NVIDIA DEEP LEARNING INSTITUTE

MULTI-GPU PROGRAMMING FOR CUDA C++

INTRODUCTION

INTRODUCTION

Main Objectives

Concurrency Strategies

Workshop Structure

MAIN OBJECTIVES

MAIN OBJECTIVES

Increase performance for Single-Node CUDA C/C++ applications by exploiting, and then combining, 2 concurrency strategies offered to CUDA programmers.

MAIN OBJECTIVES

Increase performance for Single-Node CUDA C/C++ applications by exploiting, and then combining, 2 concurrency strategies offered to CUDA programmers:

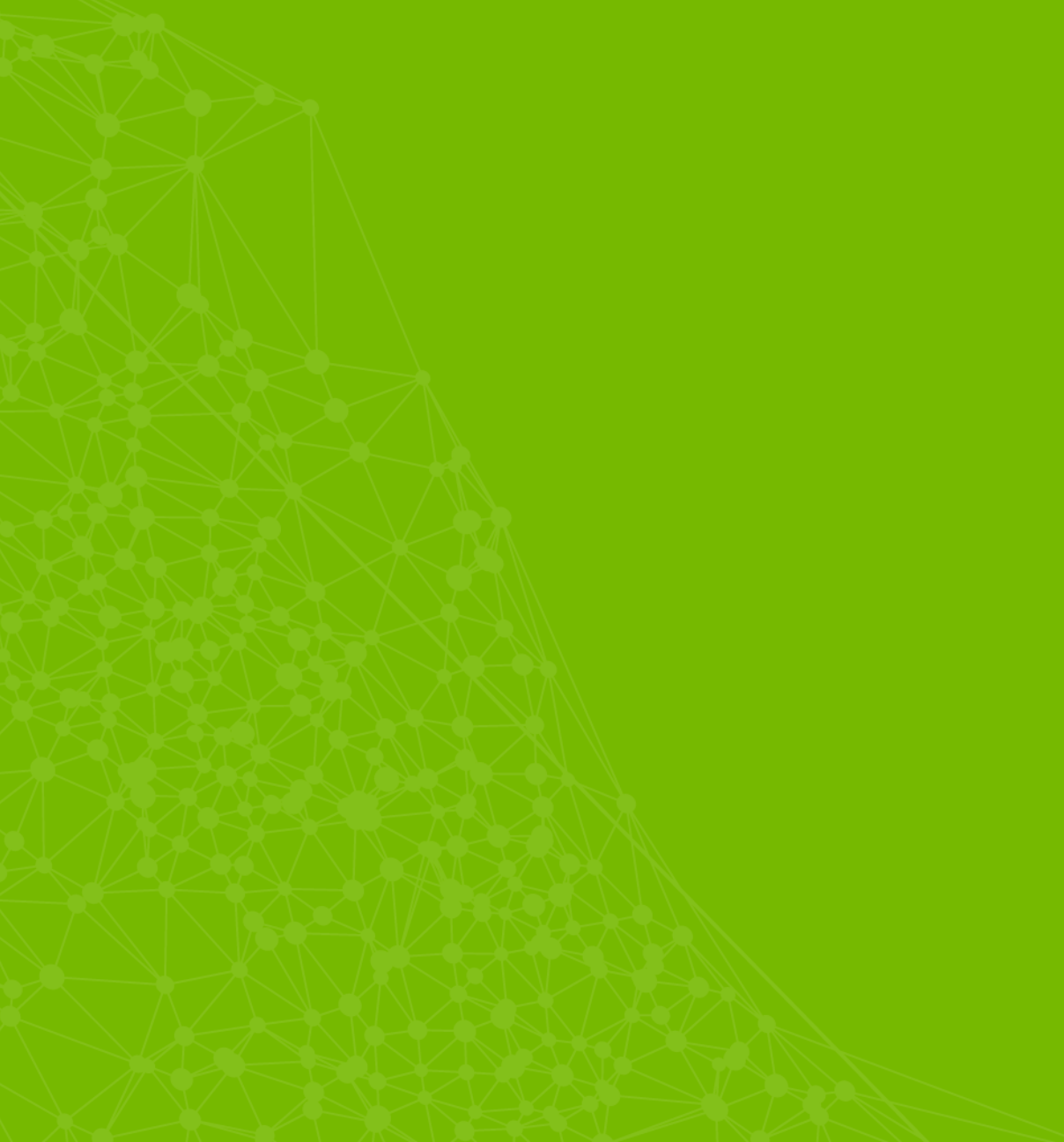
1) Overlapping memory transfers to and from the GPU with computations on the GPU

MAIN OBJECTIVES

Increase performance for Single-Node CUDA C/C++ applications by exploiting, and then combining, 2 concurrency strategies offered to CUDA programmers:

1) Overlapping memory transfers to and from the GPU with computations on the GPU

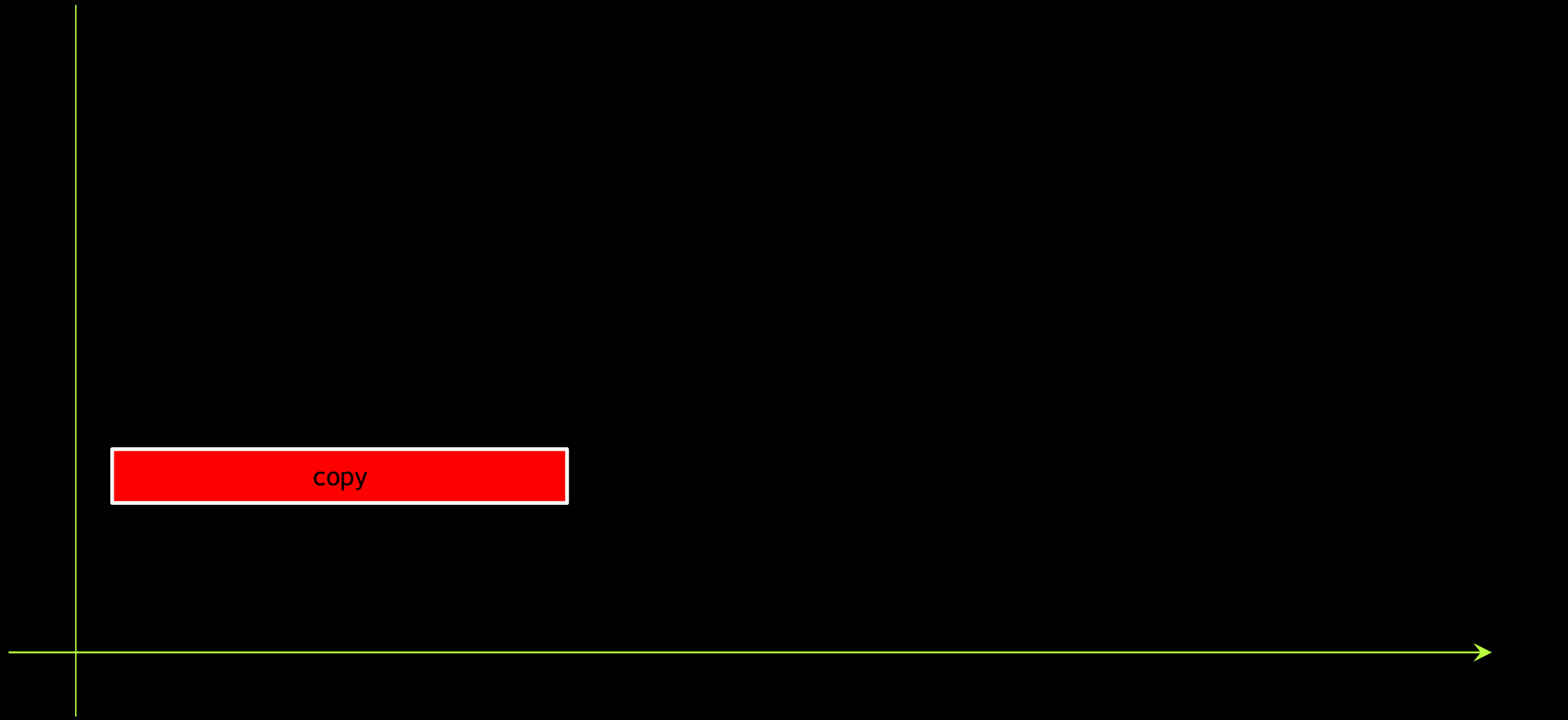
2) Performing computations concurrently on more than one GPU



