **SMs**

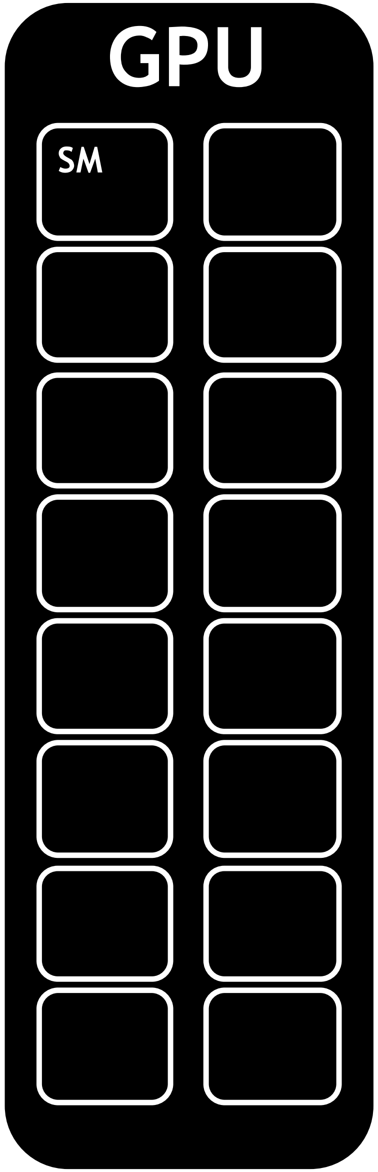


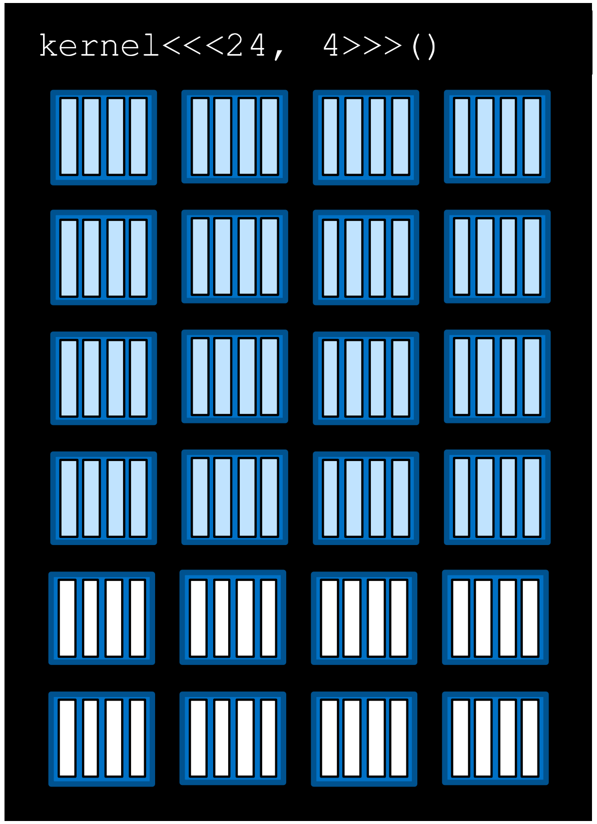

```
kernel<<<24, 4>>>()
```

Blocks of threads are scheduled to run on SMs



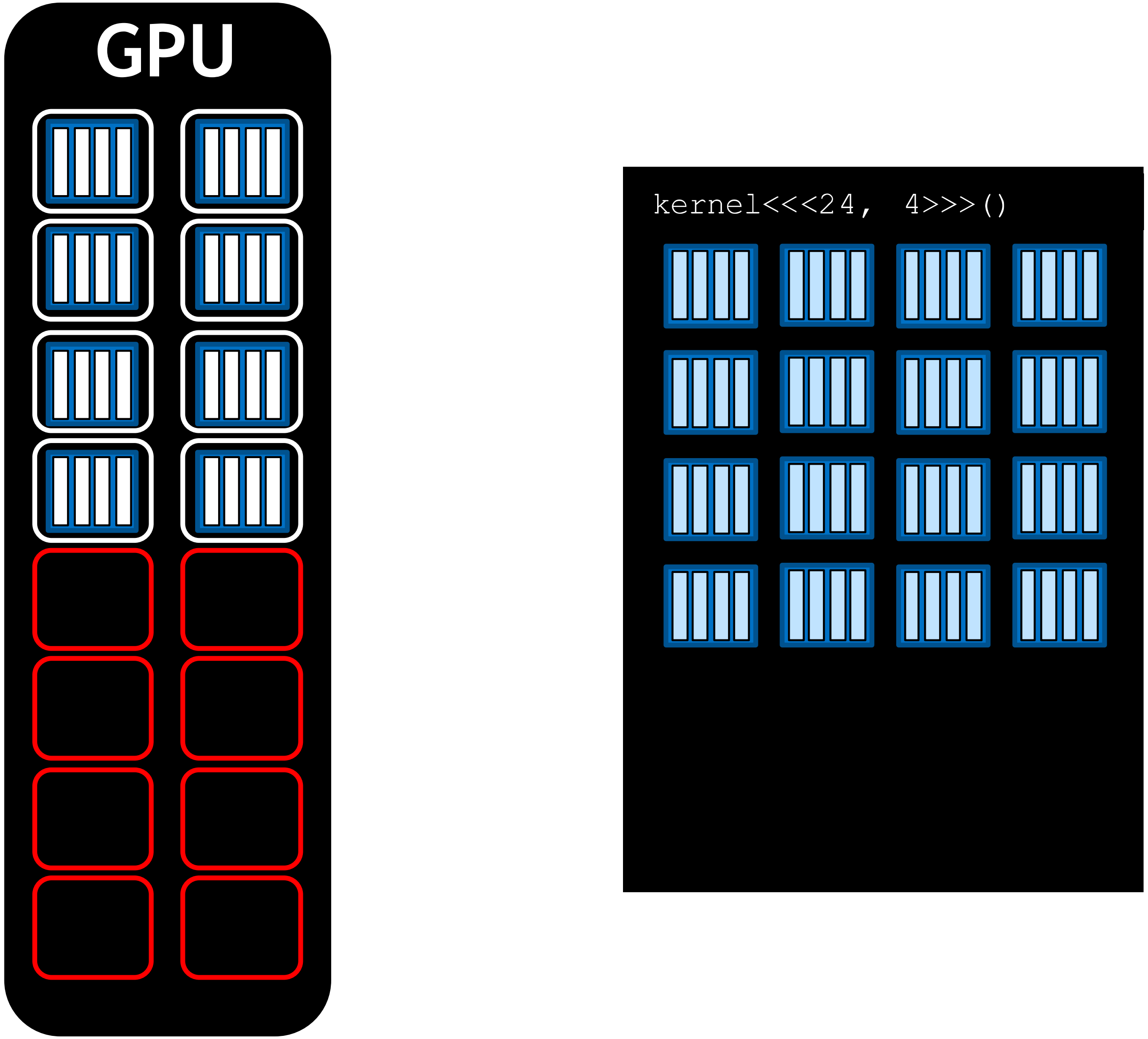
Depending on the number of SMs on a GPU, and the requirements of a block, more than one block can be scheduled on an SM





