**Министерство образования науки Российской Федерации**

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский авиационный институт (национальный исследовательский университет)»

«Информационные технологии и прикладная математика»

**РЕФЕРАТ**

По теме

«Видеокарта»

Выполнил:

Студент 1 курса группы: М80-106Б-21

Петров И. О.

Преподаватель:

Дубинин Алексей Владимирович

Москва, 2021

Оглавление

[Введение 3](#_Toc92675623)

[История создания 4](#_Toc92675624)

[1980-1990 – Зарождение графических ускорителей 4](#_Toc92675625)

[1990-2000 – Эволюция видеокарт 10](#_Toc92675626)

[Общее устройство видеокарт 21](#_Toc92675627)

[Графический процессор (GPU) 21](#_Toc92675628)

[Видеопамять 23](#_Toc92675629)

[Система питания 23](#_Toc92675630)

[Система охлаждения 24](#_Toc92675631)

[Интерфейсный разъем для подключения к материнской плате 26](#_Toc92675632)

[Разъем(ы) питания 26](#_Toc92675633)

[Устройство современных видеокарт 27](#_Toc92675634)

[CUDA 27](#_Toc92675635)

[Трассировка лучей 27](#_Toc92675636)

[Новые ядра 28](#_Toc92675637)

[Общий принцип работы видеокарты 29](#_Toc92675638)

[Этап 1. Растеризация. 29](#_Toc92675639)

[Этап 2. Текстурирование 30](#_Toc92675640)

[Этап 3. Пиксельный шейдер. 31](#_Toc92675641)

[Этап 4. Сохранение 31](#_Toc92675642)

[AMD vs Nvidia – сравнение архитектур 32](#_Toc92675643)

[Внутри кристаллов 32](#_Toc92675644)

[Управление ГП - Как все устроено внутри чипов 33](#_Toc92675645)

[Подсчет ядер по методу NVIDIA 36](#_Toc92675646)

[Трассировка лучей по-особенному 38](#_Toc92675647)

[Система памяти и многоуровневые кэши 40](#_Toc92675648)

[Так что же лучше? 42](#_Toc92675649)

[Мультимедиа-движок и видеовыход 42](#_Toc92675650)

[Созданы для вычислений, созданы для игр 43](#_Toc92675651)

[Выводы 44](#_Toc92675652)

[Список литературы 45](#_Toc92675653)

# Введение

Компьютерная графика. Услышав эти слова, мы представляем поражающие воображение спецэффекты из крупных блокбастеров, красивейшие модели персонажей из ААА-игр, и всё, что связано с визуальной красотой современных технологий. Но компьютерная графика, как и любой технологический аспект, развивалась не одно десятилетие, преодолев путь от отображения нескольких символов на монохромном дисплее до поражающих воображение пейзажей и героев, с каждым годом всё сложнее отличимых от реальности.

Видеокарта – это устройство, которое преобразует графический образ из памяти ПК или самой карты в форму, адаптированную для его демонстрации на экране монитора.

Сейчас под графическим адаптером понимают устройство с графическим процессором — графический ускоритель, который и занимается формированием самого графического образа. Современные видеокарты не ограничиваются простым выводом изображения, они имеют встроенный графический процессор, который может производить дополнительную обработку, снимая эту задачу с центрального процессора компьютера. Например, все современные видеокарты Nvidia и AMD (ATI) осуществляют рендеринг графического конвейера OpenGL, DirectX и Vulcan на аппаратном уровне. В последнее время также имеет место тенденция использовать вычислительные возможности графического процессора для решения неграфических задач.

Обычно видеокарта выполнена в виде печатной платы (плата расширения) и вставляется в слот расширения, универсальный либо специализированный (AGP, PCI Express). Также широко распространены и встроенные (интегрированные) в системную плату видеокарты — как в виде отдельного чипа, так и в качестве составляющей части северного моста чипсета или ЦПУ; в этом случае устройство, строго говоря, не может быть названо видеокартой.

# История создания

## 1980-1990 – Зарождение графических ускорителей

До 1980 персональный компьютер был роскошью, доступной не каждому. Но все начало меняться с момента выпуска IBM PC 5150 в 1981 году. Базовая версия поставлялась без монитора, жесткого диска и floppy-дисковода – их нужно было покупать отдельно. Также отдельно приобретались и платы расширения, среди которых были видеоадаптеры. Покупателю в то время предлагалось на выбор 2 видеоадаптера: MDA (Monochrome Display Adapter) и CGA (Color Graphics Adapter). Первый был более популярным – он работал с монохромными дисплеями, поддерживал исключительно текстовый режим – 80 столбцов на 25 строк и не имел графических режимов. CGA был полной противоположностью MDA – помимо работы в текстовом режиме он мог функционировать и в графическом. Причем поддерживалось как черно-белое, так и цветное изображение. По сути своей это была первая графическая карта с поддержкой 16 цветов.