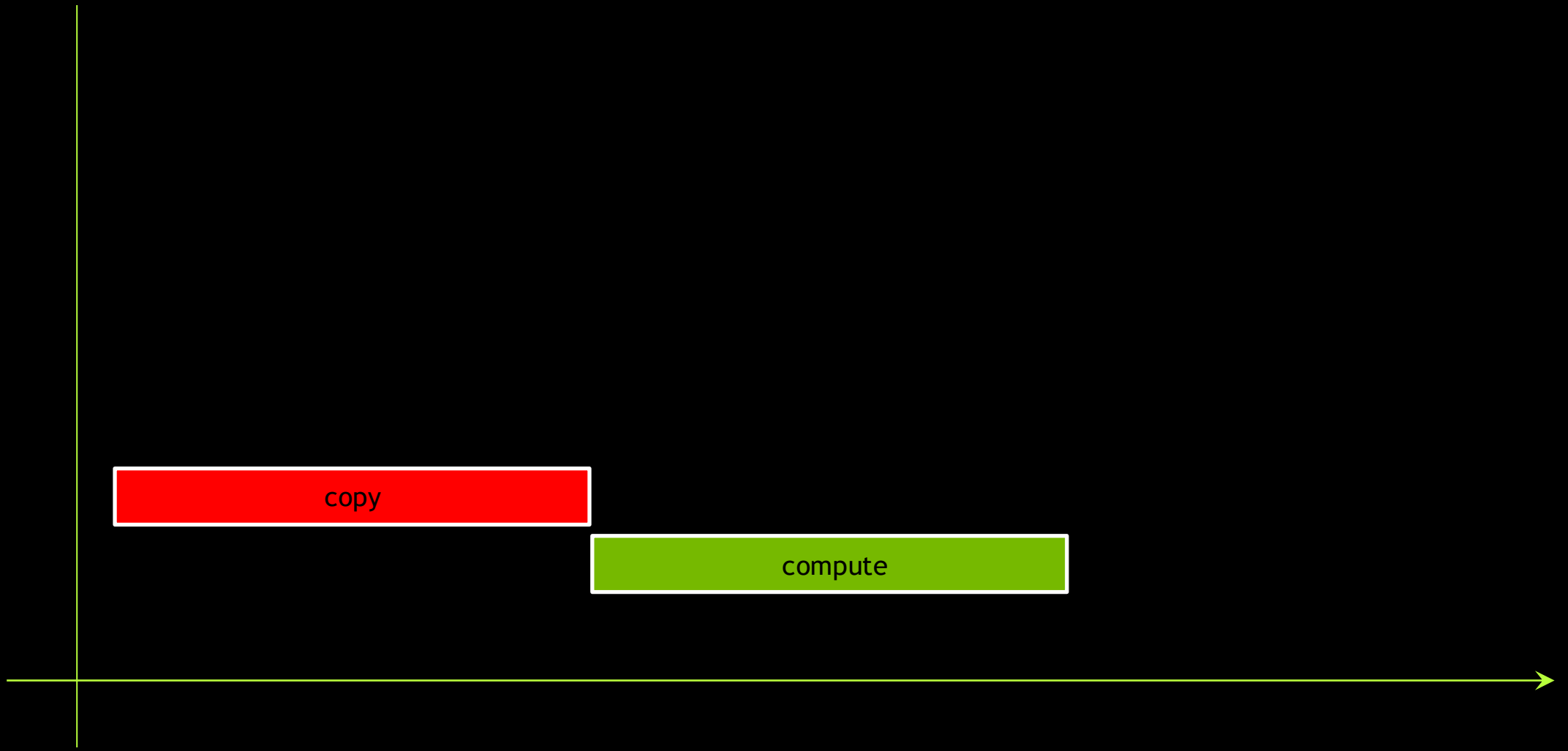
CONCURRENCY STRATEGIES

GPU programming is usually a 3-step process

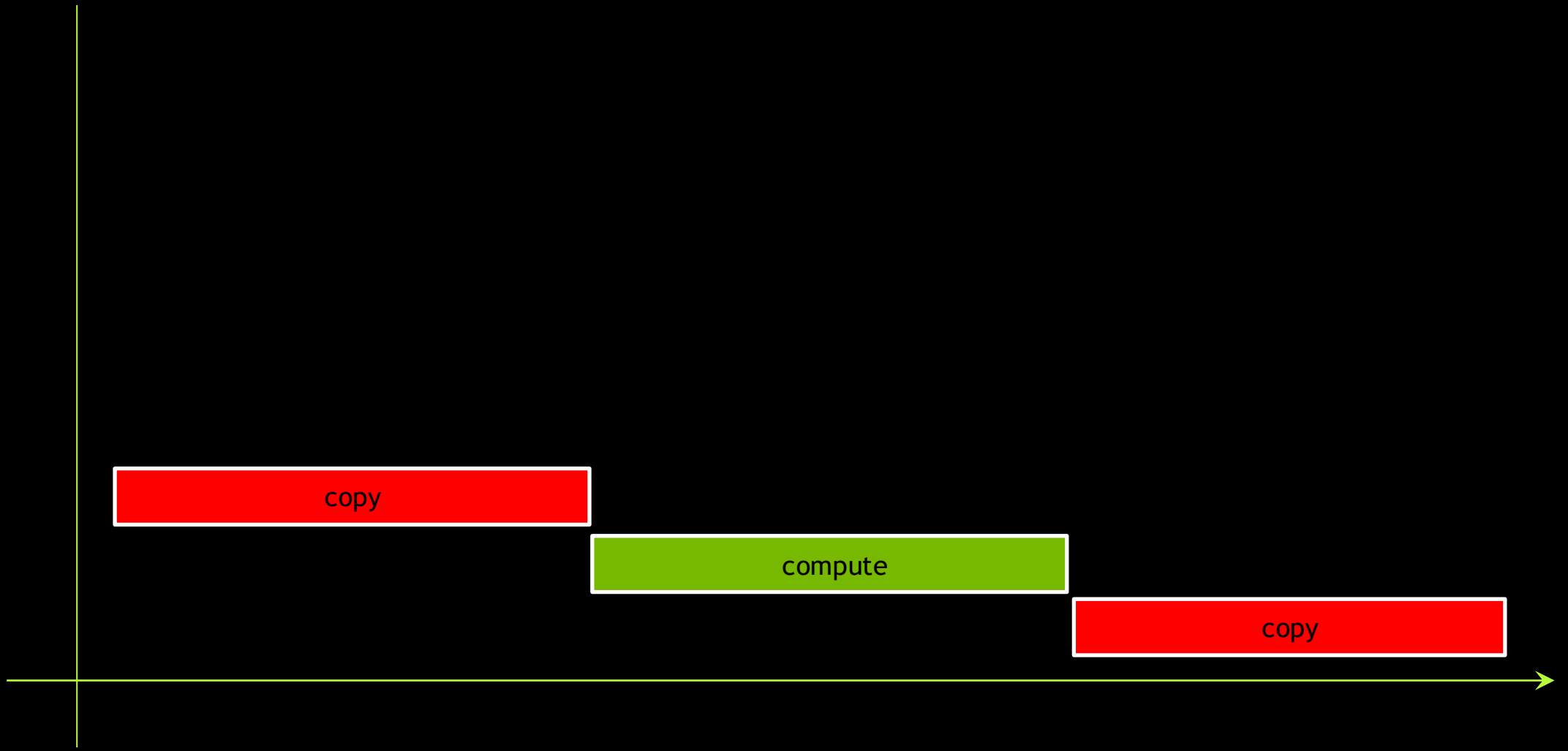
1. Transfer data to GPU device(s)