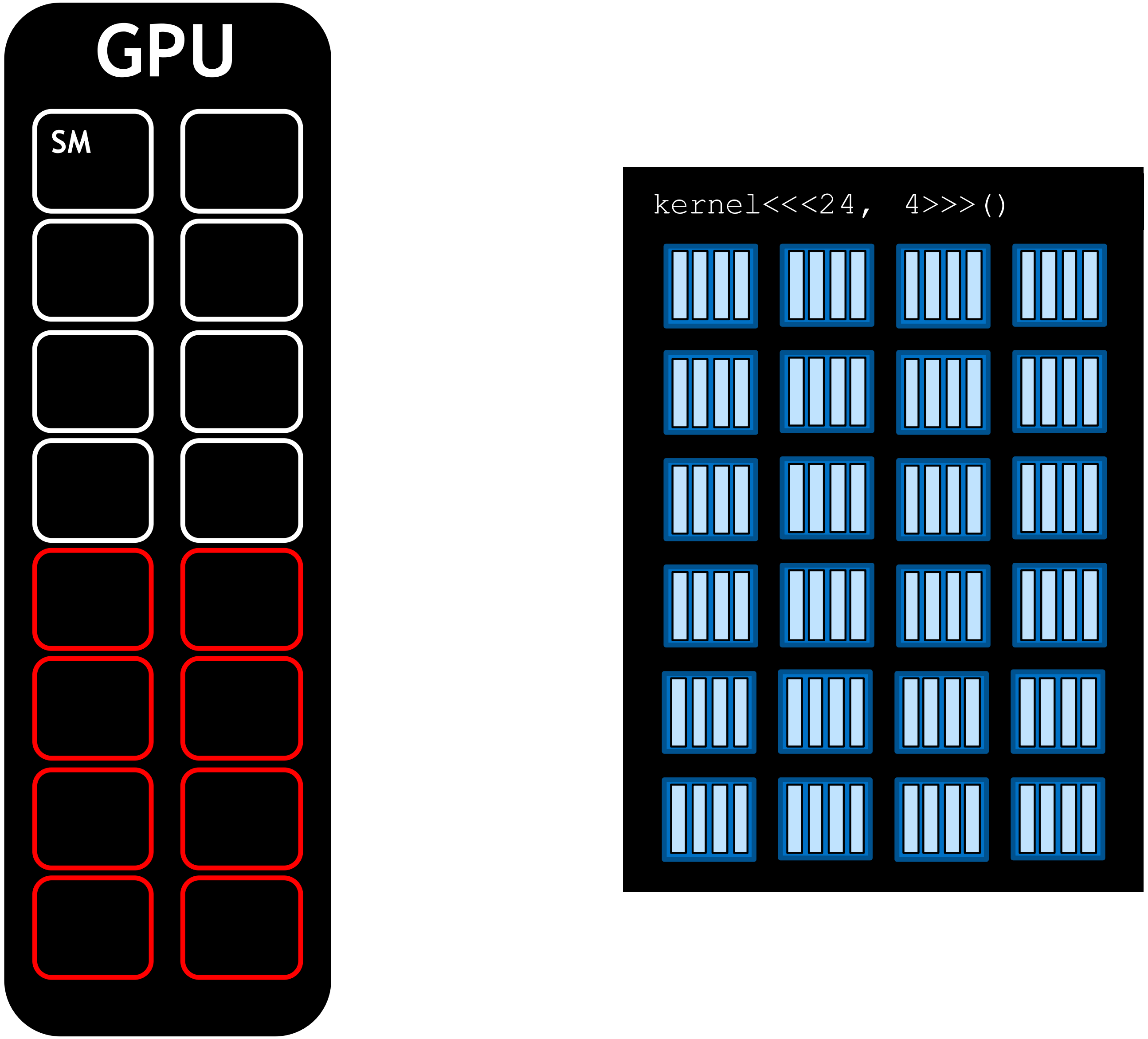
```
kernel<<<24, 4>>>()
```

Depending on the number of SMs on a GPU, and the requirements of a block, more than one block can be scheduled on an SM



```
kernel<<<24, 4>>>()
```

Grid dimensions divisible by the number of SMs on a GPU can promote full SM utilization