Логическим продолжением MDA и CGA стало тоже решение IBM под названием EGA (Enhanced Graphics Adapter), представленное в сентябре 1984 года для нового персонального компьютера IBM PC/AT. По своей сути новый видеоадаптер стал первым в своем роде решением, способным воспроизводить нормальное цветное изображение. Так же, как и CGA, EGA поддерживал текстовый и графический режимы, при этом максимальное разрешение составляло 640x350 пикселов при использовании 16 цветов из 64 возможных. Для передачи данных применялась шина ISA. Вся память подразделялась на 4 сегмента (4 цветовых слоя). Процессор умел заполнять сегменты параллельно, что значительно повысило скорость заполнения кадра. Кстати, адаптер дополнительно оснащался 16 Кбайт памяти для расширения графических функций BIOS.

Архитектурно адаптер EGA не был полностью совместим с MDA и CGA, однако ограниченно поддерживал работу этих стандартов. Например, к EGA можно было подключить MDA-монитор, но в таком случае поддерживался только один режим работы с разрешением 640x350 пикселов и монохромной графикой.

Занимательно и то, что в то время компания IBM не регистрировала никакие патенты на свои графические разработки. Это означало, что на рынке видеоадаптеров появлялось большое количество «клонов» решений IBM. В наше время такое попросту невозможно! Клоны адаптеров отличались улучшенными характеристиками и порой были избавлены от редких глюков в работе, свойственных адаптерам IBM. К примеру, некоторые клоны были обучены работе в более высоком разрешении 720x540 пикселов. Выпуском клонов занимались многие компании. В их число входили и такие известные производители, как ATI Technologies и подразделение компании Western Digital — Paradise.

Принято считать, что эра EGA продолжалась вплоть до выпуска адаптера VGA (Video Graphics Array), то есть до 1987 года. Однако в этом году было выпущено еще одно графическое решение — MCGA (Multi Colour Graphics Array). Оно было представлено как видеоадаптер для ранних моделей компьютера IBM PS/2, причем он был встроен в чипсет — дискретных решений MCGA не было выпущено вовсе. MCGA обладал полной поддержкой всех режимов CGA. Вдобавок, адаптер мог работать в монохромном режиме с разрешением 640x480 пикселов и частотой обновления 60 Гц. Кроме того, поддерживался 256-цветной режим с разрешением 320x200 и частотой развертки 70 Гц. 256-цветной режим впоследствии стал очень популярным среди игр. Однако во время появления MCGA многие игры еще не поддерживали такой режим и работали лишь в 4-цветном режиме CGA.

Кстати, именно из-за этого ограничения возникла необходимость использования аналогового сигнала. В качестве разъема MCGA использовал DB-15 семейства D-Sub, в то время как все предшественники использовали цифровой сигнал 9-контактного разъема. Стандарт MCGA так и не стал таким же распространенным, как предыдущие решения IBM. В первую очередь, это случилось из-за политики самой компании: IBM не предоставляла лицензии на производство адаптеров данного типа сторонним производителям.

Однако в том же 1987 году IBM представила в своем роде революционный продукт на базе стандарта VGA. Так же, как и MCGA, адаптеры использовались в новейших на то время компьютерах IBM PS/2. В плане архитектуры VGA схож со своим предшественником EGA. Он состоит из следующих основных подсистем: графического контроллера, памяти, секвенсора, контроллера атрибутов, синхронизатора и контроллера ЭЛТ. Что касается их функционала, то графический контроллер отвечал за обмен данными между центральным процессором и видеопамятью. Секвенсор, или, как его еще называют, сериализатор, преобразует данные из памяти в поток битов, которые впоследствии передаются контроллеру атрибутов. Контроллер атрибутов, в свою очередь, преобразует входные данные в цветовые значения. Синхронизатор управляет временными параметрами адаптера, а также переключает цветовые слои. Ну а контроллер ЭЛТ генерирует сигналы синхронизации для дисплея. Интересно, что основные подсистемы стали располагаться на одной микросхеме. Это позволило уменьшить размеры видеоадаптера. Однако тут стоит отметить, что дискретные VGA-ускорители выпускали лишь сторонние производители. Так, IBM в своих компьютерах встраивала адаптеры в материнскую плату.

Как и в случае с MCGA, видеоадаптер VGA использовал не цифровой, а аналоговый интерфейс. Выбор в пользу D-Sub пал из-за возросшего количества отображаемых цветов. Так, цифровой интерфейс передавал RGB-сигналы управления тремя основными цветами, а с использованием аналогового варианта появилась возможность присвоить каждому сигнал определенный уровень яркости. В результате количество отображаемых цветов возросло до отметки в 262 144. На то время это был большой шаг вперед в плане увеличения реалистичности изображения.

