CUDA高性能科学计算

2023-2024-1 秋季

公选-06110

Lecture 8

理学院　赖欣

授课平台/讨论QQ群

CUDA高性能科学计算(GX)
课程编号:107016


加速计算基础——CUDA C/C++
8 学时 | 中文 | 90 美元 | C/C++，CUDA ®
有培训证书

复习

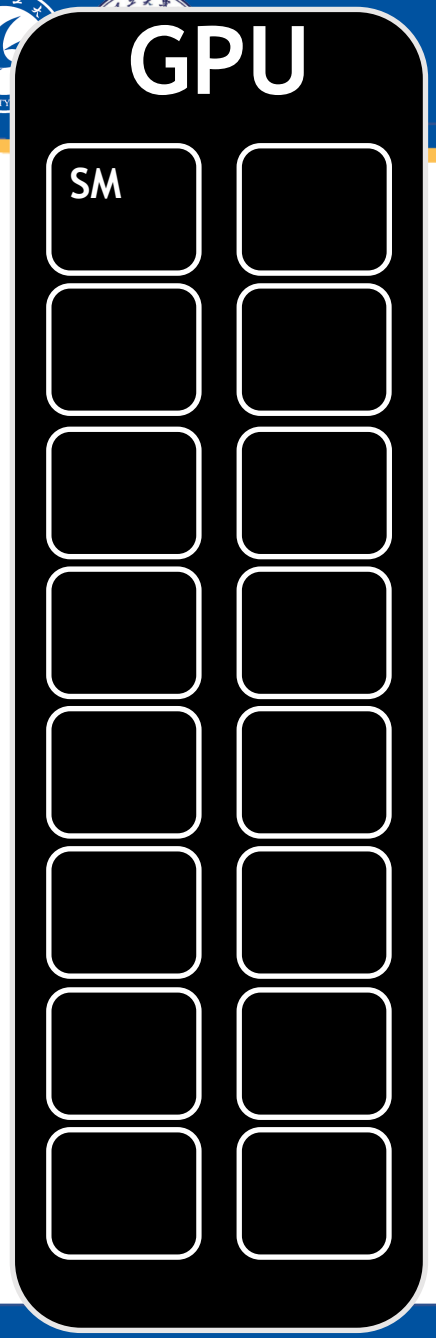
- Nsys命令行分析器
- 使用nsys性能分析器帮助应用程序迭代地进行优化
- 利用基本的 CUDA 内存管理技术来优化加速应用程序
- 流多处理器（Streaming Multiprocessors）及查询GPU的设备配置
- 统一内存行为
  - 统一内存(UM)的迁移