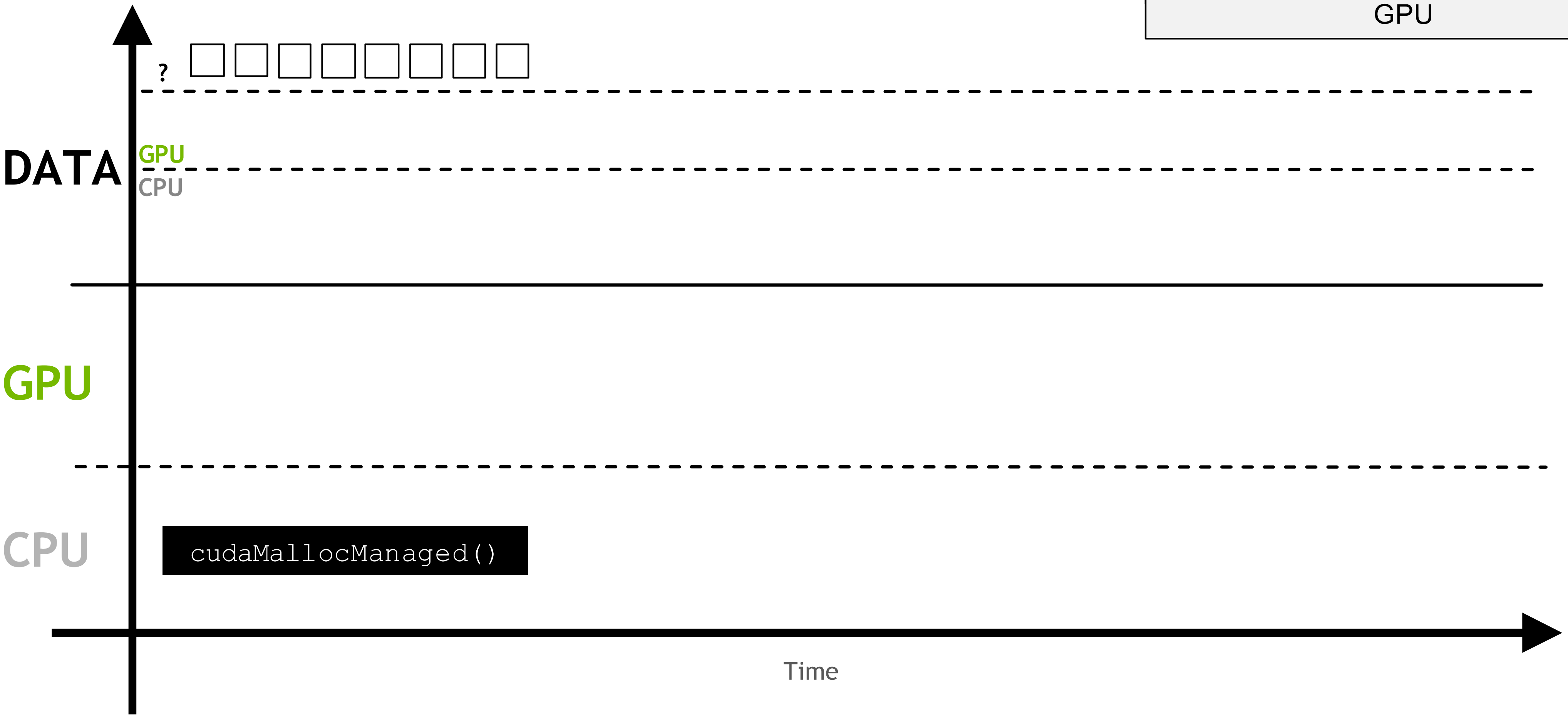


```
kernel<<<24, 4>>>()
```

Here there are fallow SMs

# Unified Memory Behavior

When **UM** is allocated, it may not be resident initially on the CPU or the GPU



?