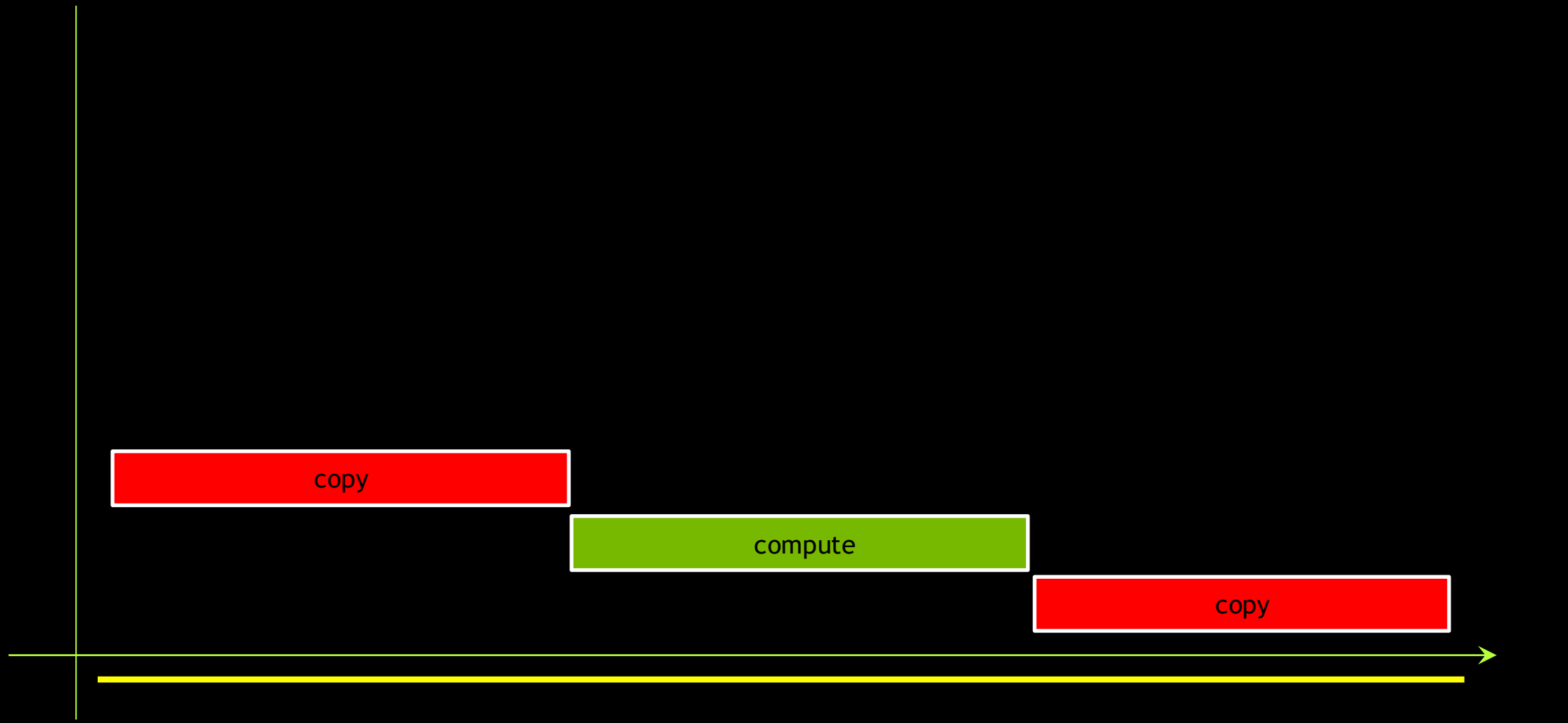


2. Perform computation on GPU device(s)

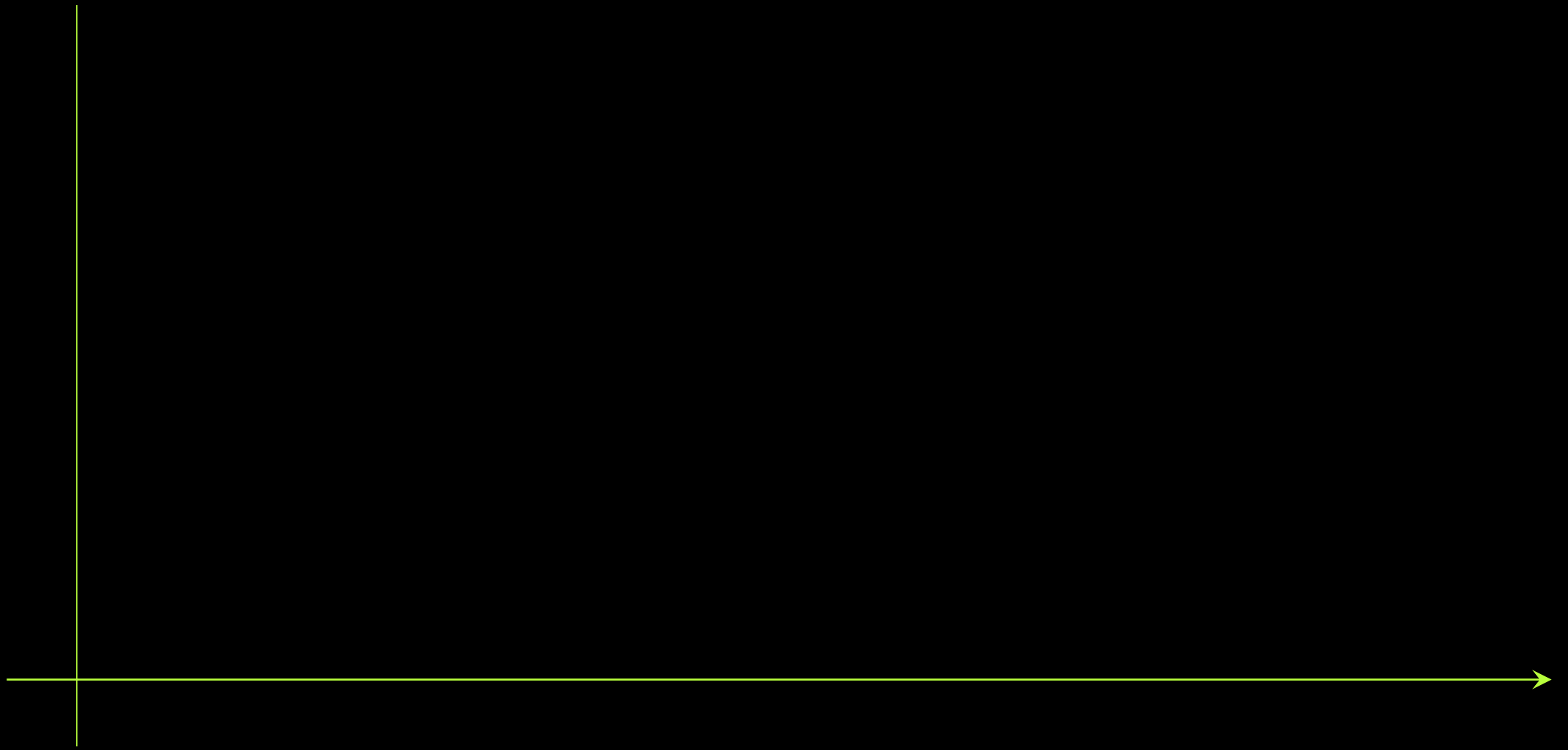


3. Transfer data back to the host

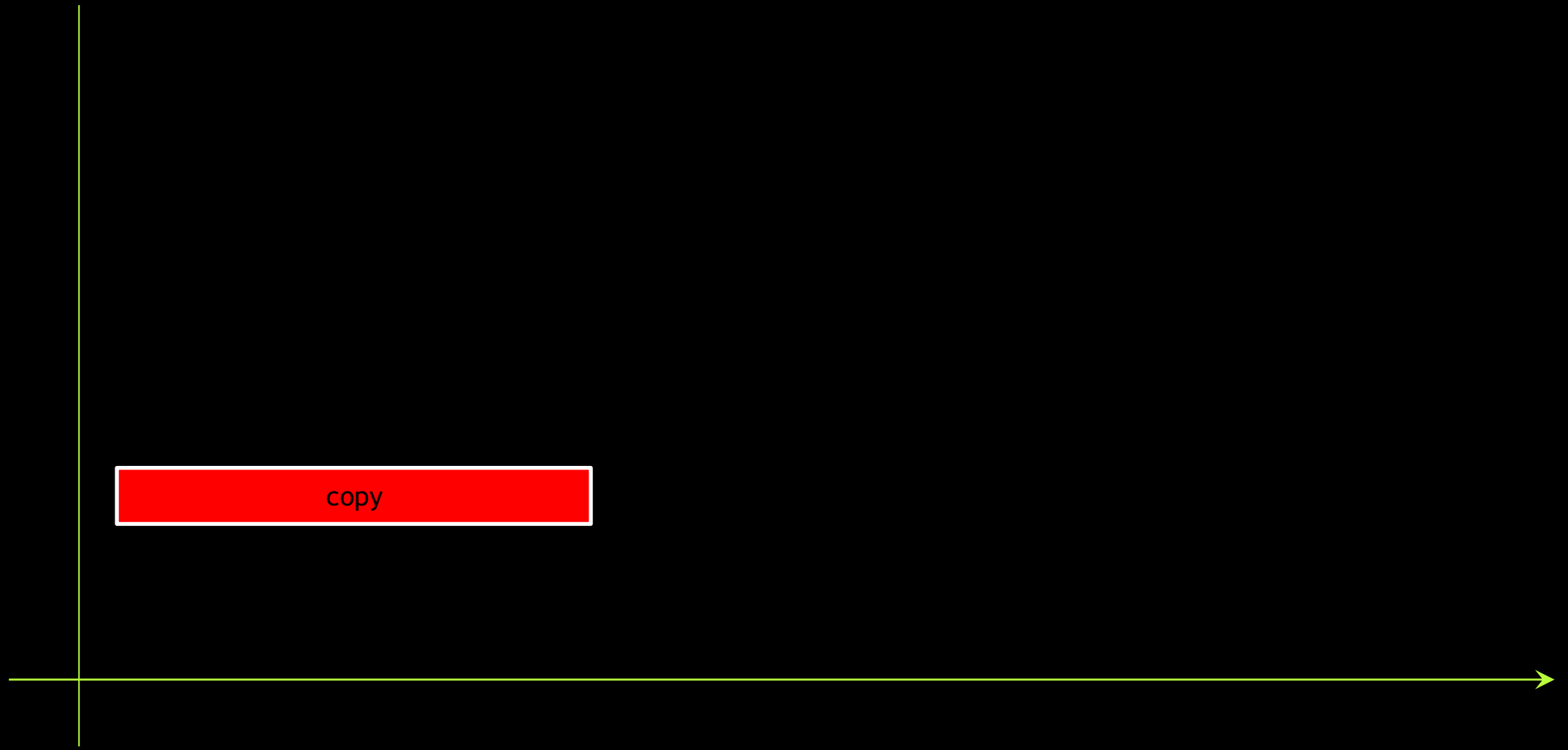



Total runtime is the sum
copy
compute
copy

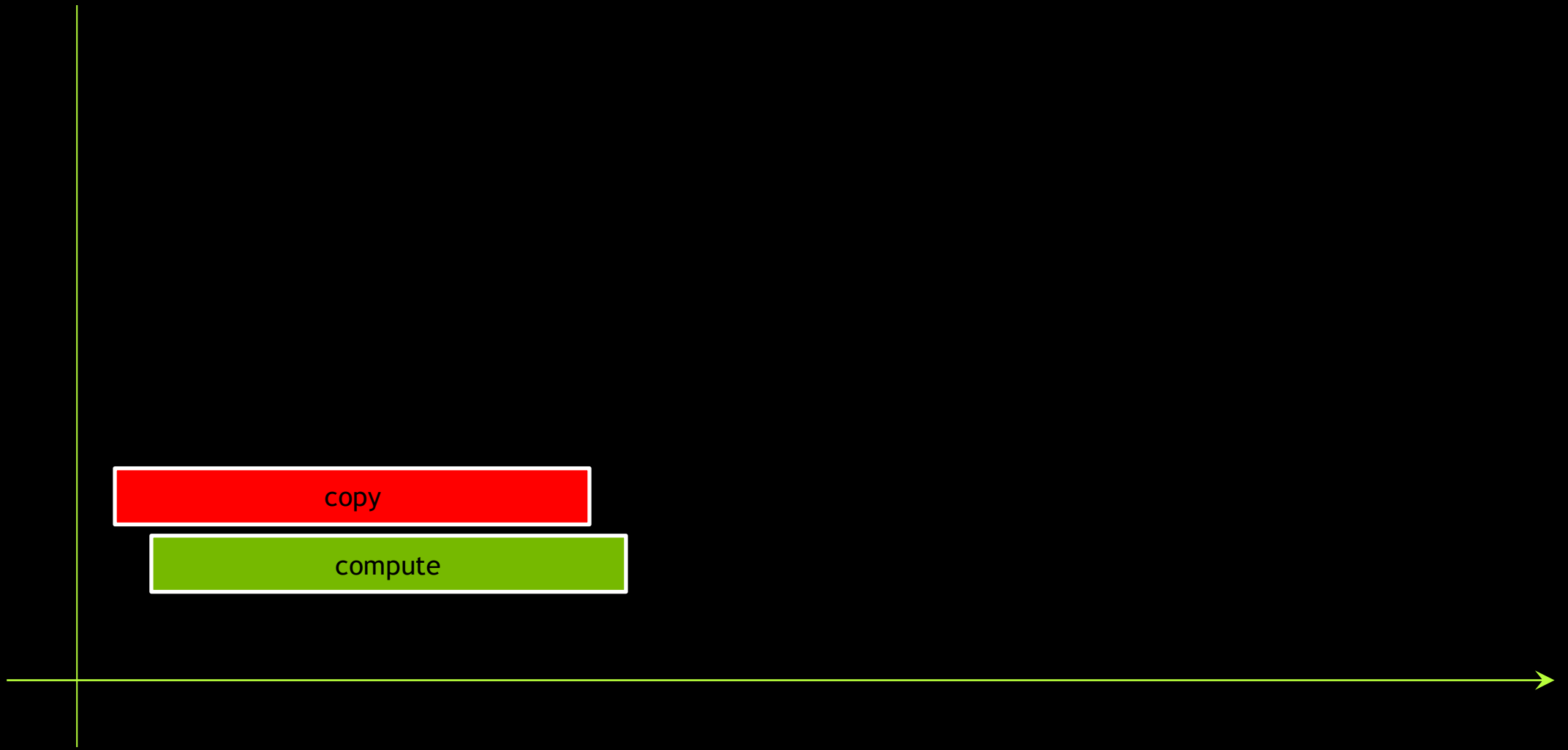
If we can overlap memory transfer and compute…