Само собой, из-за увеличения количества отображаемых цветов появились новые графические режимы. Стандартными для VGA являлись режимы: с разрешением 640x480 пикселов и 16 цветами; с разрешением 640x350 и 16 цветами (режим обратной совместимости с EGA); с разрешением 320x200 точек и 16 цветами; с разрешением 320x200 точек и 256 цветами. Последний режим допускал хранение в памяти до 4 страниц одновременно, каждая из которых весила чуть менее 64 Кбайт. В 1991 году программист Майкл Абраш сумел запустить на видеоадаптере VGA режим с разрешением 320x240 и 360x480 пикселов и 256 цветами. При этом использовалась более эффективная организация работы с памятью.

Интересно, что VGA стал первым видеоадаптером, который работал с разрешениями с отношением сторон 4:3 (прежде пикселы растягивалась по вертикали). Что касается текстовых режимов, то в VGA поддерживались несколько видов шрифтов и режимов, что позволяло использовать различные их комбинации. Стандартный шрифт VGA имеет разрешение 8x16 пикселов. Также поддерживаются шрифты с разрешением 8x14 и 8x8 точек для обратной совместимости с EGA и CGA соответственно.

Стандарт развивался очень быстро. На рынке появилось множество различных клонов видеоадаптеров от сторонних производителей. В отношении VGA компания IBM проводила такую же политику, как и с адаптерами MDA, CGA и EGA, предоставляя другим компаниям лицензию на производство доработанных решений. В целом VGA стал первым «народным» адаптером, аббревиатуру которого используют и по сей день. А под разрешением VGA подразумевается разрешение 640x480 точек

В 1987 году IBM представила еще один, но уже «профессиональный» видеоадаптер 8514/A. В отличие от других решений, данный ускоритель не был совместим ни с одним из предыдущих решений компании. На выбор пользователю предоставлялись две версии 8514/A с различным объемом видеопамяти: 512 Кбайт или 1 Мбайт. Поддерживались только два разрешения: 640x480 и 1024x768 точек, причем младшая версия адаптера работала с 16 цветами, а старшая поддерживала все 256 цветов. В высоком разрешении адаптер функционировал с малой частотой развертки (43 Гц), что привело к проблеме мерцания изображения. Инженеры IBM запрограммировали адаптер таким образом, чтобы использовать в системе более дешевые мониторы - в то время дисплеи с поддержкой высокого разрешения и высокой частоты развертки стоили очень дорого. Тем не менее стоит отметить, что у многочисленных клонов 8514/A данное ограничение было снято. Важной особенностью 8514/A считалась поддержка аппаратного ускорения рисования. Так, 2D-акселератор ускорял рисование линий, прямоугольников, заливку различных фигур, а также поддерживал функцию BitBLT.

Стоимость IBM 8514/A составляла 1290 долларов США за версию с объемом памяти 512 Кбайт. Для того, чтобы приобрести мегабайтное решение покупателю необходимо было доплатить еще 270 долларов. Однако на низком распространении адаптера сказывалась не столько цена, сколько его работа лишь в системах с шиной MCA (Micro Channel Architecture). Однако к концу 80-х годов появилось большое количество клонов. Многие из них имели поддержку шины ISA, а также дополнительных разрешений (800x600 и даже 1280x1024 точек). Среди видеоадаптеров-близнецов опять выделялись решения компаний Western Digital и ATI. Особенно популярны стали решения Mach 8 и Mach 32.

8514/A никогда не был коммерчески успешным, однако он внес большой вклад в развитие аппаратного ускорения графики, «бум» которого пришелся на 90-е годы. Развитием стандарта VGA стали адаптеры SVGA (Super Video Graphics Array) и XGA (Extended Graphics Array). Вторые по списку ускорители впервые были представлены в октябре 1990 года и использовались в компьютерах IBM PS/2. Они устанавливались в Model 90 и 90 XP, но также были доступны в качестве комплектующих. По большому счету, XGA представлял собой расширение стандарта VGA, то есть был своеобразным гибридом 8514/A.

XGA выгодно отличало использование видеопамяти типа VRAM — она была быстрее, чем применявшаяся в VGA DRAM. Ее объем составлял 512 Кбайт. Стоил ускоритель 1095 долларов США. Интересно, что за дополнительную плату в 350 долларов пользователь получал модель с увеличенным до 1 Мбайт объемом видеопамяти. XGA поддерживал разрешение 640x480 точек с 16-битным цветом (всего 65 536 оттенков). Помимо этого, была поддержка 256-цветного изображения с разрешением 1024x768 точек. Интересно, что не поддерживалось промежуточное разрешение 800x600 точек. Поскольку XGA частично перенял функции 8514/A, он также был способен аппаратно ускорять рисование. Графический ускоритель поддерживал алгоритм Брезенхэма (алгоритм, определяющий, какие точки двумерного растра нужно закрасить, чтобы получить близкое приближение прямой линии между двумя заданными точками), заливку прямоугольников, а также операцию BitBLT. Отличительной от 8514/A особенностью стала возможность рисования объектов произвольной формы.