DATA

GPU

CPU

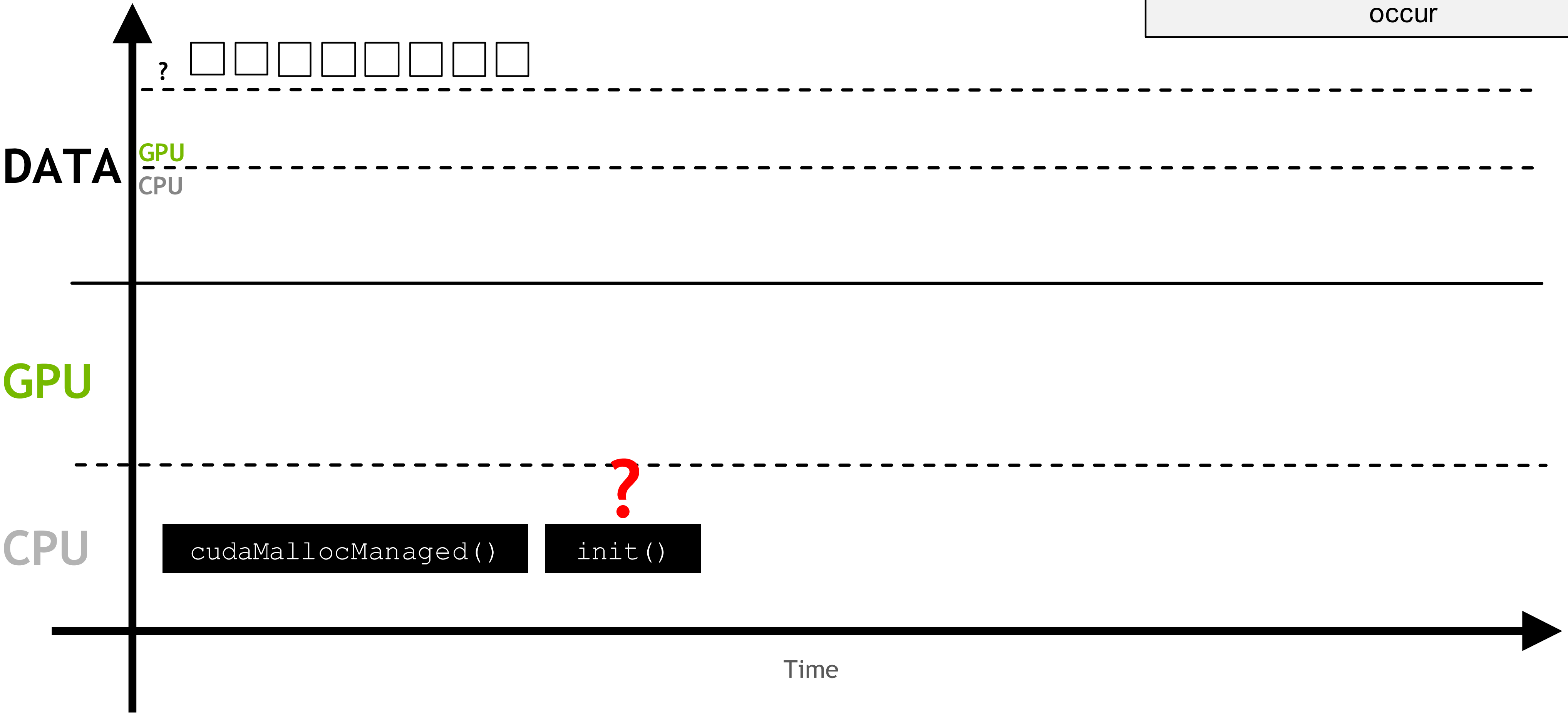
GPU

CPU

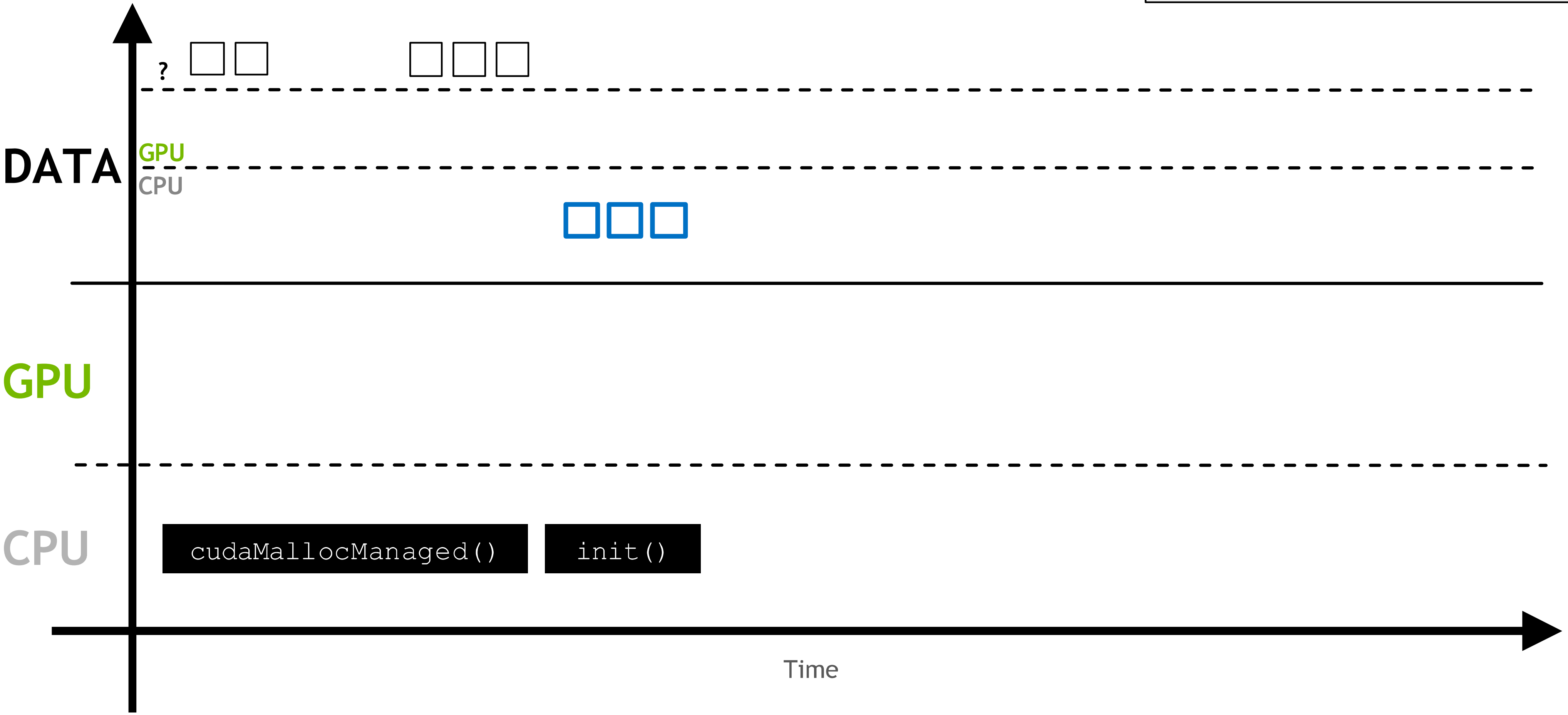
`cudaMallocManaged()`