利用基本的 CUDA 内存管理技术来优化加速应用程序

流处理器

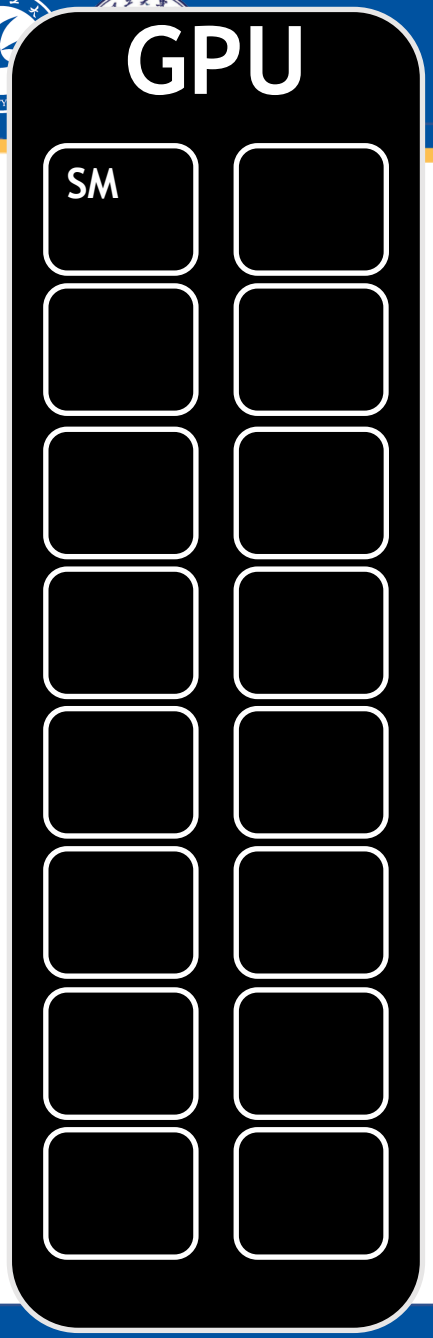
GPU

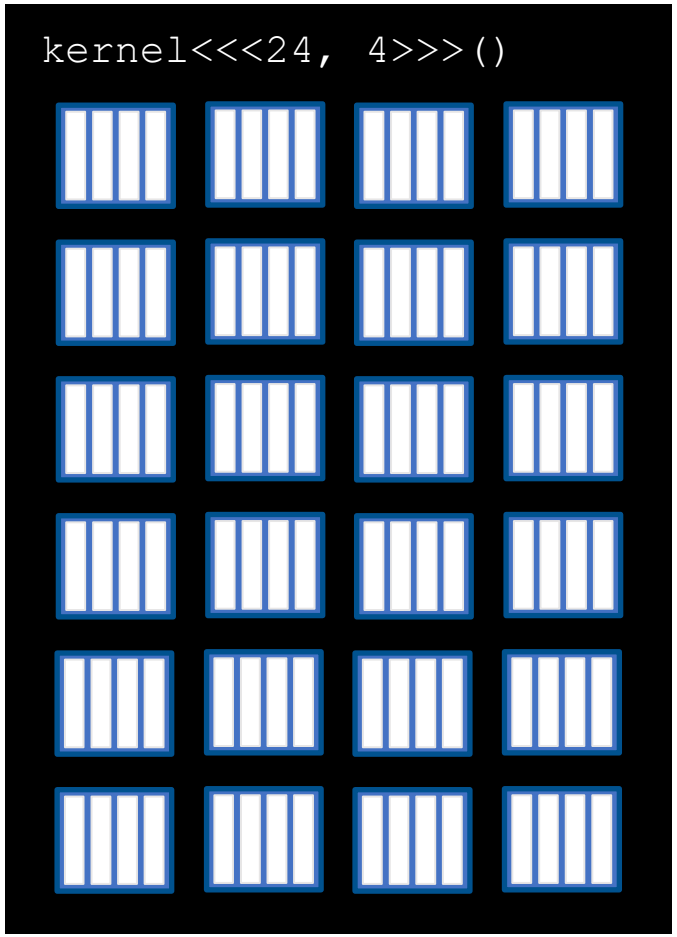
NVIDIA GPU 包含称为**流多处理器**或 **SM** 的功能单元

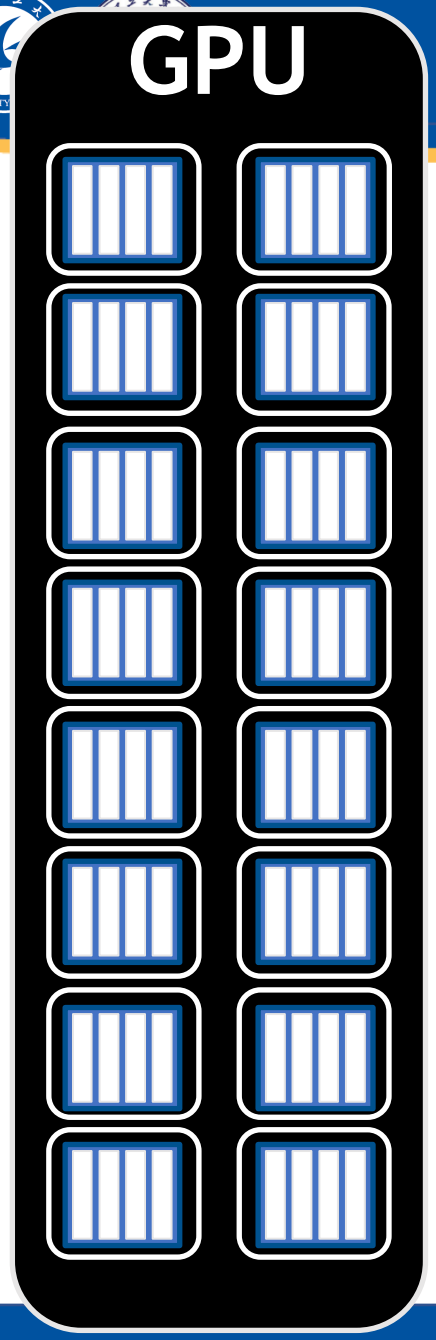