Что касается недостатков XGA, то одним из них являлось то, что адаптер использовал развертку с чередованием в высоком разрешении. Это приводило к тому, что на мониторе было заметно мерцание из-за снижения частоты регенерации. В сентябре 1992 года был представлен обновленный стандарт XGA с индексом 2. XGA-2 не претерпел никаких кардинальных изменений. Объем памяти был увеличен до 1 Мбайт, причем использовалась еще более быстрая VRAM. Вкупе с движком, отвечающим за аппаратное ускорение рисования, это дало значительную прибавку производительности в некоторых задачах. Видеоадаптер научился работать в высоком разрешении 1024x768 точек с 16-битным цветом и высокой частотой развертки — так была решена проблема мерцающего экрана. Наконец, была добавлена поддержка разрешения 800x600.

Однако с выходом стандарта XGA конкуренты IBM решили не копировать эти устройства, а производить более дешевые видеоадаптеры, но способные функционировать в более высоком разрешении и/или с большим количеством цветов. Эти адаптеры и образовали новый класс устройств под названием SVGA.

Изначально SVGA не являлся стандартом — уж больно различны были все устройства, поскольку для них не существовало четких спецификаций. Само собой, такое положение дел привело к сложности программирования. В 1989 году ассоциация производителей VESA (Video Electronic Standards Association) предложила ввести единый программный интерфейс для всех видеоадаптеров SVGA. Этот интерфейс получил название VESA BIOS Extention. С его помощью программисты могли определять специфические соответствия и использовать их в дальнейшем. При этом для работы с любым SVGA-устройством использовался единый драйвер. Стандарт VESA предусматривал использование всех разрешений, включая самое высокое 1280x1024 точек с 16777216 оттенками (24-разрядное кодирование цвета). Отличительной чертой решений SVGA стал встроенный акселератор, появление которого было связано с необходимостью качественной обработки графической составляющей новых ОС. Например, только набирающей в то время обороты Microsoft Windows.

Время шло, и качеством двухмерной графики уже было не удивить. В какой-то степени она себя исчерпала. Вполне естественно, что инженеры начали делать упор на обработку трехмерного изображения. Изначально 3D-ускорители представляли собой отдельные платы, вставляемые в свободные разъемы на материнской плате. Однако со временем микросхемы 3D-ускорителя перекочевали в графический чип, и одна плата отвечала за ускорение как двухмерной, так и трехмерной графики.

## 1990-2000 – Эволюция видеокарт

Развитие трехмерной графики и ее аппаратного ускорения существенно перекроили весь рынок видеокарт. IBM больше не являлась ведущим производителем. Так, на первый план вышли устройства сторонних разработчиков. Многие из них, кстати, прежде занимались копированием адаптеров IBM.

Принято считать, что одним из первопроходцев среди 3D-ускорителей для массового рынка стала видеокарта компании S3 под названием ViRGE (Virtual Reality Graphics Engine), выпущенная в 1995 году. В этом же году была представлена консоль Sony PlayStation, поэтому рынок персональных компьютеров, можно сказать, уже был в ожидании устройств, которые позволили бы ускорять трехмерную графику в реальном времени. S3 выпускала ViRGE именно для этой задачи. В сравнении с сегодняшними видеокартами характеристики S3 ViRGE могут лишь вызвать улыбку. Частота графического процессора составляла всего-навсего 66 МГц. Шина памяти — 64-бит, объем — 8 Мбайт. В качестве интерфейса использовался PCI.

Несмотря на то, что S3 ViRGE был выпущен как 3D-ускоритель, лучше всего он работал в режиме 2D. С задачами обработки трехмерного изображения он справлялся значительно хуже. Например, его производительность сильно проседала при применении билинейной фильтрации. Из-за своей скорости S3 ViRGE даже получил среди пользователей прозвище «деселератора» (от англ. decelerator — тормоз). Тем не менее видеокарта отлично справлялась с 2D-контентом. Например, с обработкой графического интерфейса Windows. А для комфортной игры многие пользователи использовали связку S3 ViRGE и вышедшую несколько позже 3Dfx Voodoo Graphics.

Годом позже канадская компания ATI представила собственную разработку — 3D-акселератор Rage. В основе первого поколения этой видеокарты лежал чип Mach64, который был заточен под ускорение двухмерной графики. Для использования в Rage он был немного доработан, а именно обзавелся поддержкой 3D и функцией ускорения видео формата MPEG-1.