Time

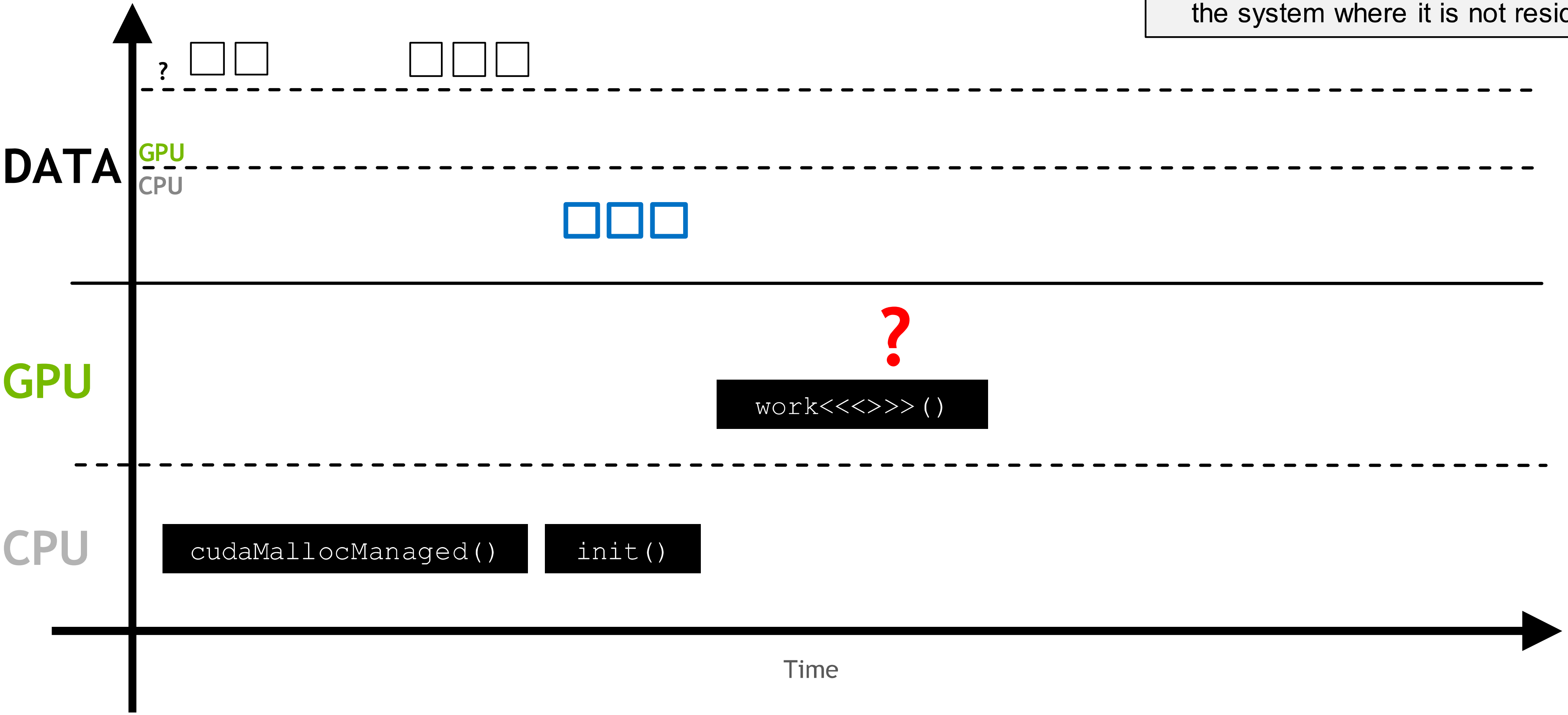
When some work asks for the memory for the first time, a **page fault** will occur



