CUDA (Computer Unified Device Architecture) — простыми словами

В этой статье вы узнаете о том, что такое *CUDA*, почему её изобретение считается революцией в мире компьютерной инженерии и как она работает изнутри.

Содержание:

Глава 1: графика

Чтобы понять, что такое CUDA и как она работает, нам нужно рассмотреть несколько важных концептов.

После тяжёлого рабочего дня я, как любой другой человек, хочу провести несколько часов, занимаясь любимым хобби. Будучи представителем профессии компьютерной инженерии, я делаю выбор в пользу ноутбука с недавно установленной на него Cyberpunk 2077. Первое, что бросается мне в глаза при запуске – графика, визуал, освещение и прорисовка 3D пространства. За всё это отвечает GPU, философию которого нам нужно понять в первую очередь.

GPU (Graphics Processing Unit) – это графический процессор, основная задача которого – обработка компьютерной графики.

Во всех современных 3D-играх для освещения, отрисовки мира используется GPU. К примеру, при игре разрешением 1080p * 1920 и частотой кадров 60 FPS на экране ежесекундно сменяется несколько миллионов пикселей. Чтобы провернуть такое, понадобится процессор, способный выполнять математические операции, простых самих по себе, но ошеломляющих количеством.

Именно эту задачу выполняет GPU: получив инструкцию, тысячи потоков процессора параллельно выполняют свою порцию задач, обеспечивая плавную смену кадров на мониторе. Как конкретно это происходит? По сути, GPU – это множество вычислительных блоков. Каждый вычислительный блок (далее – ВБ) имеет в себе ядра, отдельные блоки поменьше внутри ВБ. Такие блоки называются потоковыми и содержат в себе, кто бы мог подумать, потоки – мельчайшие единицы работы, выполняющие одну инструкцию, полученную от ядра. Таким образом, несколько вычислительных блоков могут выполнять простые задачи параллельно. Стоит сказать, что GPU имеет собственную иерархию памяти. Подробно рассматривать её в рамках этой статьи мы не будем, можно отметить лишь то, что ориентирована она на передачу большого объема данных.

Сделаем заметку на полях о том, что несмотря на свою приоритетную задачу (работа с графикой), под капотом GPU параллельно выполняет множество математических вычислений, это пригодится нам позже.

Глава 2: истоки

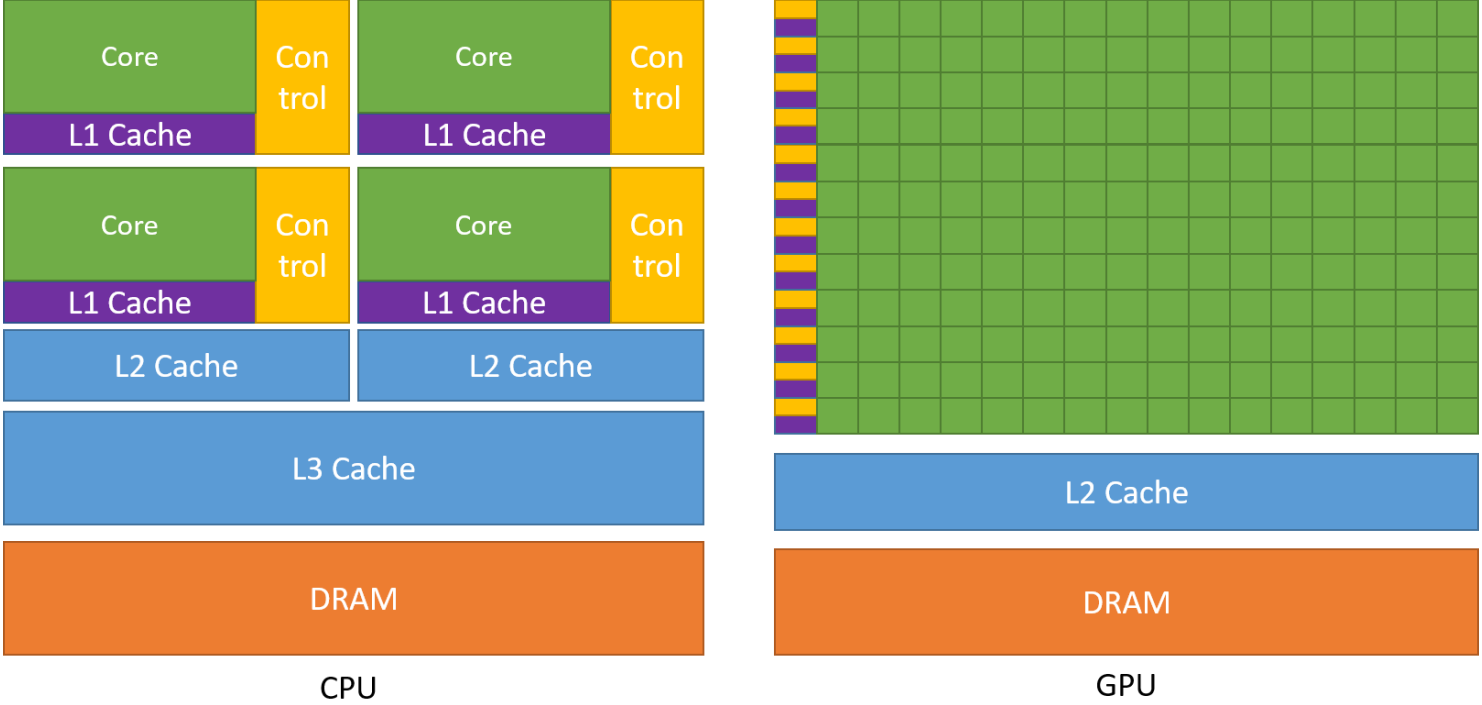
Выбрав предысторию своего персонажа и побегав по городу, я решаю начать продвигаться по сюжету игры. В этот раз меня поражает проработка поведения неигровых персонажей, то, как они обучаются уворачиваться от моих атак, заучивая мои паттерны передвижения.