Однако в истории оставило след второе поколение устройств ATI — Rage II. В его основе лежало все то же ядро Mach64, но значительно переработанное. Чип работал с различными типами памяти: EDO, SDRAM и SGRAM. Ее объем мог составлять 2, 4 или 8 Мбайт. В случае использования памяти типа SGRAM ее частота составляла 83 МГц, в то время как графическое ядро функционировало на частоте 60 МГц. В целом оптимизация архитектуры графического процессора позволила повысить производительность второго поколения видеокарт Rage примерно на 20% в двухмерном режиме в сравнении с первым поколением. Rage II поддерживал библиотеки Direct3D и OpenGL. Также была добавлена поддержка формата MPEG-2 и некоторых полезных функций для рендеринга: фильтрации текстур, альфа-смешения и некоторых других возможностей. Интересно, что на видеокартах Rage II также располагался специальный чип ImpacTV для оцифровки ТВ-сигнала. Помимо использования в IBM-совместимых компьютерах, она использовалась в некоторых компьютерах Apple Macintosh, а именно в моделях G3 и Power Mac 6500.

Сама линейка видеокарт Rage II была представлена тремя моделями: IIC, II+, II+DVD. Различались они частотой процессора и памяти, а также ее типом. Производительность каждой из моделей разнилась, но даже 2-мегабайтная версия Rage II в среднем была на 20% производительнее ViRGE при обработке трехмерного изображения. Тем не менее продукты ATI уступали в объеме продаж адаптерам S3. Более того, у видеокарт Rage II были проблемы с драйверами для ОС Windows. Как видите, распространенное мнение о нестабильности драйверов ATI (а теперь и AMD) берет свое начало в 90-х годах.

А в 1996 году свет увидела легендарная видеокарта Voodoo Graphics (или Voodoo1) от компании 3Dfx. В то время фирма являлась самой молодой на рынке графических ускорителей. Она была основана только в 1994 году тремя выходцами из компании Silicon Graphics. На первых порах 3Dfx занималась созданием чипов для игровых автоматов. Первым таким устройством стал симулятор бейсбола ICE Home Run Derby, а затем появились и другие игровые решения, среди которых были San Francisco Rush и 3D Hockey. Чипы 3Dfx обеспечивали великолепную на то время графику — разработки компании заинтересовали даже производителей приставок.

Изначально в планах 3Dfx не значился выход на рынок персональных компьютеров. Однако в 1996 году цены на производительную память типа EDO сильно упали. Это позволило 3Dfx создать высокопроизводительное решение с умеренной ценой. Так, была выпущена видеокарта Voodoo Graphics. К выпуску видеокарты приложила руку также компания Diamond Multimedia, которая обеспечивала 3Dfx своими производственными мощностями. Графический процессор и память Voodoo Graphics работали на частоте 50 МГц, объем памяти типа EDO составлял 4 Мбайт. Интерфейс памяти был 64-битным. Интересно, что Voodoo обладала отдельным текстурным модулем. Такая архитектура позволила видеокарте работать быстро и показывать картинку очень высокого качества.

Вместе с Voodoo компания 3Dfx представила программный интерфейс Glide — альтернативу другим API в лице Direct3D и OpenGL. Glide основывался на библиотеке OpenGL. По своей сути, он был урезанной версией OpenGL. Данный стандарт содержал огромное количество различных функций, Инженеры 3Dfx собрали в Glide только самое необходимое для визуализации трехмерных игр в реальном времени, причем все функции их библиотеки в полной мере поддерживались видеокартами Voodoo, что значительно облегчало программирование. В то же время это накладывало определенные ограничения на разработчиков: к примеру, максимальная глубина цвета в приложениях составляла лишь 16 бит. Сотрудничество 3Dfx с разработчиками игр привело к тому, что видеокарты Voodoo доминировали на рынке игровых продуктов для персональных компьютеров. Так, под работу с Voodoo были заточены игры серии Tomb Raider от Eidos. Еще больше упрочил лидирующие позиции Voodoo выпуск драйвера MiniGL. Он предназначался исключительно для аппаратного ускорения новейшей игры Quake.

Сегодня трудно представить рынок видеокарт без компании NVIDIA. Но в середине 90-х годов она только начинала набирать свои обороты. Ее восхождение началось с выпущенной все в том же 1997 году видеокарты RIVA 128 (Real-time Interactive Video and Animation accelerator). Сама компания была основана в 1993 году, но в первые годы ее устройства не снискали никакого успеха. Так, выпущенный в 1995 году чип NVIDIA NV1 использовал иную технологию рендеринга, основанную на квадратичном маппинге текстур. Такой тип рендеринга был несовместим с библиотекой Direct3D, и чип NV1 так и остался непопулярным среди пользователей. RIVA 128 стала первым успешным продуктом NVIDIA.