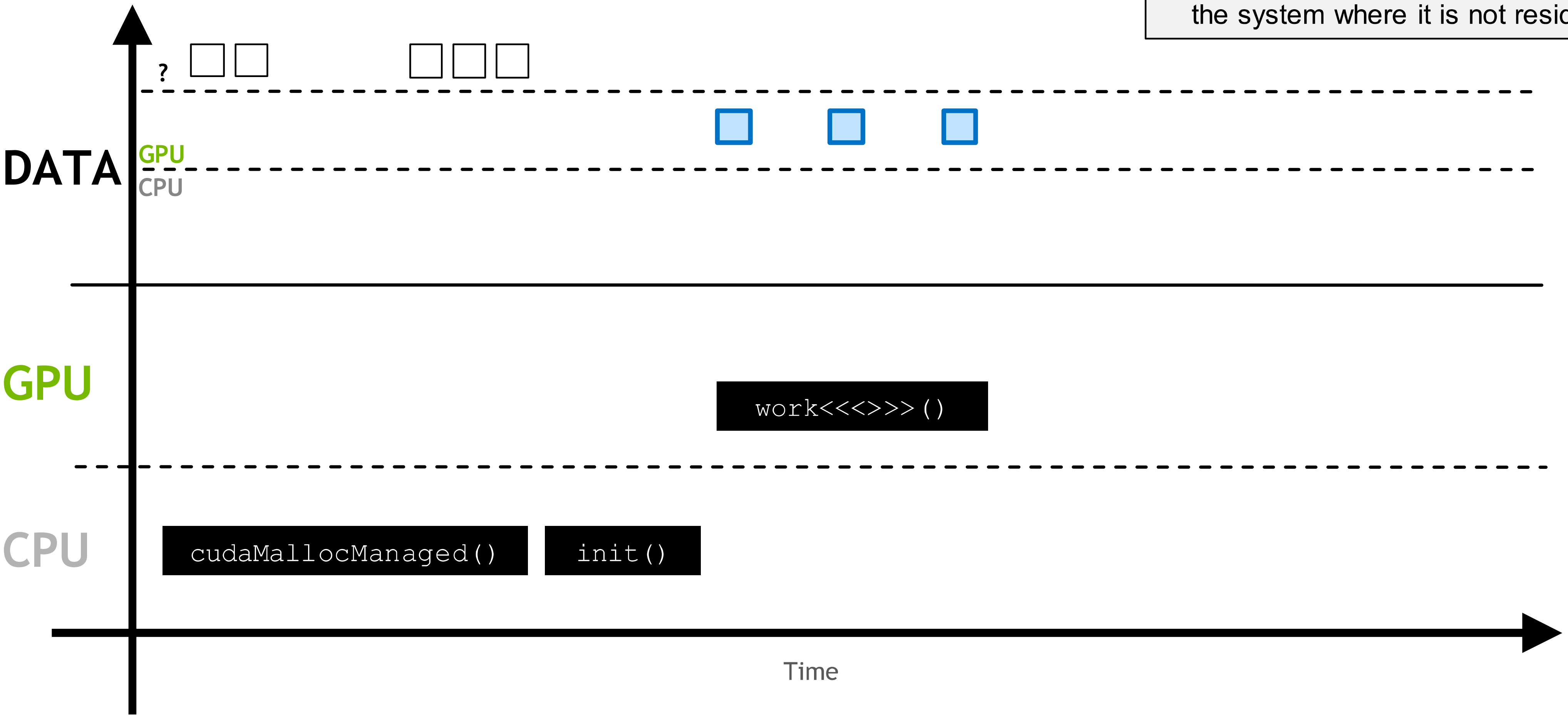
The page fault will trigger the migration of the demanded memory



