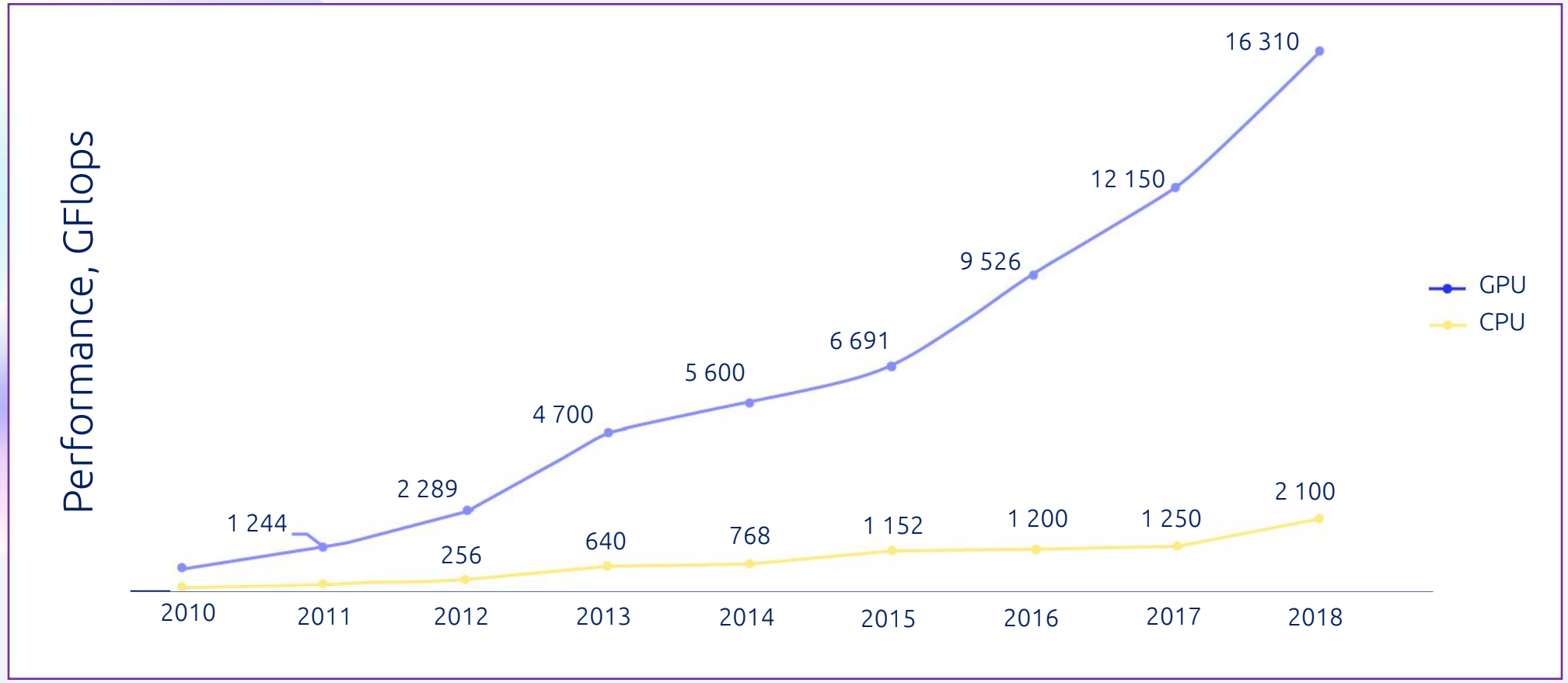