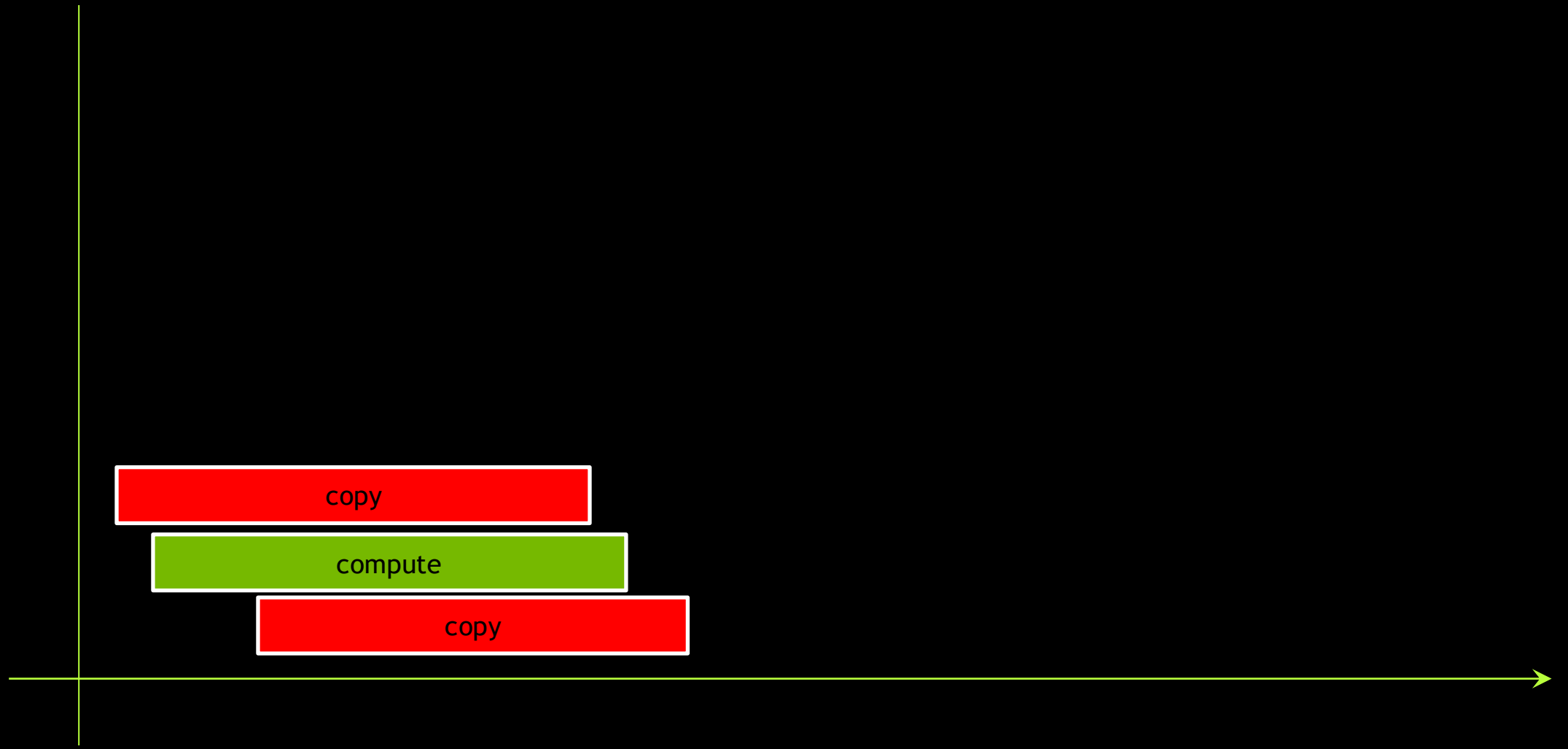
If we can overlap memory transfer and compute…



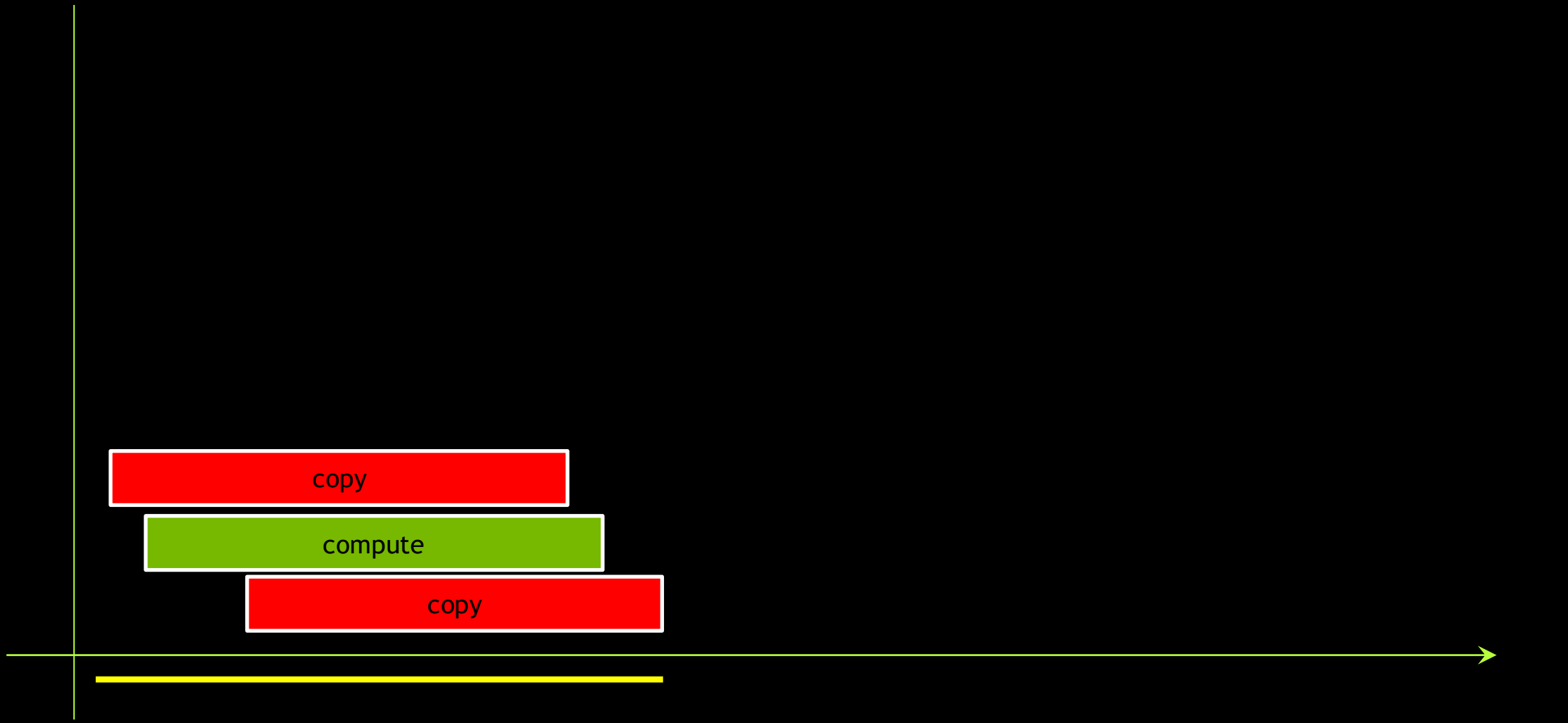
If we can overlap memory transfer and compute…