NVIDIA GPU 包含称为**流多处理器**或 **SM** 的功能单元

线程块均可安排在 SM 上运行

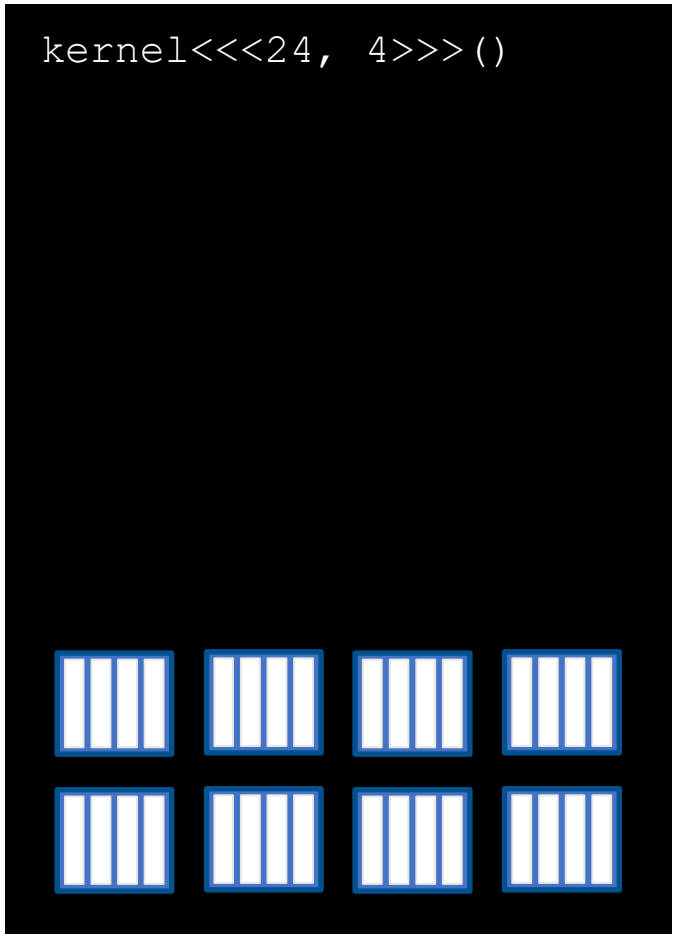


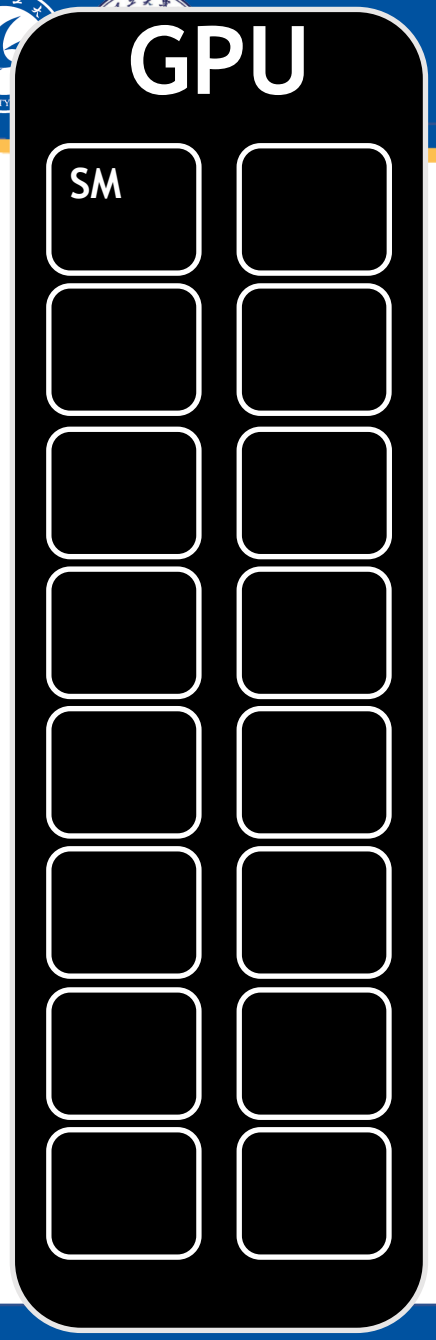


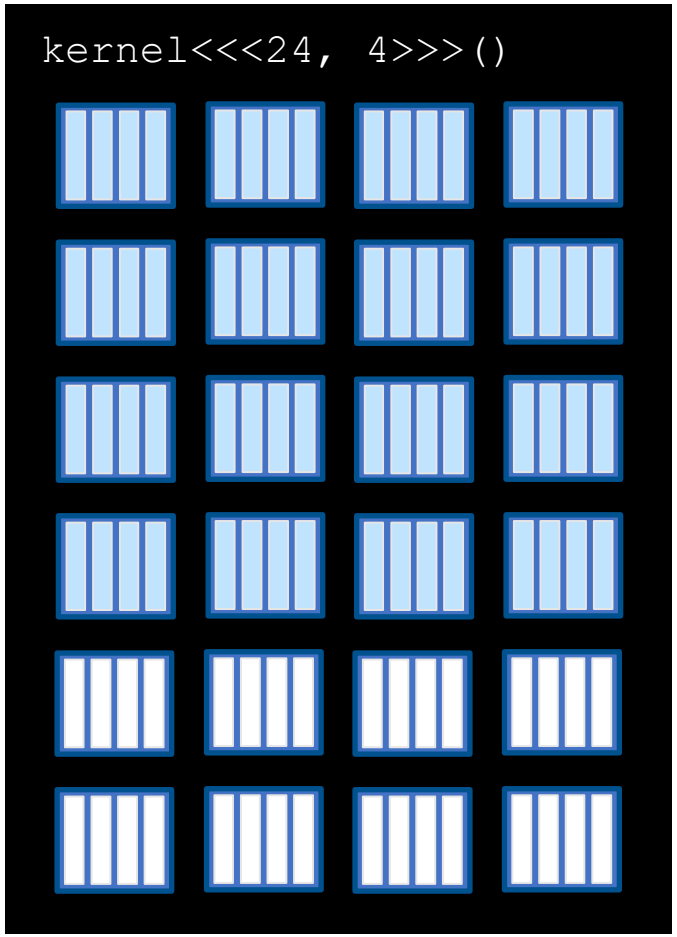


根据 GPU 上的 SM 数量以及线程块要求，可在 SM 上安排运行多个线程块