RIVA 128 во многом была полной противоположностью NV1. Она была заточена под работу с Direct3D и стандартом OpenGL. Сам графический процессор RIVA 128 изготавливался по 350-нм технологическим нормам и содержал 3,5 миллиона транзисторов. Его частота составляла 100 МГц. Видеокарта использовала память SGRAM, объем которой составлял 4 Мбайт, а частота — 100 МГц. Ширина шины памяти составляла 128 бит. В отличие от видеокарты Voodoo Graphics, RIVA 128 работала не только через интерфейс PCI, но и новейший на то время порт AGP 1x. Главным качественным отличием RIVA 128 являлось то, что видеокарта сочетала в себе функции как 2D-, так и 3D-ускорителя.

3D-ускоритель использовал прогрессивную разделяемую память. Это обеспечило поддержку высоких разрешений (800x600, 960x720 точек) в трехмерном режиме. Для сравнения: видеокарта Voodoo в режиме 3D работала с максимальным разрешением 640x480 пикселов. В плане производительности и качества изображения RIVA 128 немного уступала Voodoo, но и цена устройства NVIDIA была ниже. Кстати, благодаря низкой стоимости и одновременной поддержке 2D и 3D многие OEM-производители делали выбор в пользу решения NVIDIA. В начале 1998 года NVIDIA выпустила небольшую модификацию видеокарты с индексом ZX. RIVA 128ZX поддерживала интерфейс AGP 2x, а объем ее видеопамяти составлял 8 Мбайт вместо прежних 4. Плюс ко всему, чип RAMDAC работал на частоте 250 МГц, в то время как у RIVA 128 частота RAMDAC составляла 206 МГц.

В феврале 1998 года на рынок был выведен новый графический ускоритель Voodoo2, который стал логическим продолжением адаптера Voodoo Graphics. После провала Voodoo Rush в 3Dfx решили остаться верными собственной философии. Поэтому чип Voodoo2, как и Voodoo Graphics, работал только с трехмерным изображением. Был добавлен еще один отдельный текстурный модуль, что значительно увеличило производительность видеокарты. Особенно это было заметно в играх Quake II и Unreal. В остальном Voodoo2 не преподнес никаких сюрпризов, хотя его заявленные характеристики были более чем впечатляющие. Частота чипа составляла 90-100 МГц, в качестве памяти использовалась EDO DRAM. Voodoo2 выпускалась в двух версиях: с 8 и 12 Мбайт «мозгов». На практике объем памяти несильно сказывался на производительности — оба устройства были очень быстры. Каждый из чипов имел 64-разрядную шину памяти. Максимальное разрешение в режиме 3D было увеличено до 800x600 точек. Поддерживалась трилинейная фильтрация. Что касается интерфейса, то Voodoo2 использовал уже уходящий в прошлое PCI.

Вместе с Voodoo2 компания 3Dfx представила нам знакомую и эффективную по сей день технологию SLI (Scan-Line Interleave). SLI, как и в настоящее время, обеспечивала одновременную работу двух графических ускорителей Voodoo2 в системе. Ускорители распределяли всю нагрузку по обработке изображения между собой: одна карта обрабатывала верхнюю часть картинки, другая — нижнюю. Это позволило увеличить максимальное разрешение в 3D-режиме до 1024x768 точек. Само собой, теоретически вырастала и производительность всей системы. На практике же дела обстояли не так хорошо. Во-первых, учитывая 3D-ориентированность видеокарт Voodoo2, в системе необходимо было обеспечить стабильную работу сразу трех видеоадаптеров. Элементарно не все материнские платы имели столько свободных портов. Во-вторых, проблемы возникали на программном уровне: далеко не все приложения корректно работали при использовании SLI. По этим причинам SLI не получил в то время должного распространения. Однако уже в двухтысячных годах под крылом компании NVIDIA технология раскрыла весь свой потенциал.

Интересно, что в третьем поколении видеокарт Voodoo произошла своеобразная дифференциация: существовало три версии Voodoo3 с индексами 2000, 3000 и 3500. Различались они частотой ядра, памяти и блока RAMDAC. Так, Voodoo3 2000 функционировал на частоте 143 МГц (ядро и память) и 300 МГц (RAMDAC). Модель с индексом 3000 имела частоты 166 МГц и 350 МГц для ядра/памяти и RAMDAC. Частоты топовой версии Voodoo3 3500 составляли 183 МГц и 350 МГц соответственно. К сожалению, видеокарта Voodoo3 стала последним успешным решением 3Dfx. После этого эпоха компании начала стремительно приближаться к своему завершению.