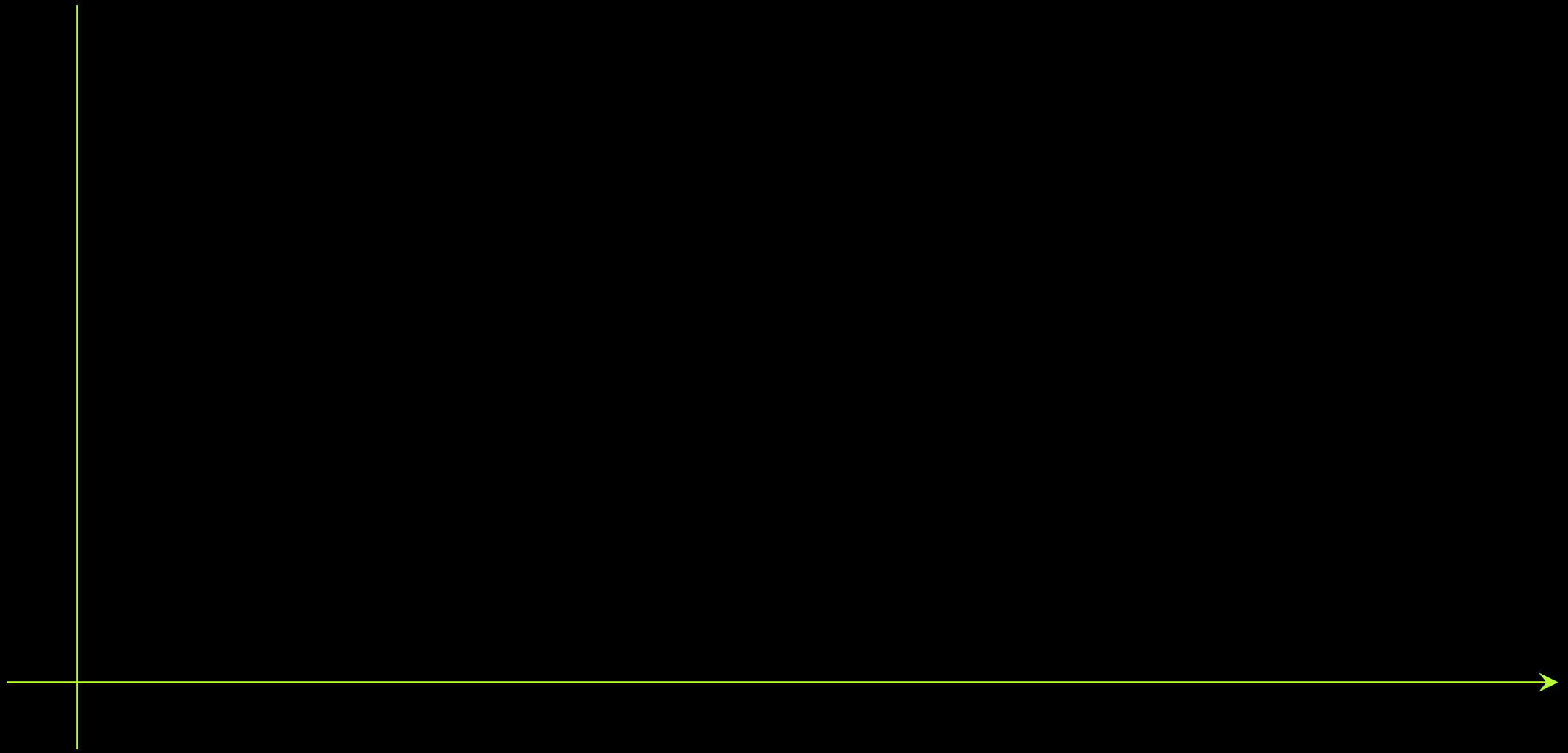
If we can overlap memory transfer and compute…



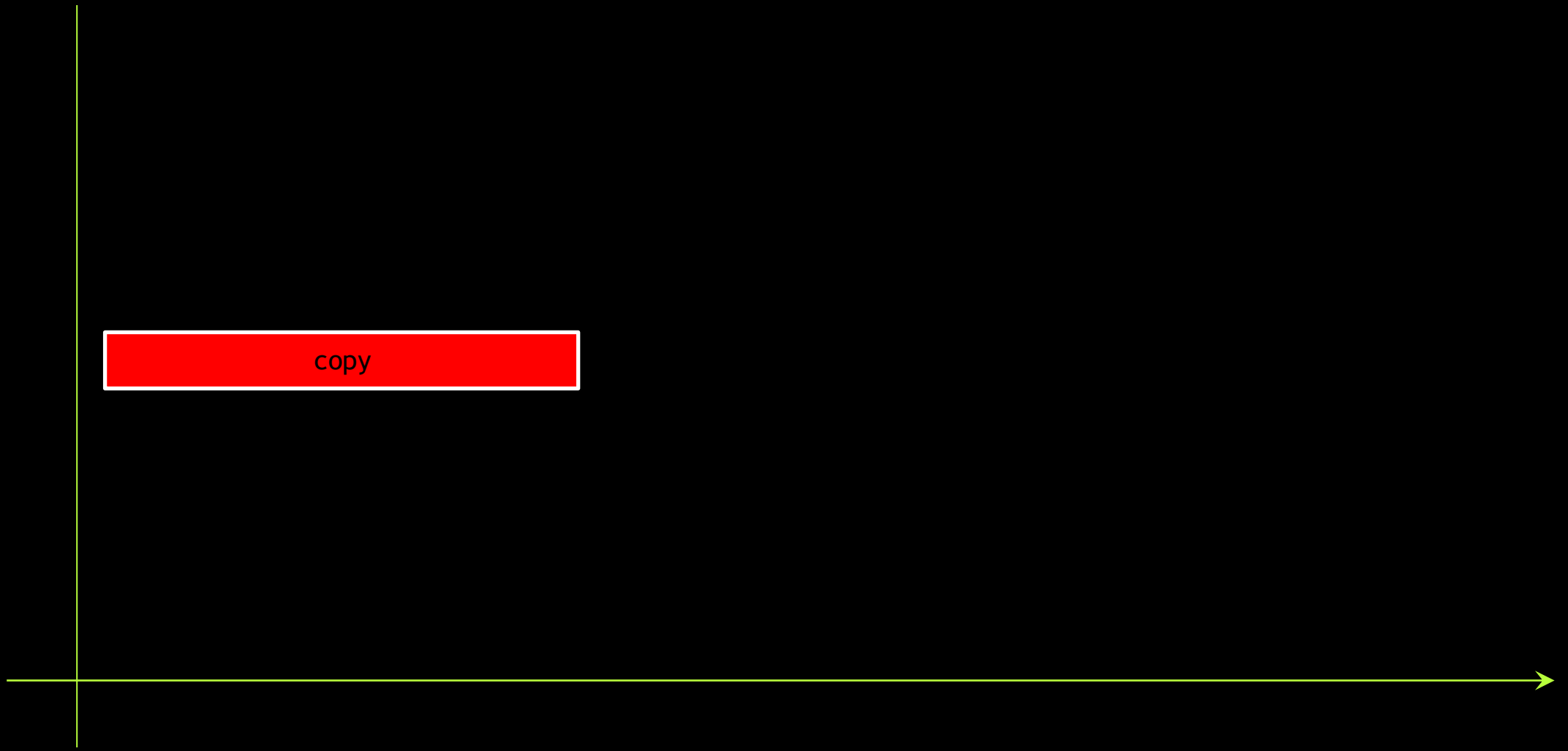
...total application time will be less



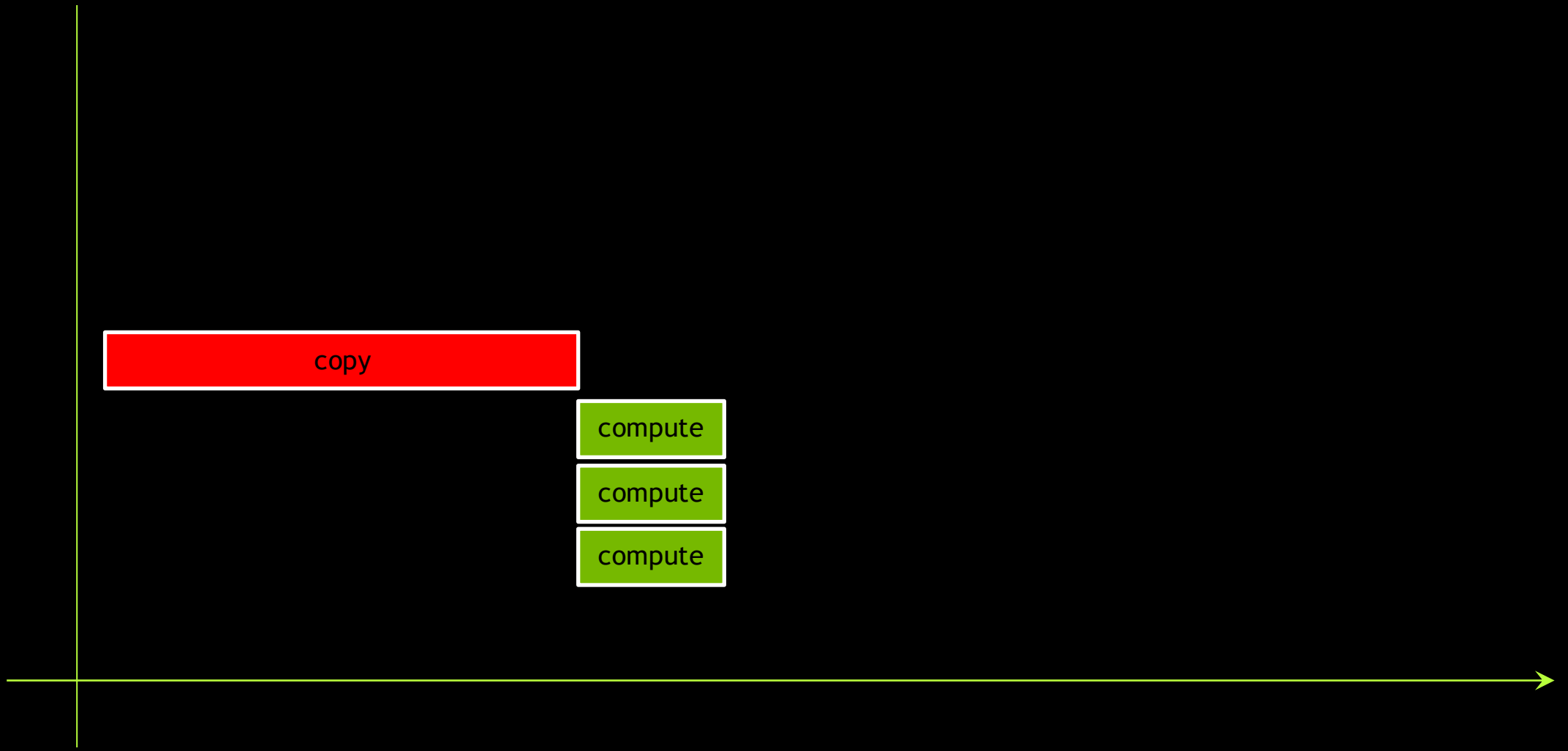
If we can overlap computation on multiple devices…