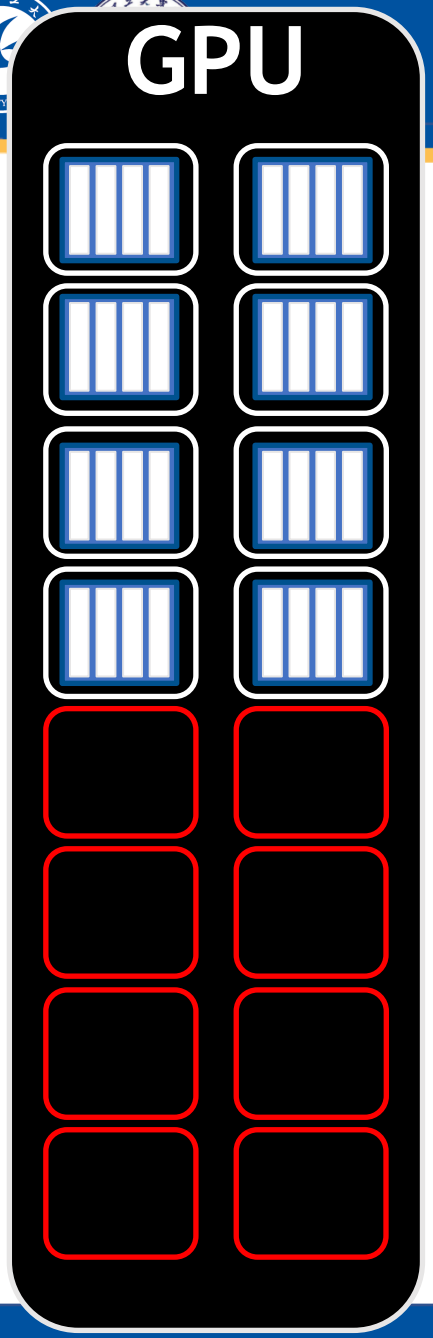
```
kernel<<<24, 4>>>()
```

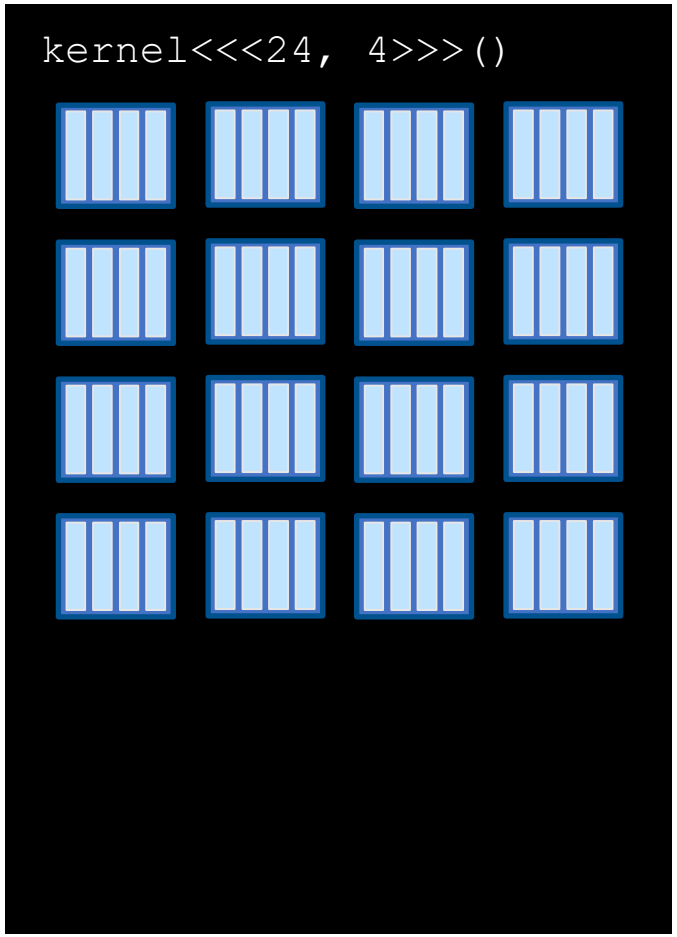






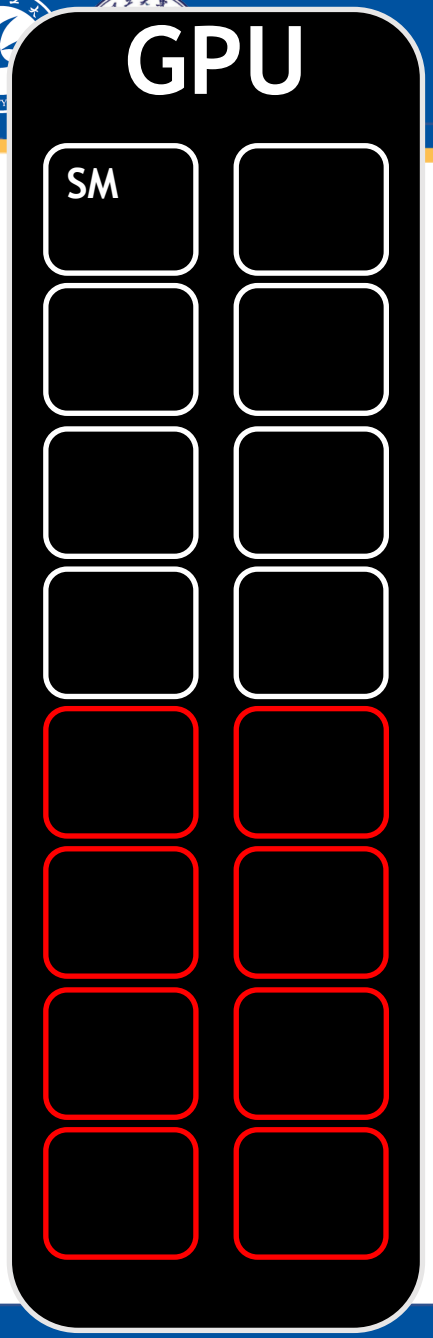
根据 GPU 上的 SM 数量以及**线程块**要求，可在 SM 上安排运行多个**线程块**