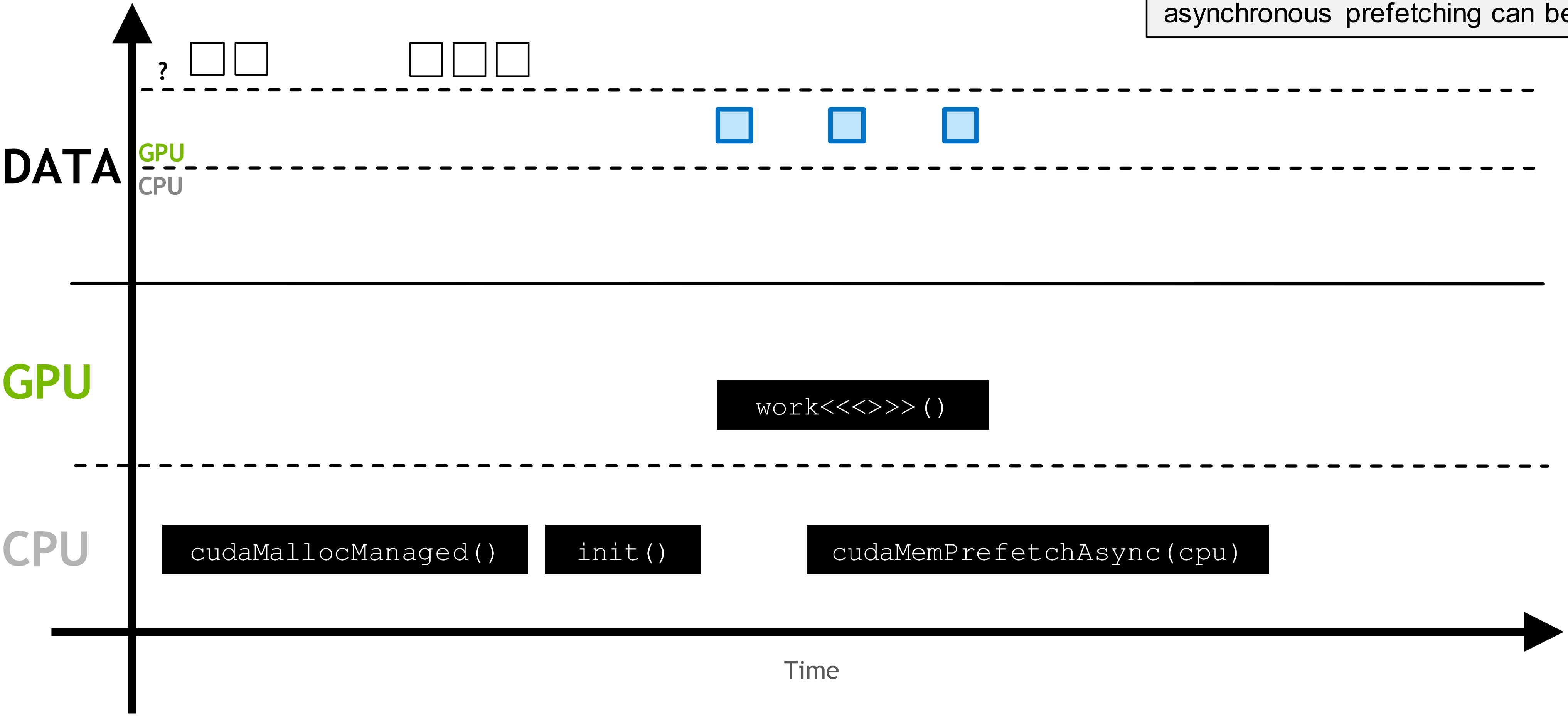
This process repeats anytime the memory is requested somewhere in the system where it is not resident