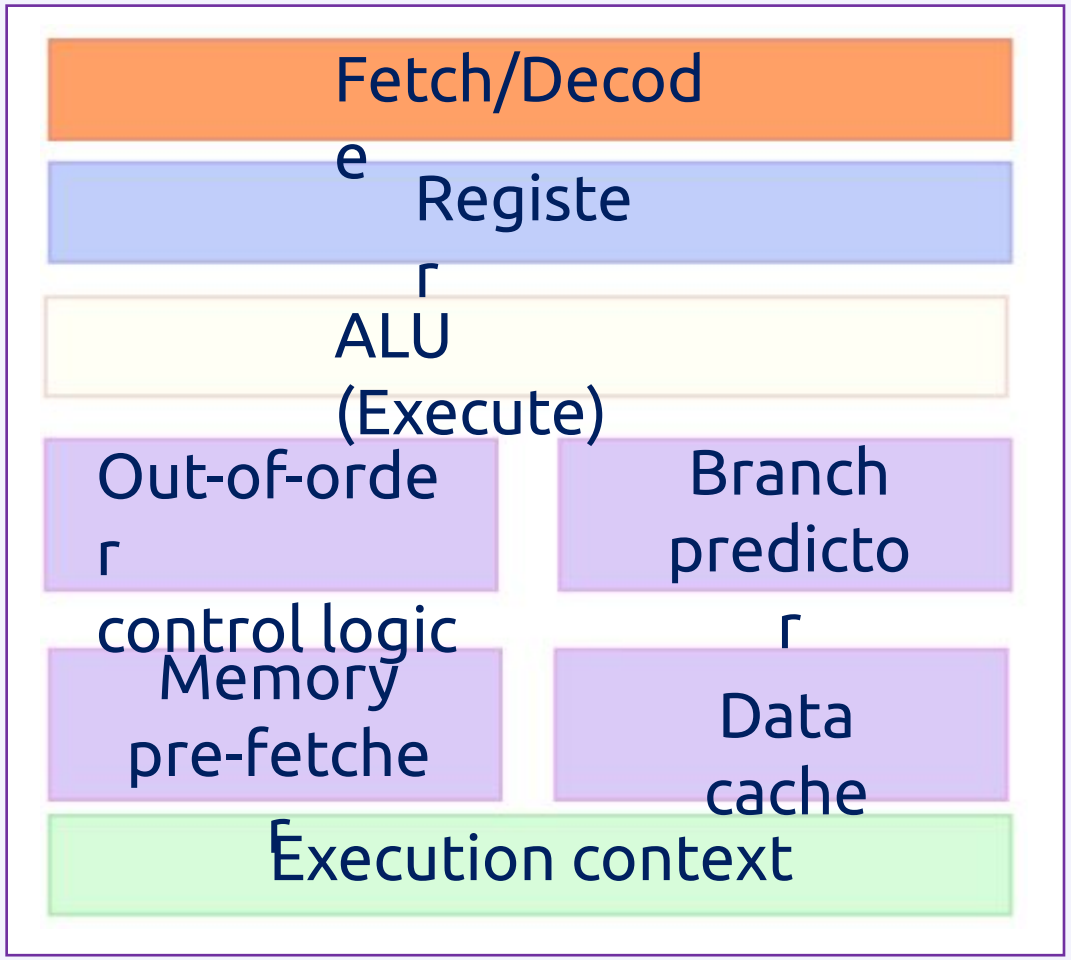
ОСОБЕННОСТИ ПРОГРАММИРОВАНИЯ НА GPU