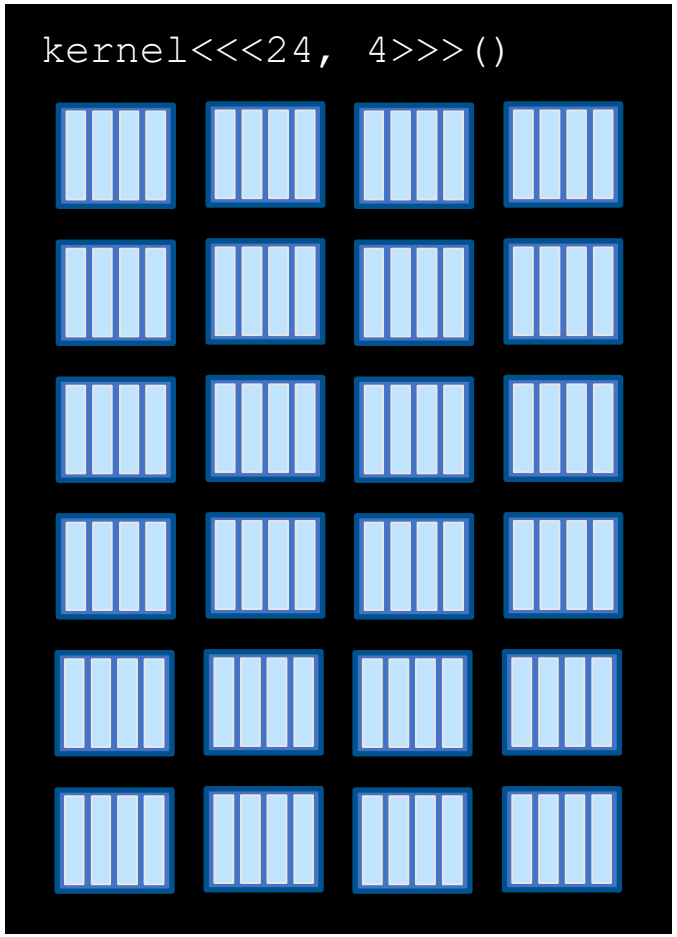




如果网格维度能被 GPU 上的 SM 数量整除，则可充分提高 SM 的利用率

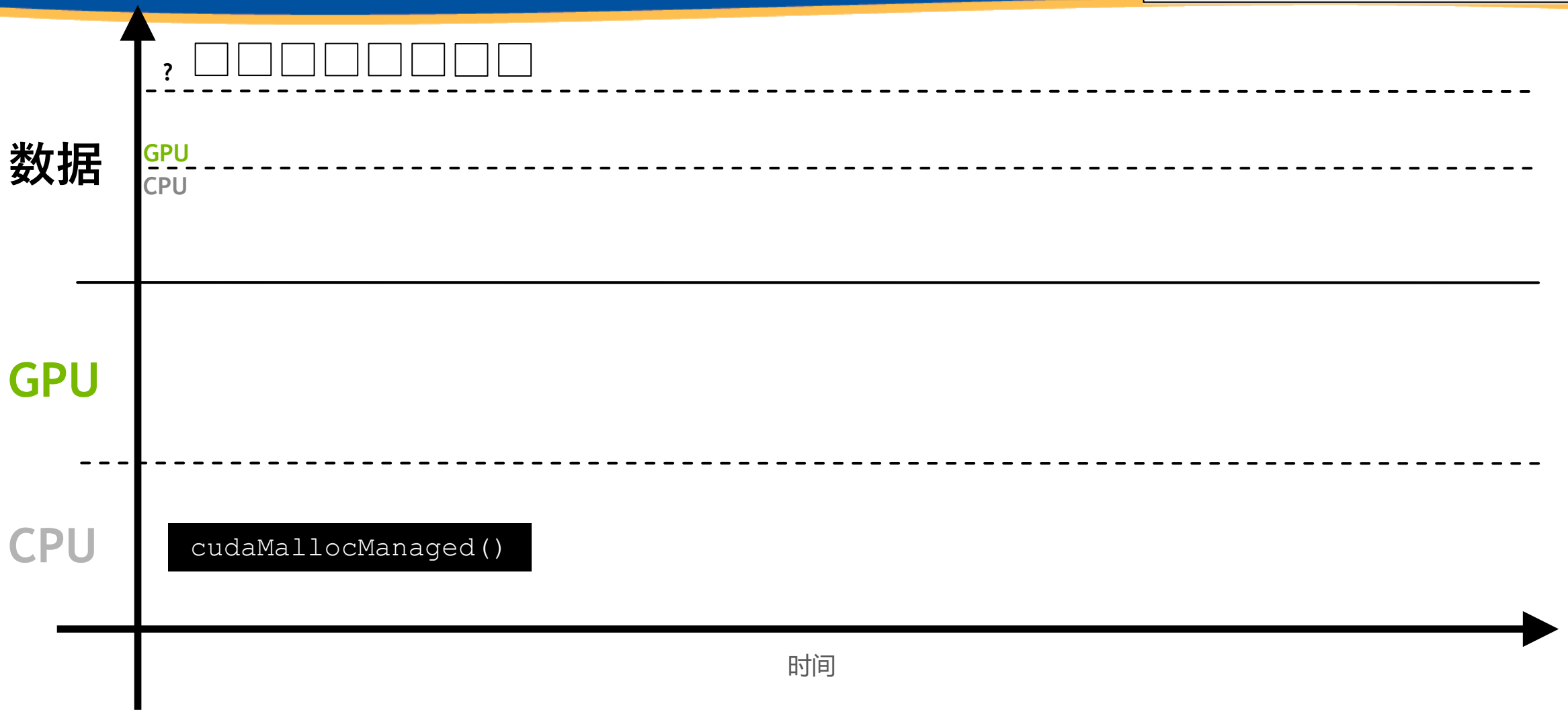
以下是闲置的 SM