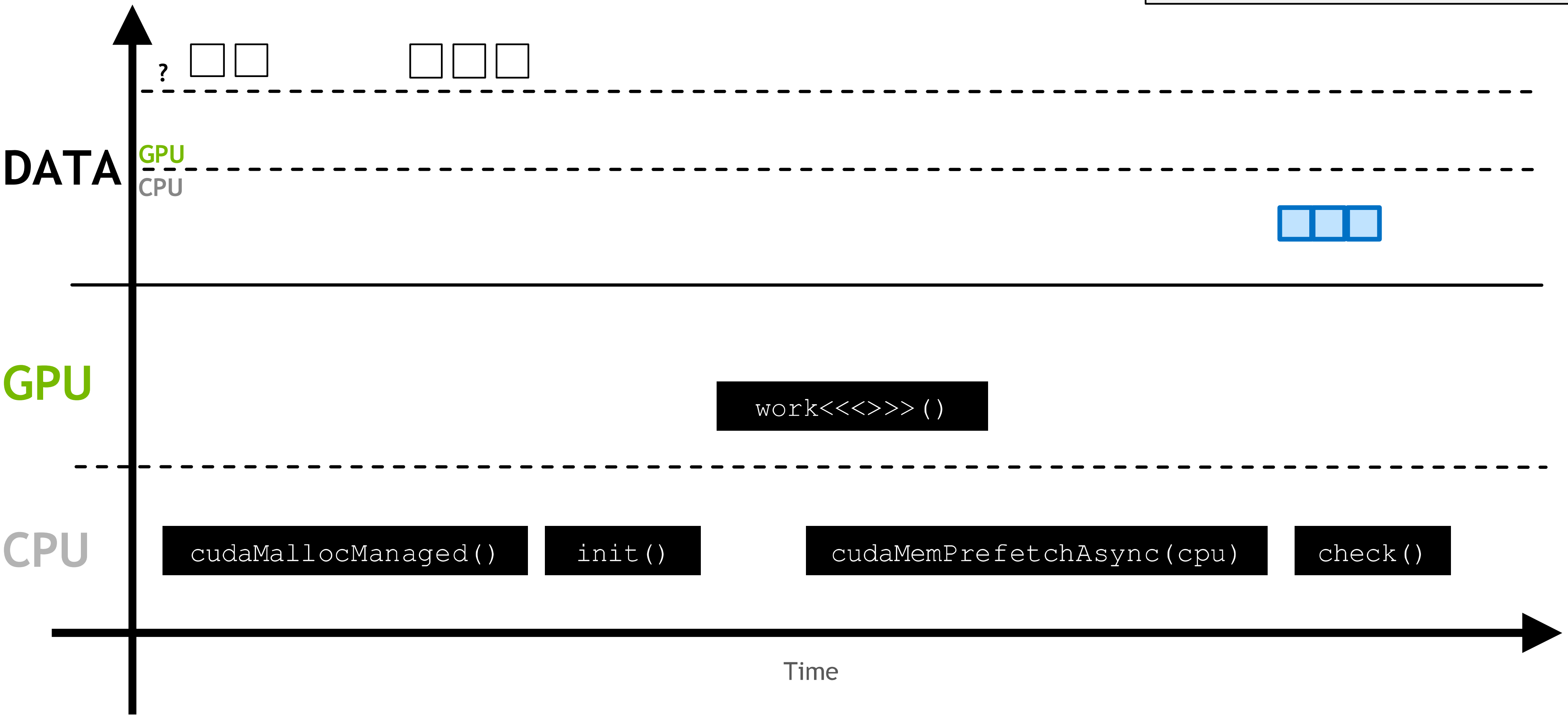
This process repeats anytime the memory is requested somewhere in the system where it is not resident



If it is known that the memory **will be** accessed somewhere it is not resident, asynchronous prefetching can be used



This moves the memory in larger batches, and prevents page faulting



?

DATA

GPU

CPU

GPU

work<<<>>>()

CPU

cudaMallocManaged()

init()

cudaMemPrefetchAsync(cpu)

check()

Time

DEEP
LEARNING
INSTITUTE

www.nvidia.com/dli