ЗАЧЕМ ЧТО-ТО СЧИТАТЬ НА GPU?

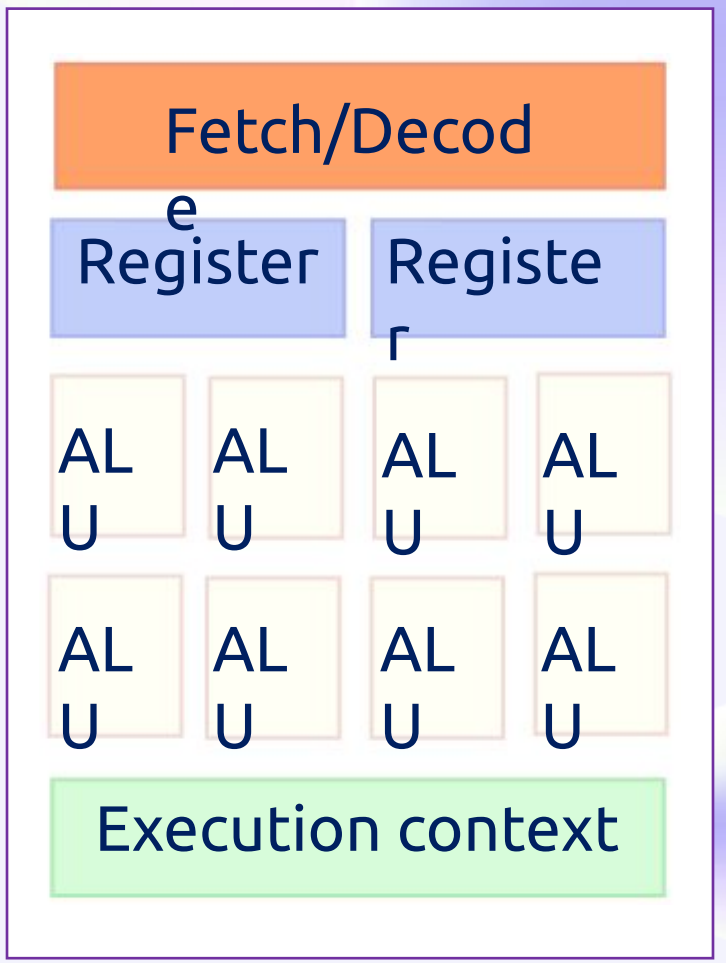


На графике изображены рост флопсов с течением времени для процессоров и для видеокарт.

АРХИТЕКТУРА GPU И ЕЕ СРАВНЕНИЕ С CPU



CPU Core



GPU Core

ОГРАНИЧЕНИЯ И ВОЗМОЖНОСТИ ПРИ РАБОТЕ С GPU

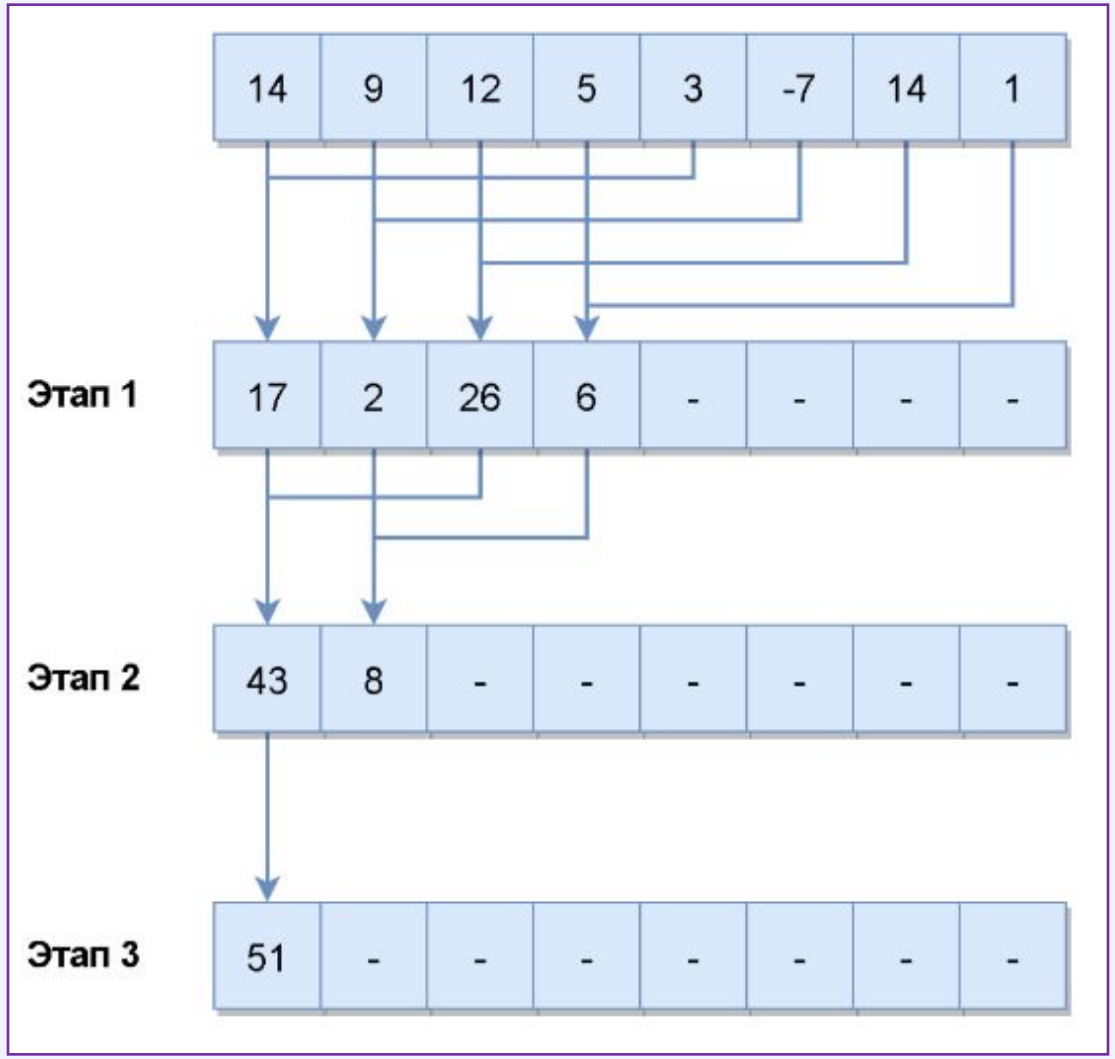
Ограничения на выполняемые алгоритмы при работе с GPU:

- Если мы выполняем расчет на GPU, то не можем выделить только одно ядро, выделен будет целый блок ядер.
- Все ядра выполняют одни и те же инструкции, но с разными данными, такие вычисления называются Single-Instruction-Multiple-Data или SIMD.
- Из-за относительно простого набора логических блоков и общих регистров, GPU очень не любит ветвлений.

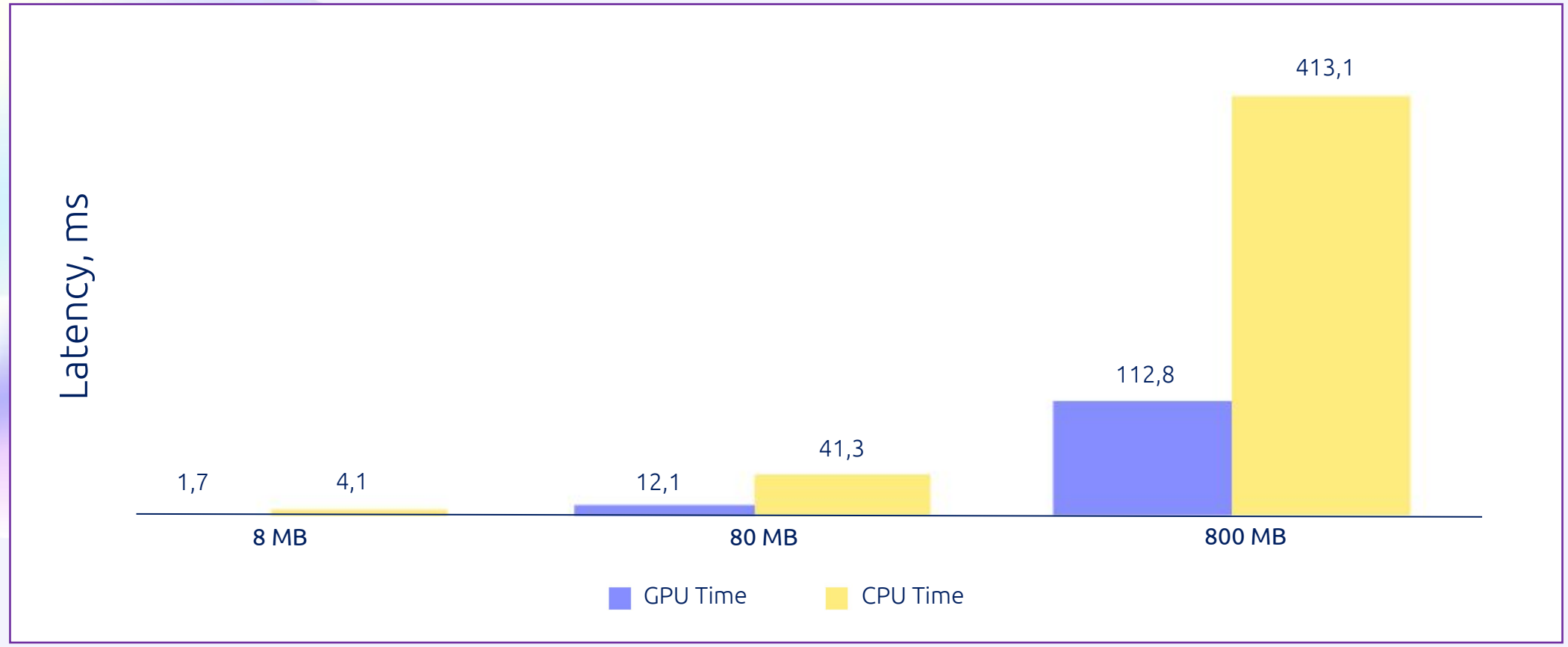
Возможности:

- Ускорение SIMD-вычислений

ПРИВЕДЕНИЕ КЛАССИЧЕСКИХ АЛГОРИТМОВ К SIMD-ПРЕДСТАВЛЕНИЮ



РЕЗУЛЬТАТЫ ВЫПОЛНЕНИЯ АЛГОРИТМОВ НА GPU



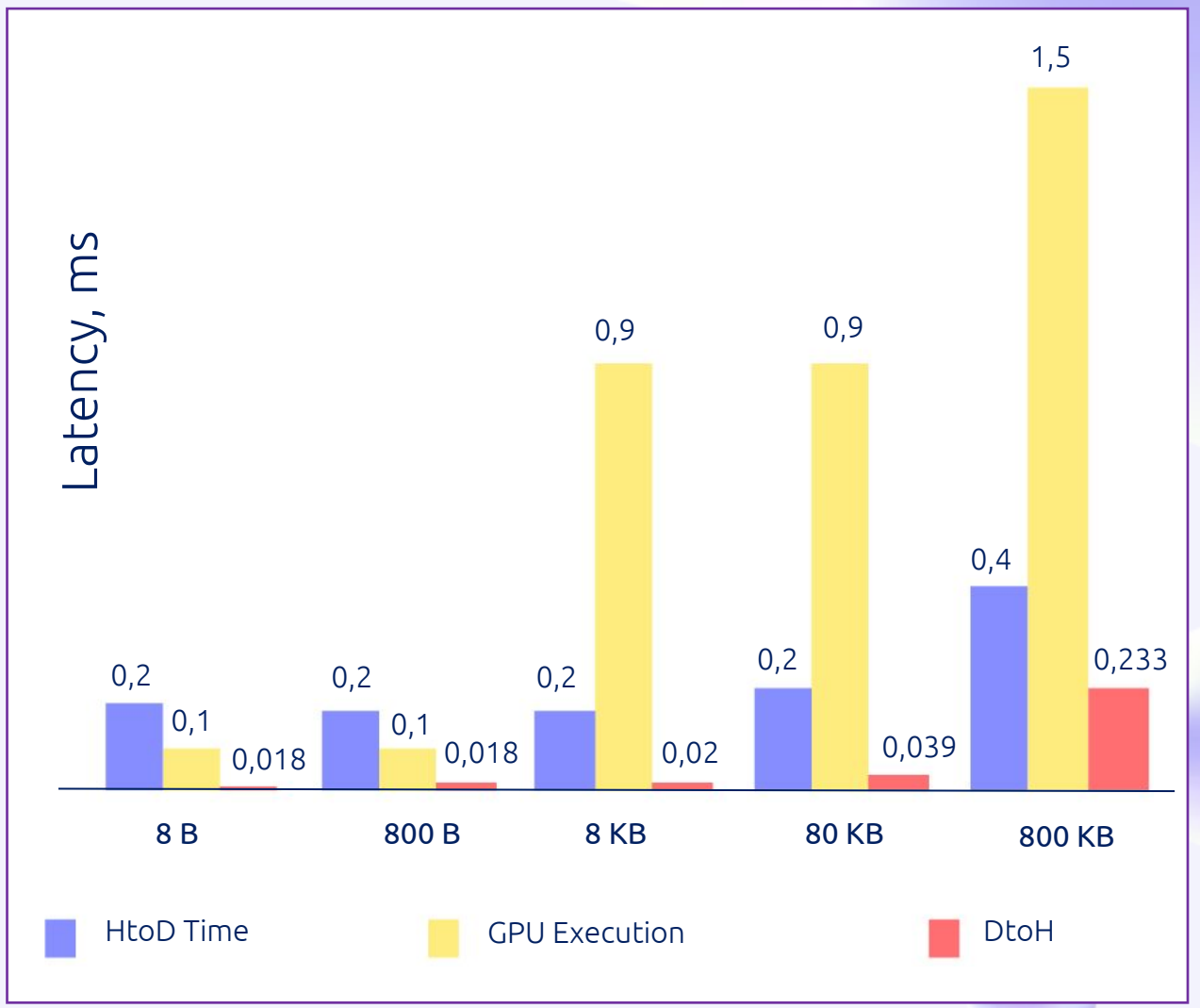
Время выполнения агрегации на GPU и CPU в мс