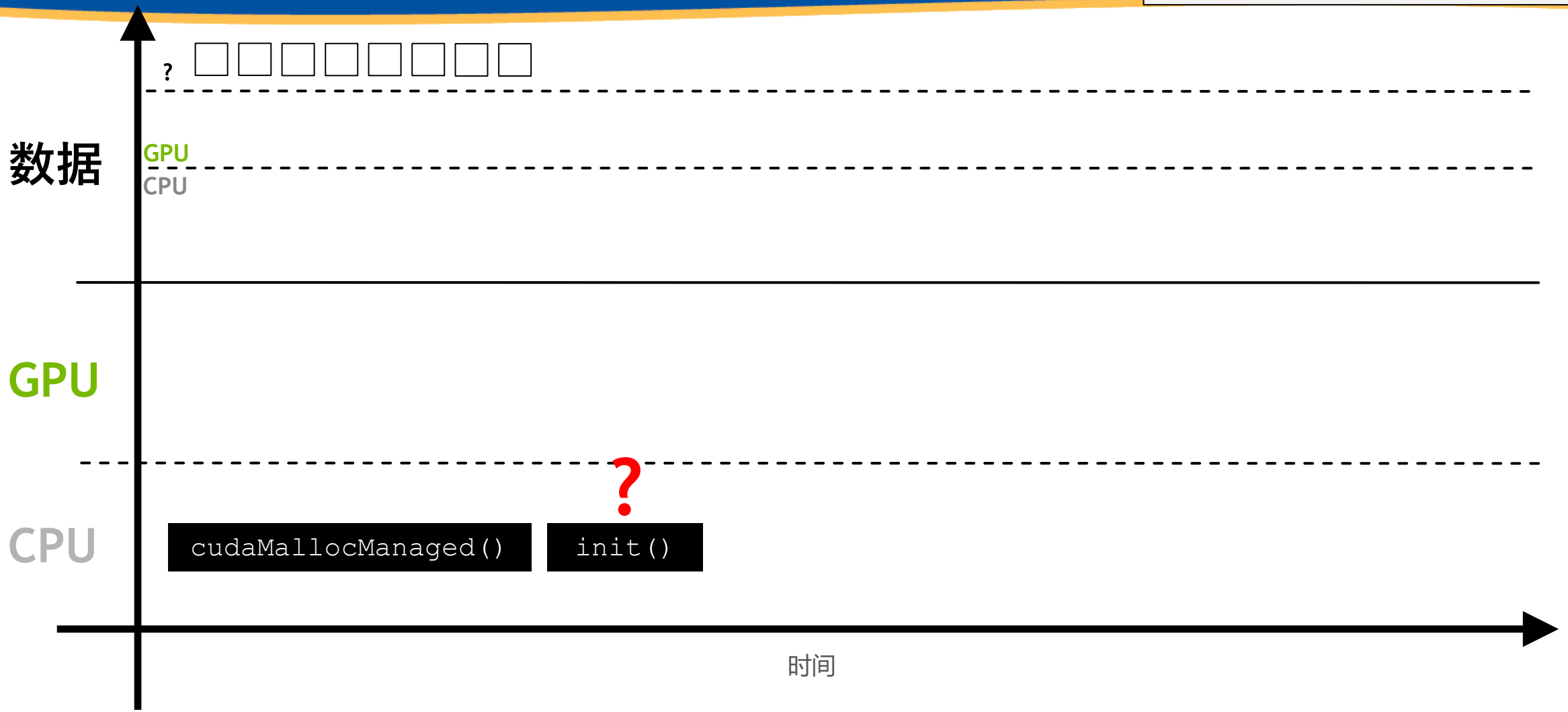


统一内存行为

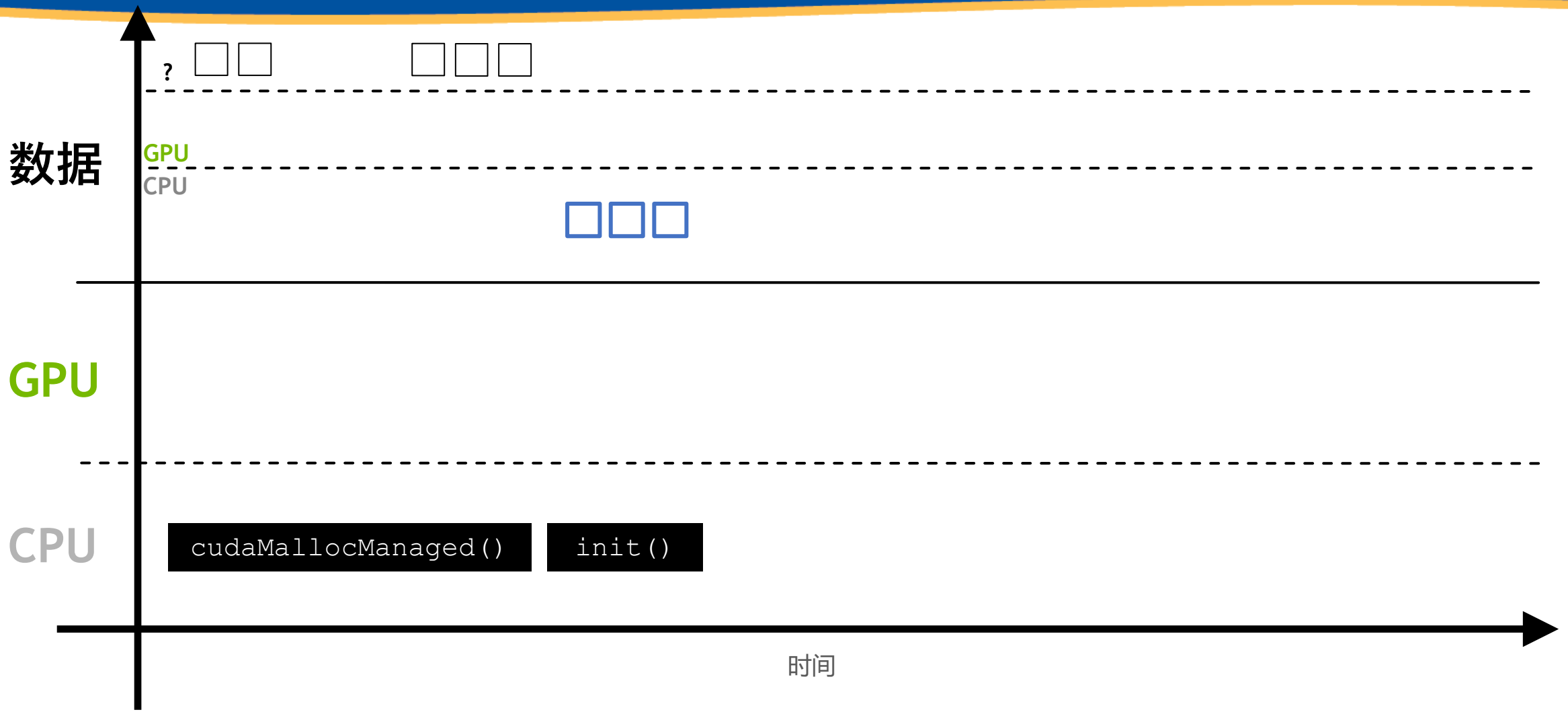
分配 UM 时，它最初可能并未驻留在 CPU 或 GPU 上



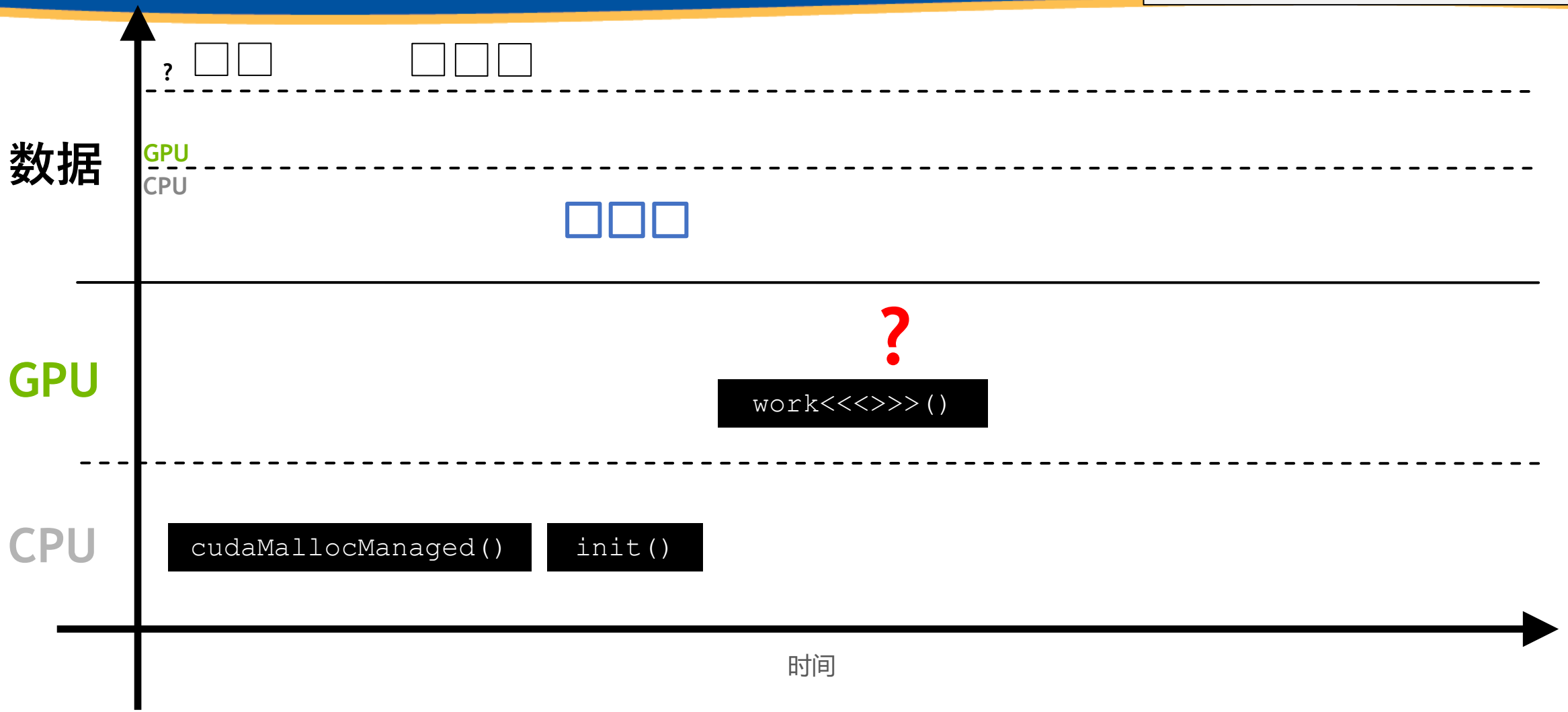
当某些工作首次请求内存时，将会发生