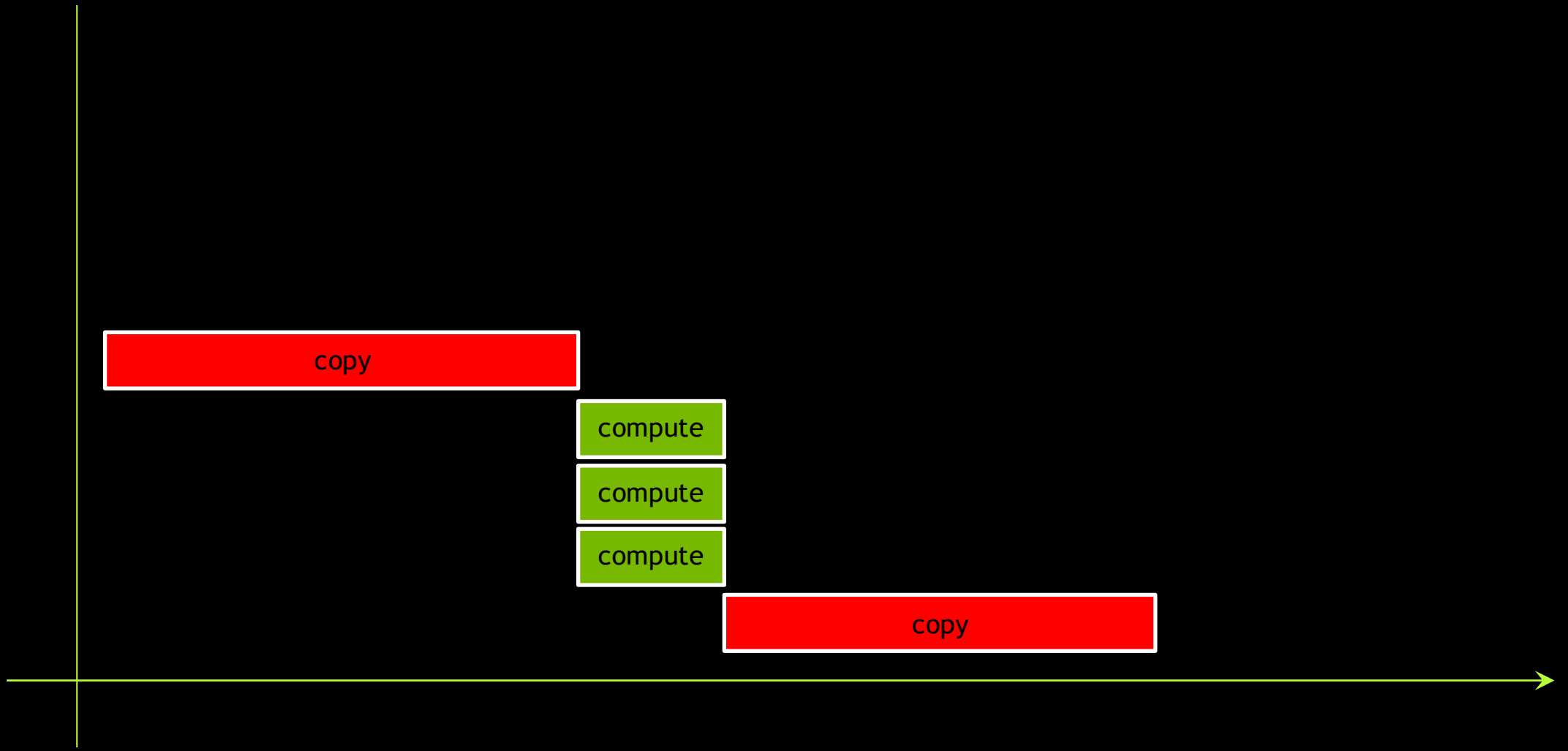
If we can overlap computation on multiple devices…



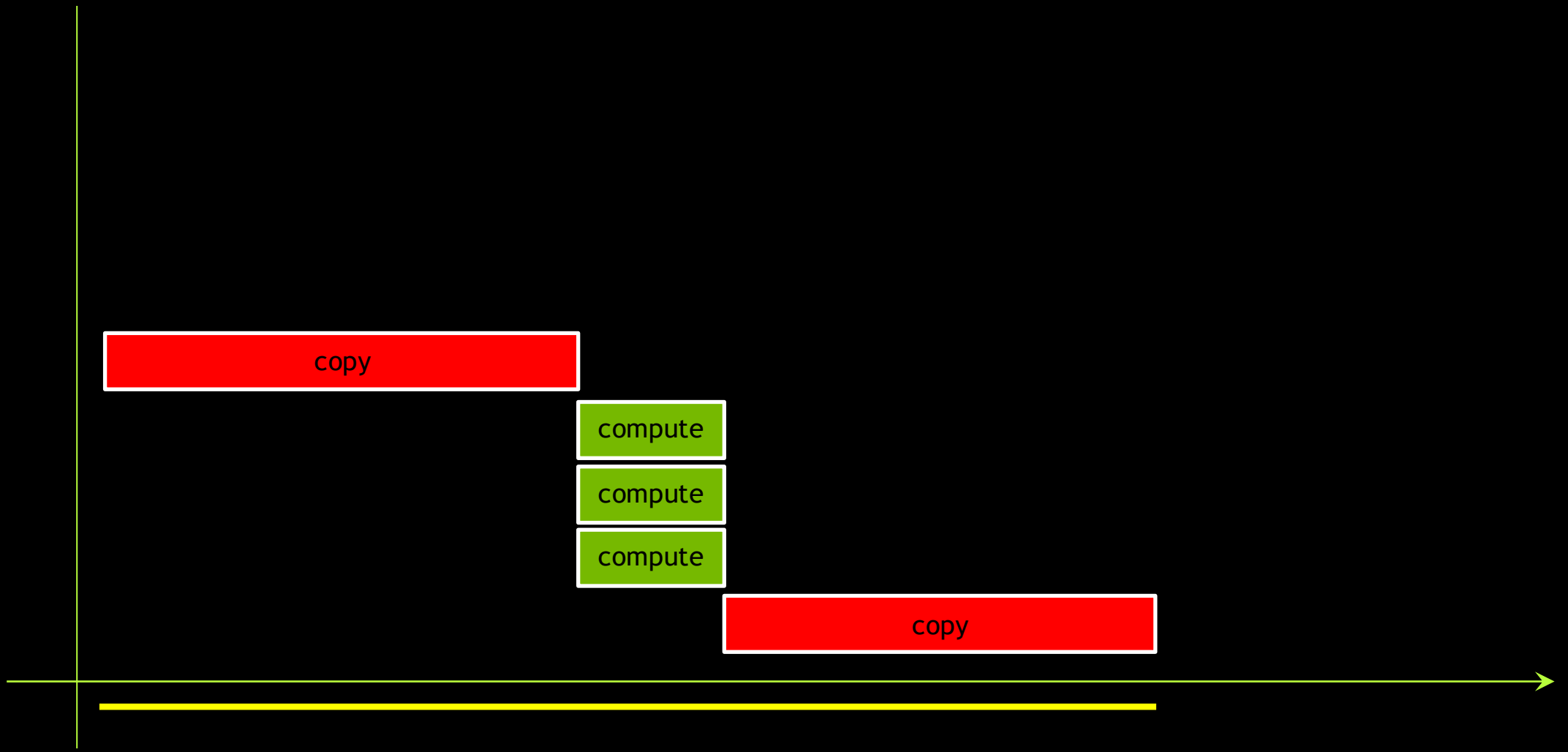
If we can overlap computation on multiple devices…