- 4992 CUDA ядра
- 24 GB памяти
- 480 Gb/s — пропускная способность памяти

ОВЕРХЕД НА ПЕРЕСЫЛКУ ДАННЫХ

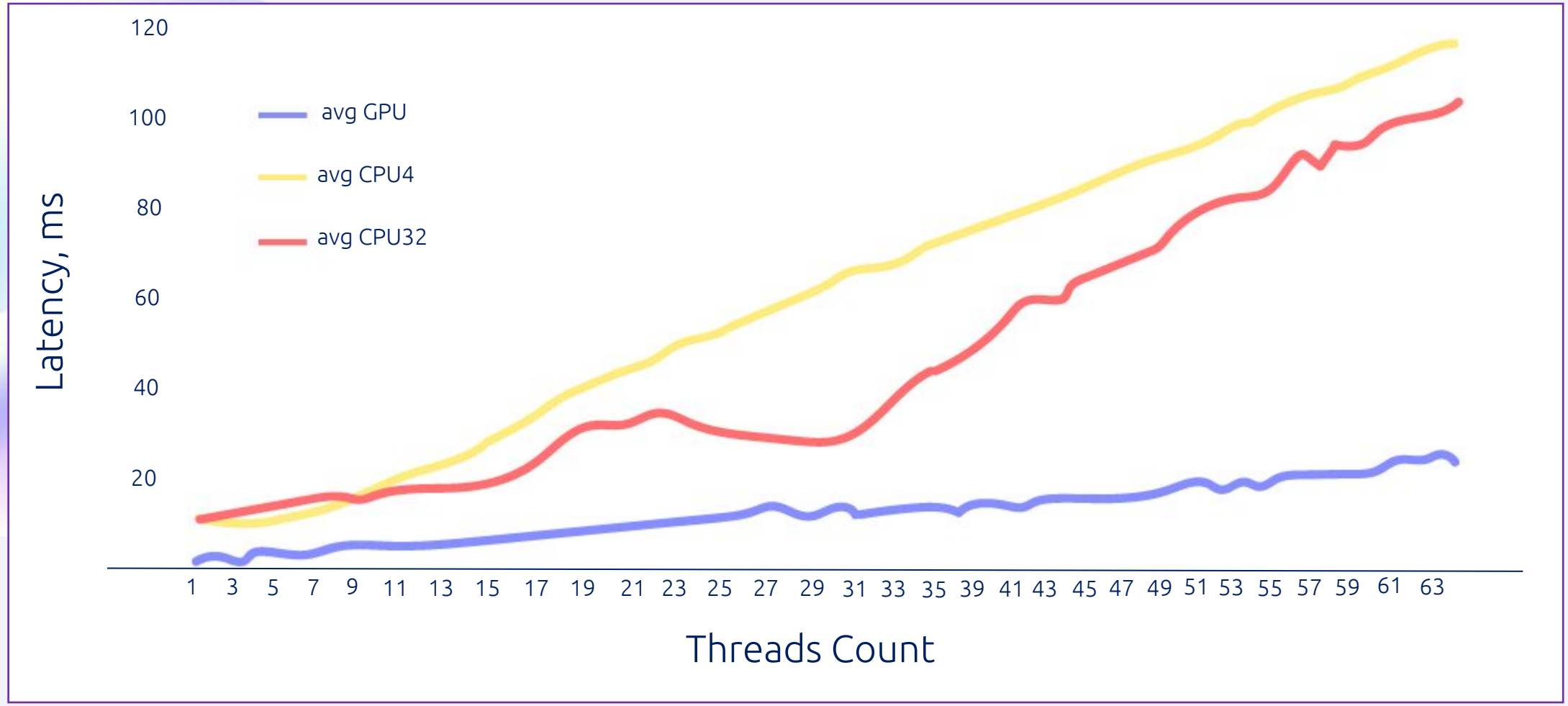
Memory bandwidth, или **пропускная способность памяти**, определяет теоретическую пропускную способность карты

HtoD — передаем данные на видеокарту
GPU Execution — сортировка на видеокарте
DtoH — копирование данных из видеокарты в оперативную память