分页错误



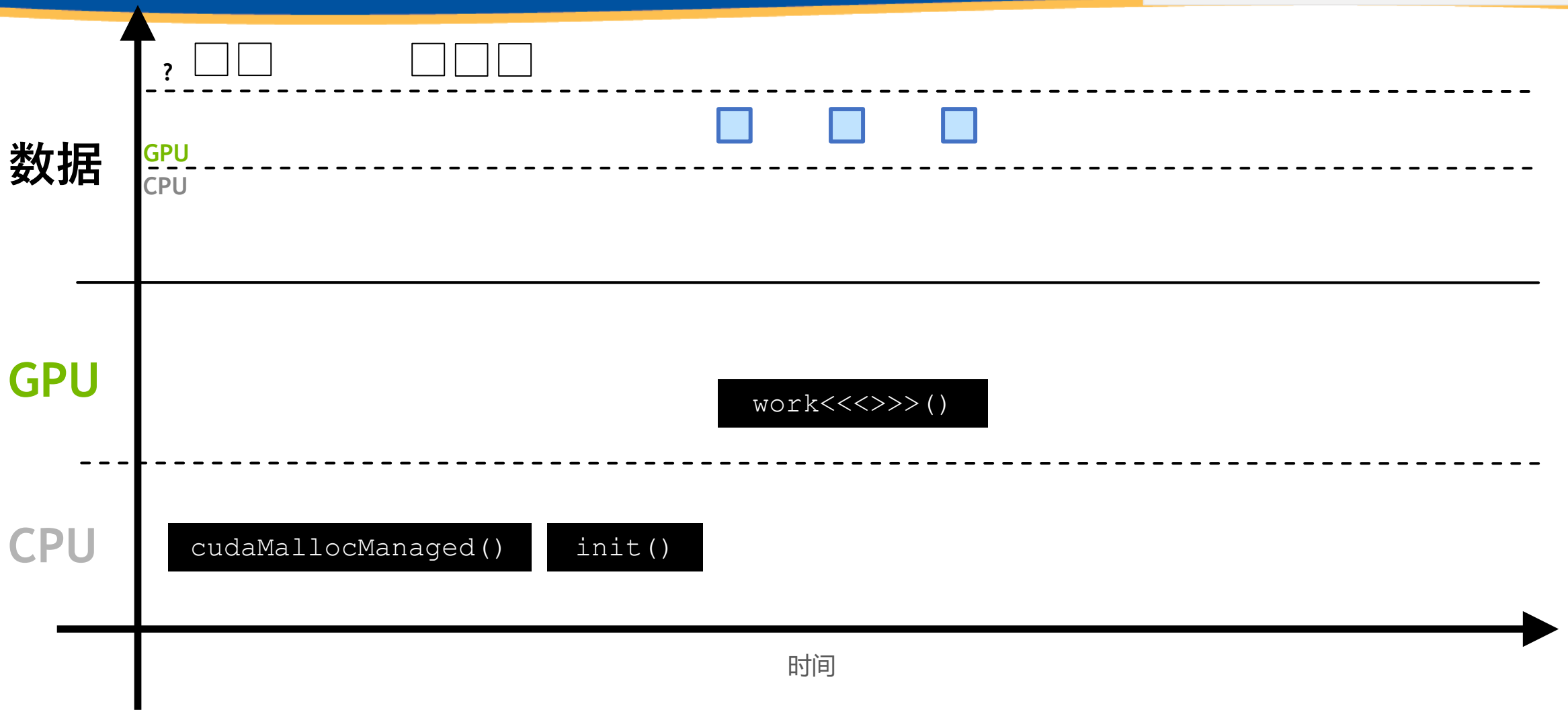
分页错误将触发所请求的内存发生迁移



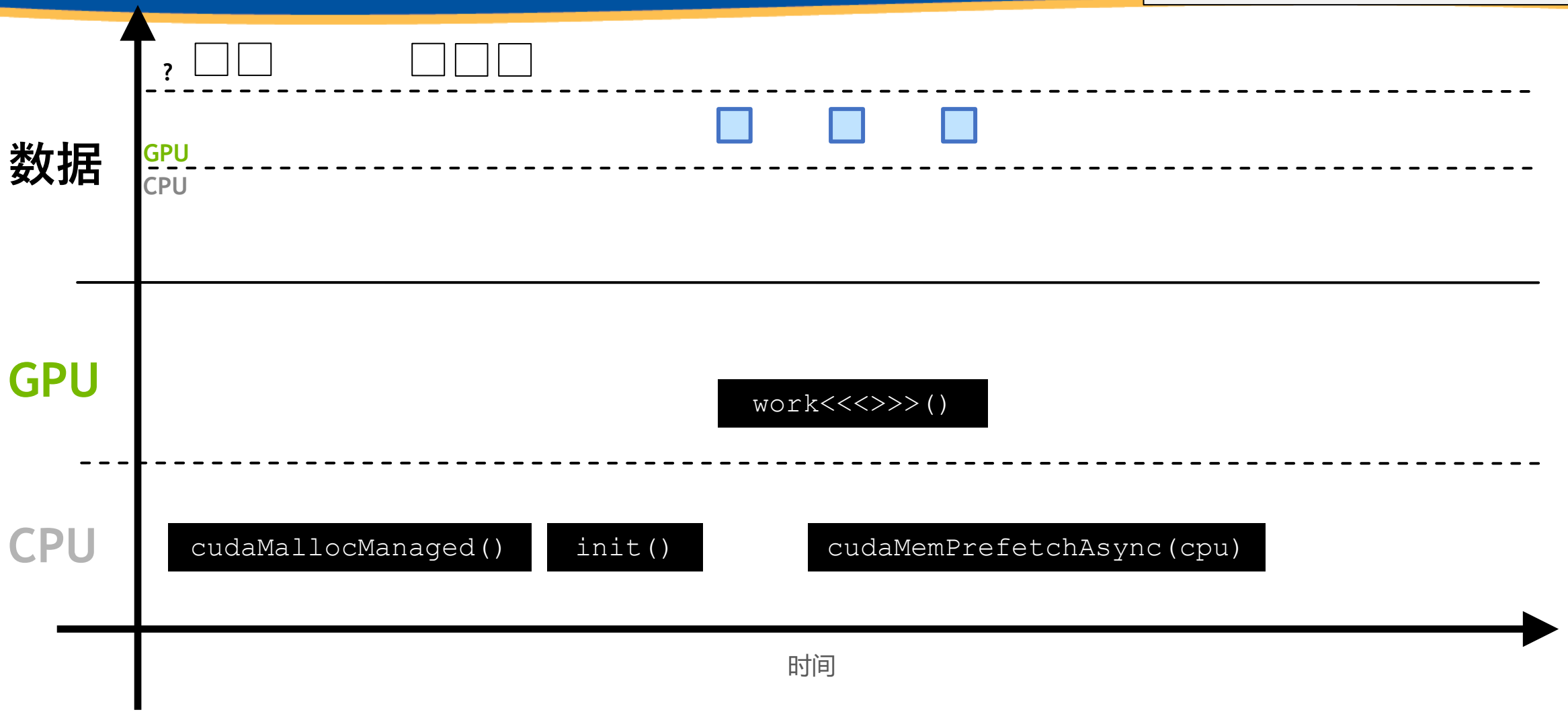
只要在系统中并未驻留内存的位置请求内存，此过程便会重复