Что касается компании NVIDIA, то после успешного запуска RIVA 128 они ничуть не думали снижать обороты. На выход Voodoo2 NVIDIA подготовила достойный ответ в лице Riva TNT. Новый чип, имеющий кодовое название NV4, был изготовлен по 350-нм технологическому процессу, содержал 7 миллионов транзисторов, а его частота составила 90 МГц. Изначально NVIDIA планировала установить частоту на отметке 110 МГц, но из-за большого нагрева тестовых образцов было решено немного опустить эту планку. В качестве чипов памяти использовались модули SDRAM, их суммарный объем составлял 16 Мбайт. Частота памяти равнялась 110 МГц, а ширина шины памяти — 128 бит. В сравнении с Voodoo2, RIVA TNT поддерживала больше функций: например, 32-битный цвет и текстуры с разрешением 1024x1024 точек. Также видеокарта обзавелась поддержкой интересной технологии Twin-Texel, которая позволяла накладывать две текстуры на один пиксел за такт в режиме мультитекстурирования. Благодаря Twin-Texel скорость заполнения выросла примерно на 40%. Также RIVA TNT поддерживала три типа фильтрации: билинейную, трилинейную и анизотропную.

Несмотря на выдающиеся характеристики, RIVA TNT оказалась чуть медленнее, чем Voodoo2. Более того, популярность устройства NVIDIA также была ниже. Ситуация во многом повторяла таковую с RIVA 128. Большинство игр работали посредством библиотеки Glide, а ее официально поддерживали лишь видеокарты Voodoo. Однако в других приложениях RIVA TNT совсем чуть-чуть уступала Voodoo2, а благодаря поддержке 32-битного цвета порой выигрывала у нее в качестве изображения.

На рынке также появилась менее производительная версия RIVA TNT под названием Vanta. По характеристикам она, естественно, уступала RIVA TNT: были снижены частоты ядра и памяти, а также урезана до 64 бит шина памяти (использовались дефектные чипы, которые не могли стабильно работать на более высоких скоростях). Это позволило NVIDIA занять нишу менее производительных продуктов. В пользу Vanta выбор делали производители OEM-компьютеров.

Еще одним интересным нововведением стало то, что NVIDIA начала выпускать собственные драйверы Detonator для линейки RIVA TNT. В то время видеокарты «зеленых» отлично работали в связке с системами Intel, однако не могли проявить себя с лучшей стороны в системах AMD (в отличие от адаптеров Voodoo). Драйверы Detonator включали в себя определенные оптимизации, которые позволили поднять производительность RIVA TNT в системах AMD на 30% в игре Quake II. Да и в целом драйверы повысили скорость видеокарт NVIDIA в различных OpenGL- и DirectX-приложениях. Также Detonator решал вопрос совместимости RIVA TNT с различными материнскими платами.

В 1999 году NVIDIA выпустила видеокарту TNT2. Инженеры «зеленых» не преследовали цель сделать абсолютно новое устройство. TNT2 был своего рода ремейком RIVA TNT, работой над ошибками. Так, чип NV5 изготавливался уже в соответствии с 250-нм технологическими нормами, что позволило поднять его частоту до 150 МГц. Был доработан блок рендеринга и поднята частота RAMDAC до 300 МГц, что обеспечило работу видеокарты в сверхвысоких разрешениях. Также добавилась поддержка интерфейса AGP 4x. В остальном TNT2 была схожа со своим предшественником. Объем SDRAM составлял до 32 Мбайт, использовалась 128-битная шина памяти. Видеокарта поддерживала 32-битный цвет, а также текстуры с разрешением 2048x2048.

Всего на рынок было выведено четыре модификации TNT2. Одна из них являлась стандартной, вторая — топовой (с приставкой Ultra), третья и четвертая — урезанные (Vanta и M64). TNT2 Ultra отличалась от стандартной версии лишь повышенными частотами. Так, скорость работы ядра и памяти у нее составляли 150 МГц и 183 МГц против 125 МГц и 150 МГц у TNT2. Что касается версий Vanta и M64, то так же, как и в предыдущем поколении видеокарт TNT, у этих модификаций была урезана шина памяти до 64 бит, но и снижены частоты ядра и памяти. Различия между ними заключается в том, что Vanta — еще более упрощенный вариант видеокарты с более низкими частотами. Опять же, эти устройства стали крайне популярны в OEM-сегменте.

В целом видеокарта TNT2 стала самым успешным продуктом NVIDIA на то время. Да, в 1999 году большую роль все еще играла библиотека Glide, за счет чего Voodoo3 еще удерживала лидирующие позиции. Однако становилось очевидным то, что продукты NVIDIA проигрывают в производительности решениям 3Dfx лишь самую малость, предлагая при этом более богатую функциональность. Пока 3Dfx и NVIDIA боролись за лидирующие позиции, несколько компаний тихо создавали и выпускали свои графические решения, пытаясь противостоять топовым устройствам.