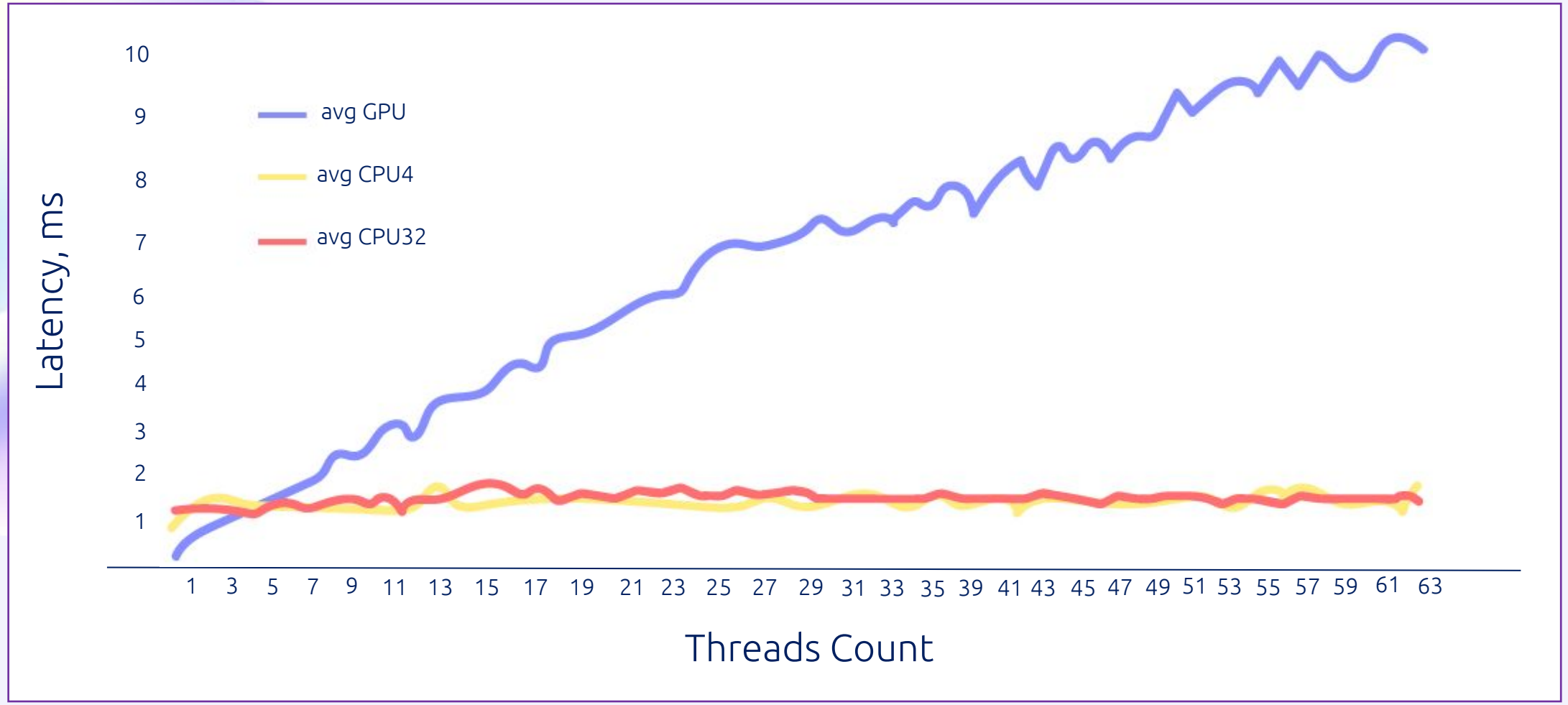
Тест видеокарты Tesla k80 в облаке Amazon

МНОГОПОТОЧНОСТЬ



Время выполнения математических расчетов на GPU и CPU с матрицами размером 1000 x 60 в мс

МНОГОПОТОЧНОСТЬ



Время выполнения математических расчетов на GPU и CPU с матрицами 10 000 x 60 в мс

РЕКОМЕНДАЦИИ

Если вы размышляете об использовании GPU в своих проектах, то GPU, скорее всего, вам подойдет если:

- Вашу задачу можно привести к SIMD-виду
- Есть возможность загрузить большую часть данных на карту до вычислений (закешировать)
- Задача подразумевает интенсивные вычисления

Заранее также стоит задаться вопросами:

- Сколько будет параллельных запросов
- На какое latency вы рассчитываете
- Достаточно ли вам одной карты для вашей нагрузки, нужен сервер с несколькими картами или кластер GPU-серверов

СПАСИБО ЗА ВНИМАНИ