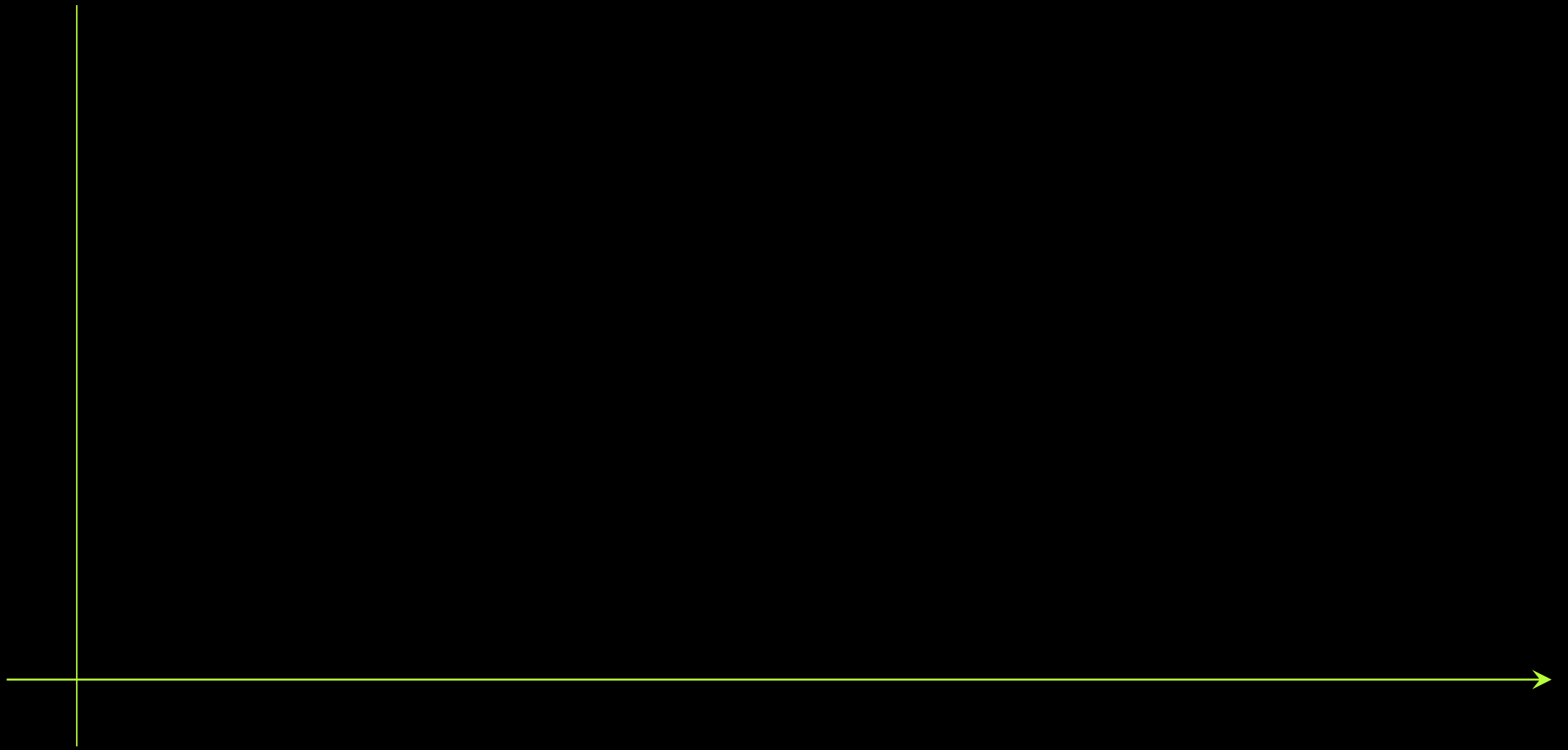
If we can overlap computation on multiple devices…



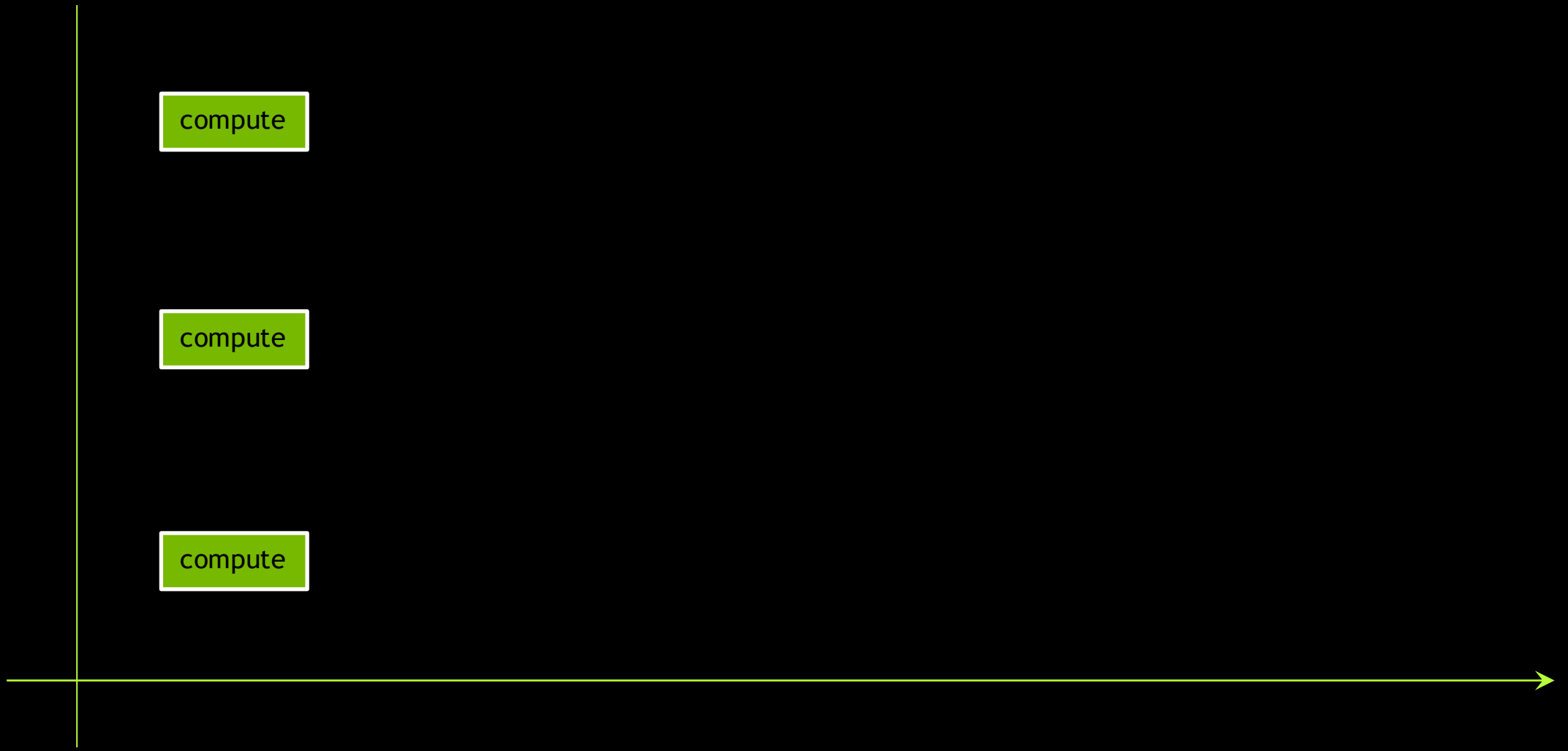
...total application time will also be less