За это отвечает следующий компонент философии CUDA, о котором сейчас поговорим – CPU. *CPU (Central Processing Unit)* – это центральный процессор, управляющий всеми компонентами системы компьютера, обрабатывающий данные и выполняющий инструкции программ.

Упрощённо механизм работы CPU выглядит как:

- Fetch (получение выбранной команды из оперативно памяти компьютера)
- Decode (расшифровка)
- Execute (выполнение команды)

До 2006 года за все вычислительные задачи внутри компьютера, в том числе за рендеринг графики, отвечал CPU. Несмотря на свою способность выполнять сложные операции, центральный процессор имеет в себе лишь несколько ядер, в то время как GPU, с которым мы уже знакомы, несколько тысяч. Низкая скорость обработки большого количества данных побудила компанию Nvidia создать новый тип процессора, рассчитанный на выполнение простых задач, но в большом количестве – так и появился GPU.



Ранее я упоминал, что для работы ядер графического процессора ему нужно получить инструкцию извне. Именно CPU говорит GPU, что надо делать, затем забирает результат работы в свою память.

Глава 3: CUDA

Наконец, имеем на руках оба камня бесконечности, необходимые нам для понимания CUDA и того, как она работает.

15 февраля 2007 года компанией Nvidia (да-да, той же самой, что год назад представила публике GPU) была выпущена первая версия CUDA, представляющая собой архитектуру параллельных вычислений. Разработка считается революционной, потому что позволяет использовать графические процессоры для решения общих вычислительных задач.

Вспомним сделанную нами чуть ранее заметку на полях: под капотом GPU – огромное кол-во математических операций. CUDA позволяет использовать механику GPU не только для работы с графикой, а куда шире, к примеру:

- ИИ и машинное обучение. Подобно рендерингу графики, это масса операций с матрицами.
- Анализ данных, конкретно работа с большим их объёмом.
- Другие сферы: медицина, метеорология, астрофизика, биоинформатика.

Глава 4: как?