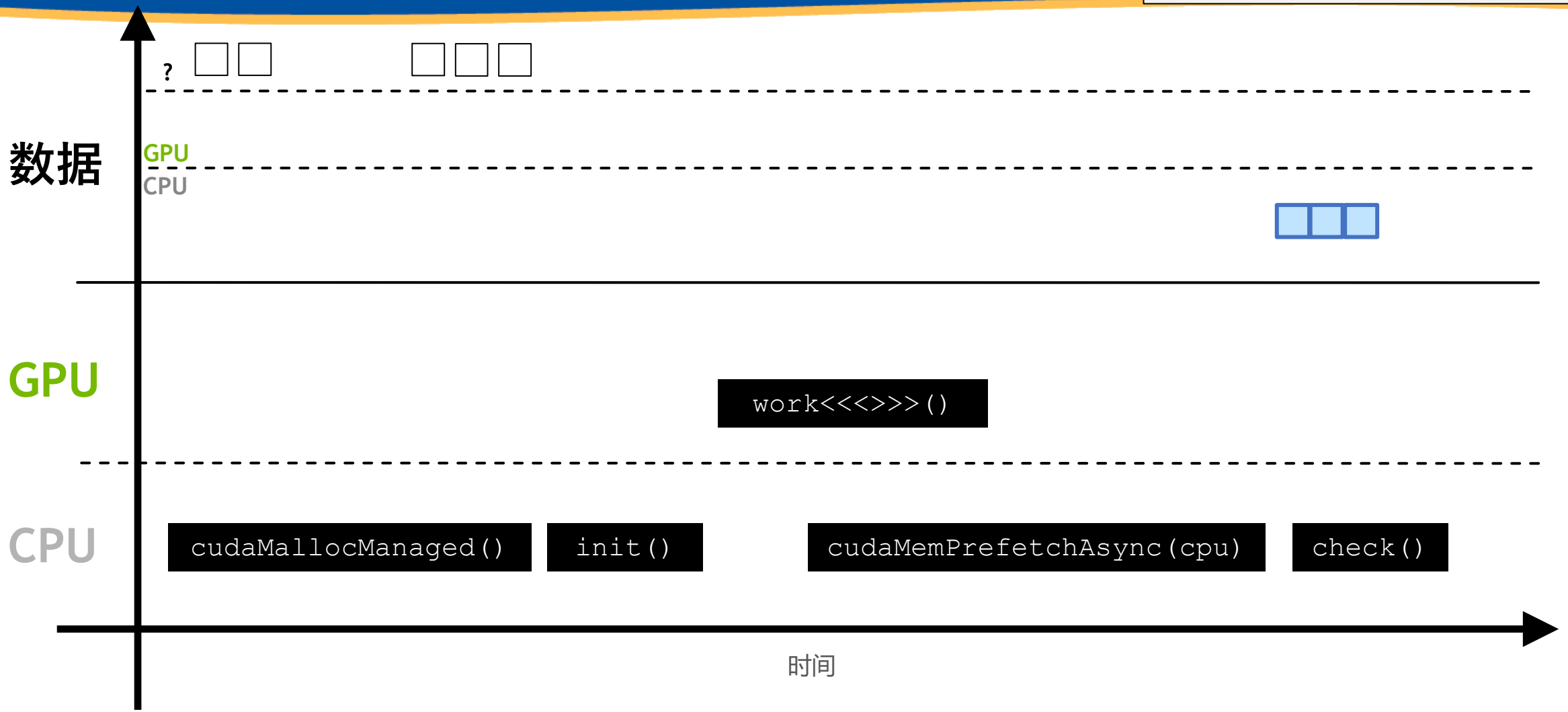
只要在系统中并未驻留内存的位置请求内存，此过程便会重复



如果已知**将在**未驻留内存的位置**访问**内存，**则**可使用异步**预**取



异步预取能以更大批量移动内存，
并会防止发生分页错误




异步内存预取


NVIDIA