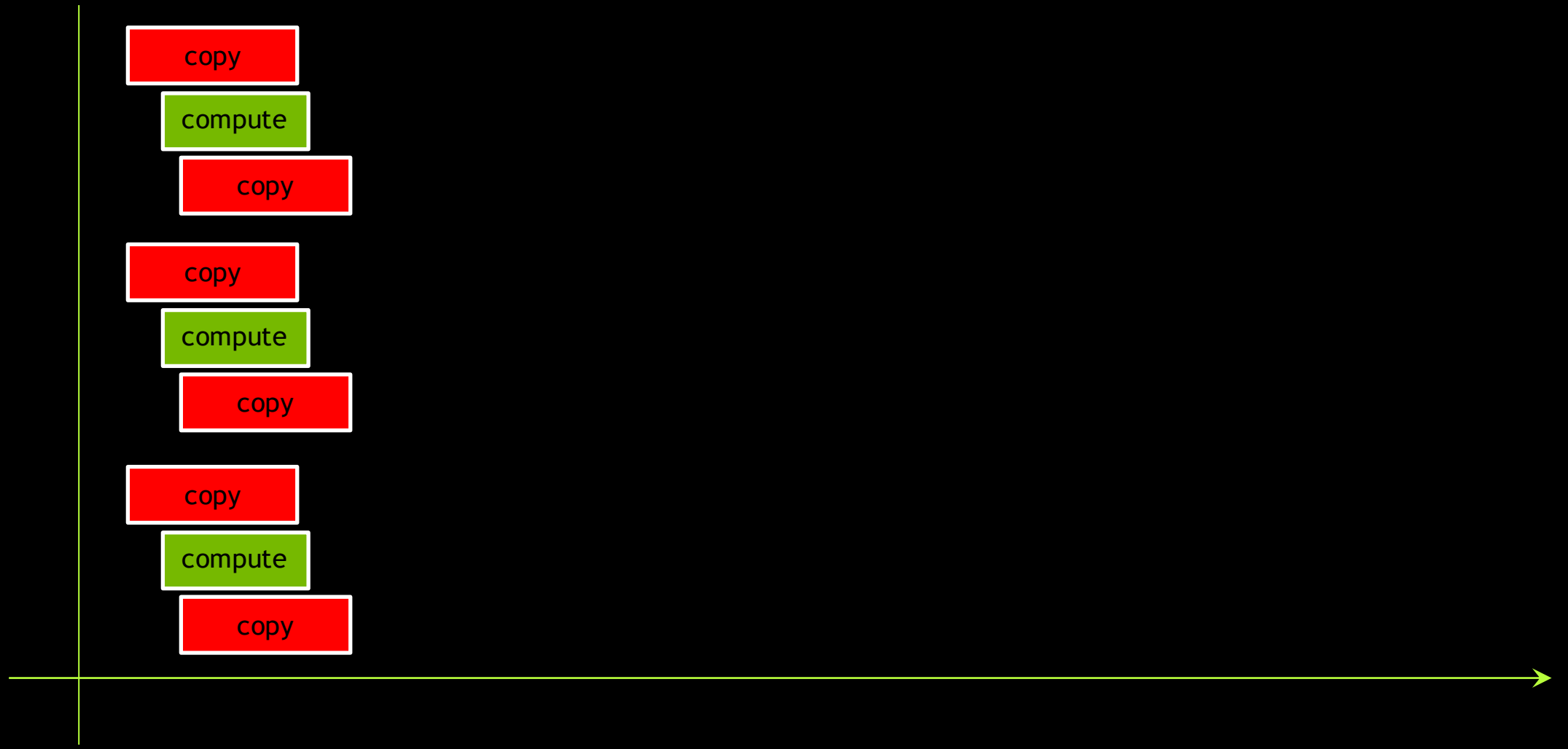
Combining the 2 strategies…



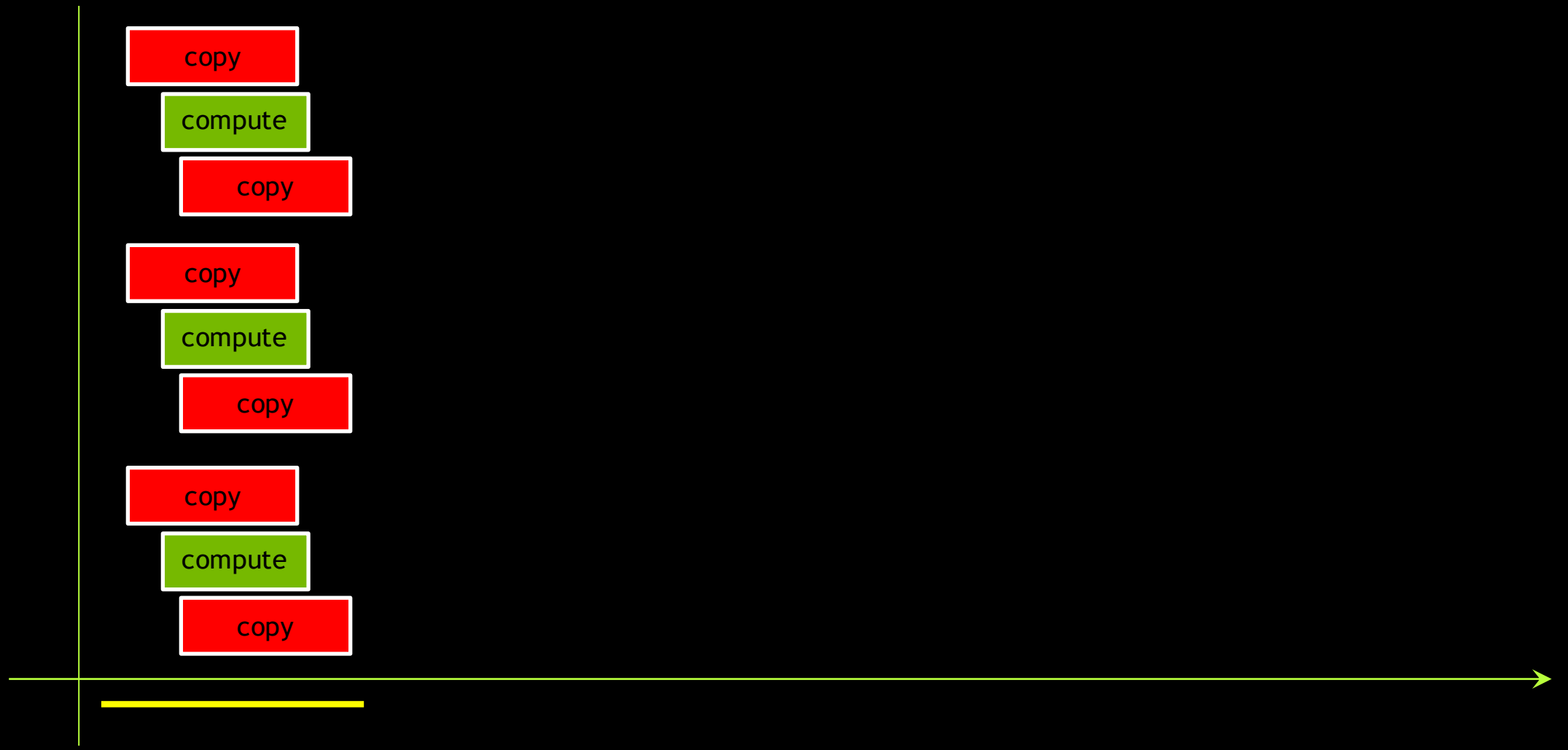
…overlapping compute on multiple devices



…and copy with each device’s compute



...total application time will be even less



MAIN OBJECTIVES

Increase performance for Single-Node CUDA C/C++ applications by exploiting, and then combining, 2 concurrency strategies offered to CUDA programmers:

1) Overlapping memory transfers to and from the GPU with computations on the GPU

2) Performing computations concurrently on more than one GPU

WORKSHOP STRUCTURE

WORKSHOP STRUCTURE

Introduction (this section)

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

Multiple GPUs

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

Multiple GPUs

Considerations for Multiple GPUs

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

Multiple GPUs

Considerations for Multiple GPUs

Exercise: Multiple GPUs

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

Multiple GPUs

Considerations for Multiple GPUs

Exercise: Multiple GPUs

Exercise: Multiple GPUs with Copy/Compute Overlap

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

Multiple GPUs

Considerations for Multiple GPUs

Exercise: Multiple GPUs

Exercise: Multiple GPUs with Copy/Compute Overlap

Course Survey

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

Multiple GPUs

Considerations for Multiple GPUs

Exercise: Multiple GPUs

Exercise: Multiple GPUs with Copy/Compute Overlap

Course Survey

Course Assessment

WORKSHOP STRUCTURE

Introduction (this section)

Using JupyterLab

Cipher Application Overview

Nsight Systems Setup

CUDA Streams

Kernel Launches in Non-Default Streams

Memory Copies in Non-Default Streams

Considerations for Copy/Compute Overlap

Exercise: Copy/Compute Overlap

Multiple GPUs

Considerations for Multiple GPUs

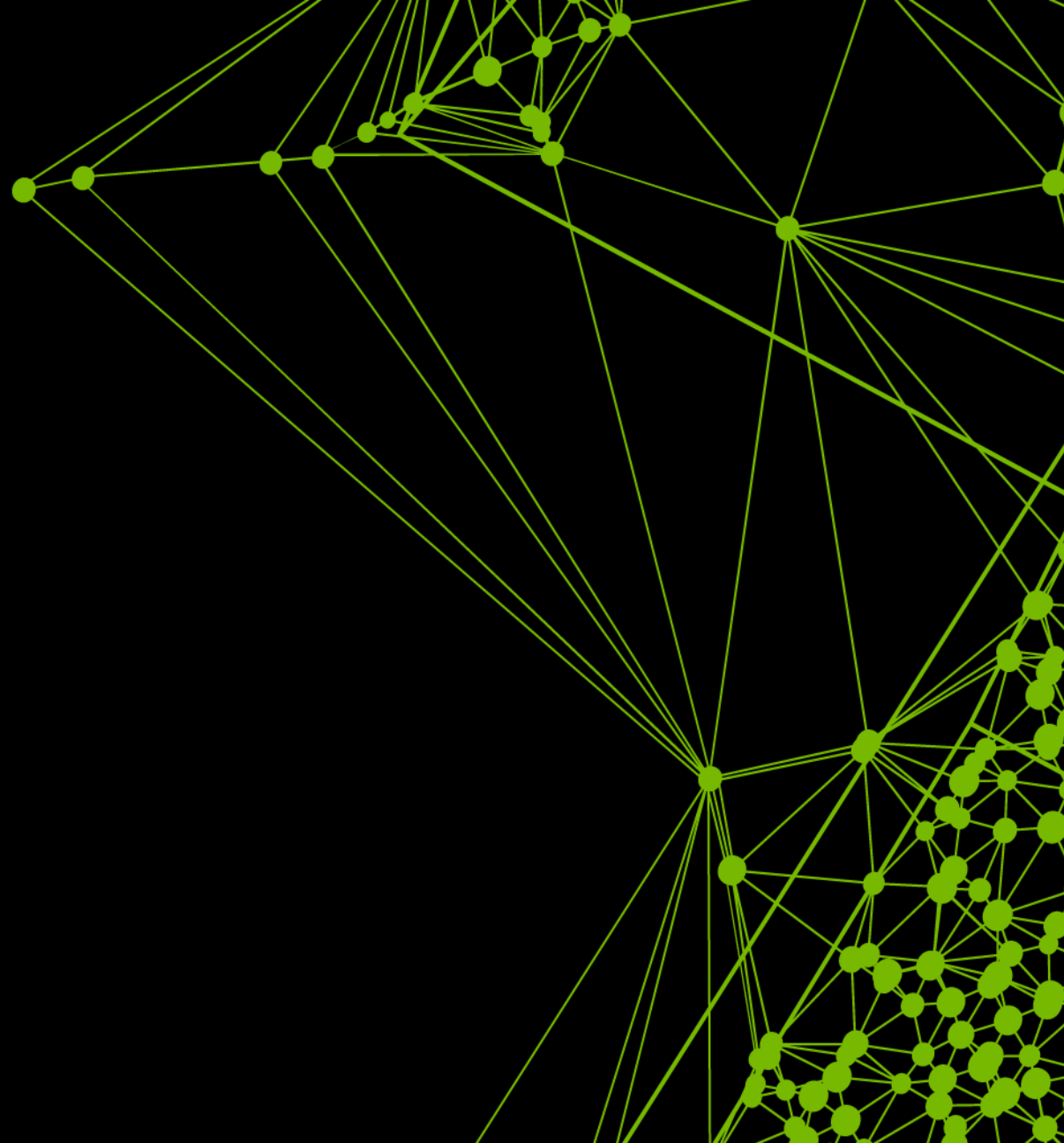
Exercise: Multiple GPUs

Exercise: Multiple GPUs with Copy/Compute Overlap

Course Survey

Course Assessment

Next Steps



NVIDIA | DEEP LEARNING INSTITUTE

www.nvidia.com/dli