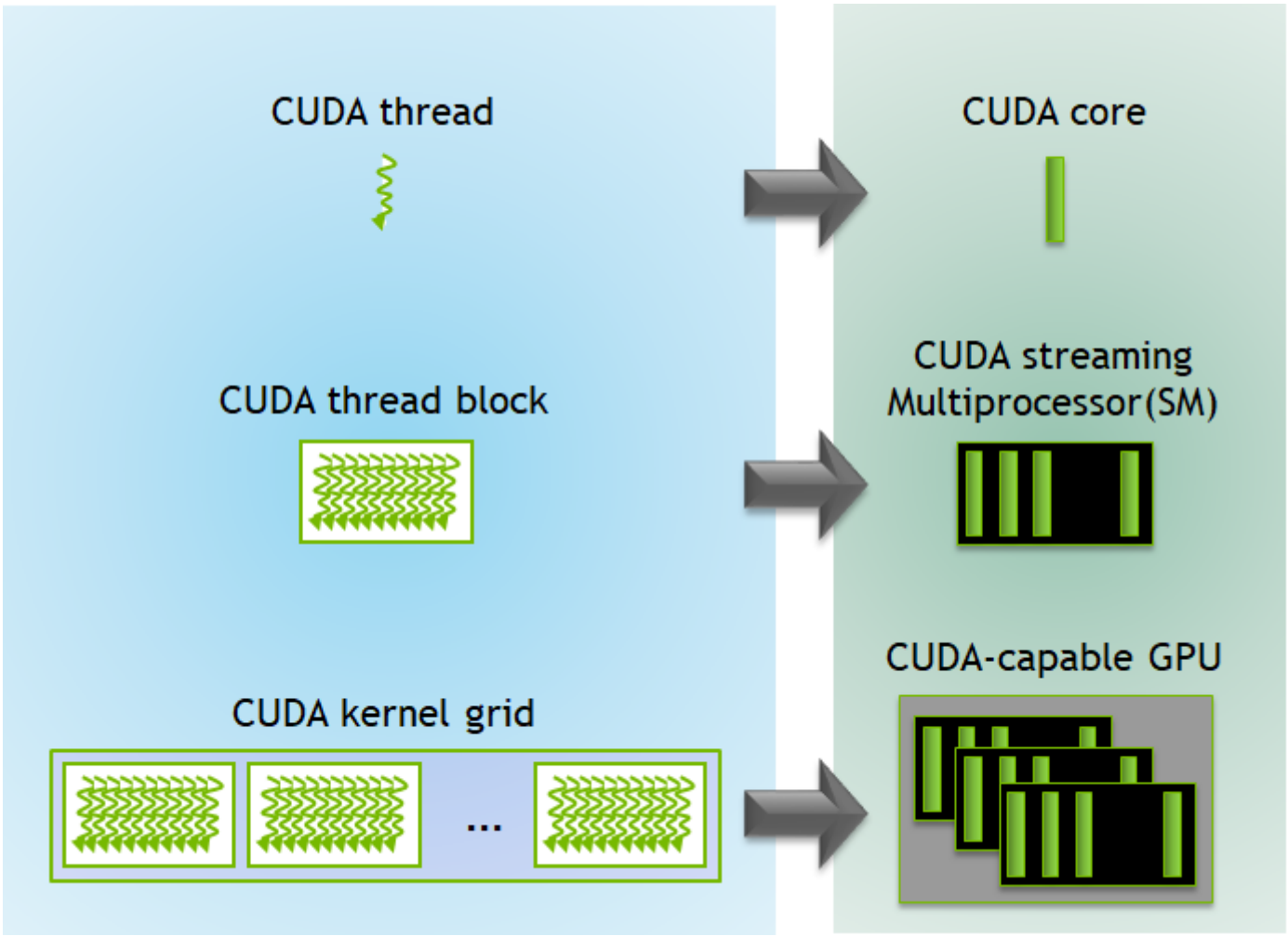
Чтобы научиться пользоваться столь мощным инструментом, для начала неплохо понять, как он работает, хотя бы упрощённо.

Сама по себе CUDA – это набор библиотек и инструментов, основанный на технологии GPGPU, которая, в свою очередь, и позволяет совершать операции над графическим процессором.

Архитектура CUDA состоит из:

- Ядра (программируемый процессор, выполняющий вычисления параллельно с другими потоками).
- Драйвера (программа, с помощью которой и происходит управление потоками ядер).
- Библиотек (коллекция алгоритмов для работы с CUDA, математических вычислений).

Механизм работы архитектуры довольно простой: сперва, пишем программу, которую будут выполнять ядра GPU на драйвер CUDA. Затем, центральный процессор даёт команду выполнять вычисления на GPU параллельно, начинают работать тысячи вычислительных блоков. Наконец, результат отдаётся центральному процессору для последующей записи в память.



Бинго!

Заключение

Основной задачей этой статьи было дать понимание того, что есть CUDA и вдохновить включить этот инструмент в свой арсенал, так как потенциал её огромен. Несмотря на стремительное развитие технологий и внедрение со стороны той же Nvidia инноваций, базовый принцип работы архитектуры будет оставаться неизменным; надеюсь, мне удалось объяснить его простым языком.

Источники:

- CUDA: как работает GPU

- CUDA Toolkit Documentation
- Все, что нужно знать про GPU: история технологии, архитектура графических процессоров и сферы их применения
- Что такое центральный процессор
- GPU architecture