В 1999 году ATI выпустила видеокарту Rage 128, которая должна была стать конкурентом RIVA TNT. Чип изготавливался по 350-нм техпроцессу и имел 128-битную шину памяти. Частоты ядра и памяти составляли 103 МГц, RAMDAC — 250 МГц. Объем памяти составлял до 32 Мбайт. Видеокарта имела очень богатую функциональность. Поддерживалась однопроходная трилинейная фильтрация и аппаратное ускорение DVD-видео. Rage 128, как и RIVA TNT, поддерживал 32-битный цветовой режим. Более того, он успешно конкурировал с видеокартой NVIDIA в этом режиме. Также Rage 128 обладала поддержкой технологии Twin Cache Architecture. Она объединяла кэш-память пикселов и текстур для увеличения полосы пропускания. Тем не менее, видеокарту ATI в худшую сторону отличало качество изображения.

Годом ранее канадская компания Matrox также представила свой 3D-ускоритель G200. В первой половине 90-х годов Matrox являлась одним из лидеров по производству 2D-ускорителей, однако затем отстала от своих новоиспеченных конкурентов. В архитектуре G200 использовалось множество интересных технологий. Так, одной из них являлась SRA (Symmetric Rendering Architecture). Она обеспечивала чтение и запись графических данных в системную память, что обеспечивало очень высокую скорость работы. Также видеокарта поддерживала технологию VCQ (Vibrant Color Quality). Ее суть заключалась в том, что при визуализации использовался 32-битный цвет вне зависимости от глубины цвета окончательного изображения. Например, даже если конечная картинка является 16-битной, все операции все равно проводятся с использованием 32-битного цвета, и лишь в конце палитра сжимается. Считается, что G200 выдавала самую качественную картинку среди всех графических ускорителей того времени. Для высокой производительности в двухмерном режиме использовалась архитектура шины памяти под названием DualBus. Она предусматривала использование двух 64-битных шин и двух командных конвейеров. G200 использовала шину AGP и память типа SGRAM, объем которой составлял 8 или 16 Мбайт. Поддерживались очень высокие разрешения. Например, в режиме 3D G200 мог функционировать с разрешением 1280x1024 точек и 32-битным цветом. Производительность видеокарты находилась на высоком уровне: G200 лишь немного уступала лидерам в лице RIVA TNT и Voodoo2.

Пожалуй, единственной проблемой Matrox G200 являлось ПО. Для запуска многих популярных игр видеокарта использовала специальный OpenGL-to-Direct3D драйвер, который значительно снижал производительность адаптера. Несмотря на обещания программистов Matrox, проблема была решена лишь в следующем поколении видеокарт G400, вышедшем в 2000 году. Однако серия G200 успела пережить небольшой апгрейд. Графический процессор перевели на новый 250-нм техпроцесс, что позволило снизить тепловыделение и энергопотребление. Обновленные видеокарты получили название G200A. Также появилась версия G200A с повышенными частотами, получившая название G250.

1990-е породили большое количество компаний, производящих дискретные видеокарты. Однако с ростом темпа развития 3D-технологий многим из них пришлось предпринять невероятные усилия, чтобы хотя бы просто удержаться в «седле»: кто-то полностью осел в бюджетном сегменте, кто-то попросту сгинул, а кто-то, наоборот, на фоне осечек конкурентов стал сильнее.

# Общее устройство видеокарт

Основные части видеокарты:

Графический процессор (GPU).

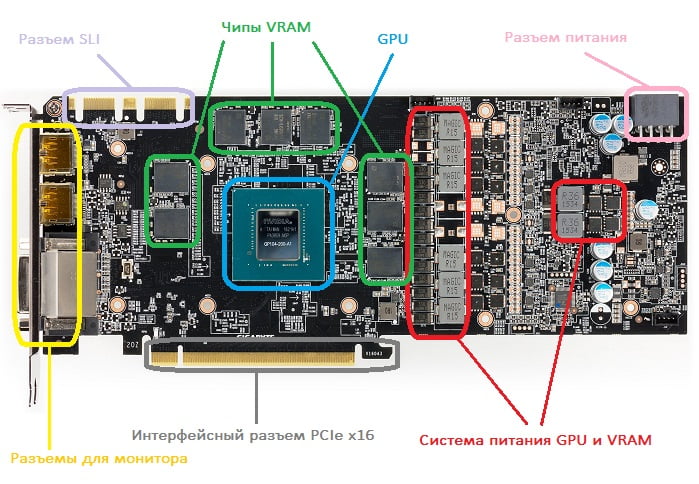
Микросхемы памяти (GDDR).

Система питания GPU и GDDR (ШИМ контроллер(ы) и цепи VRM).

Система охлаждения видеокарты.

Интерфейсный разъем для подключения к материнской плате ([PCIe](https://andiriney.ru/linii-pcie-chto-eto/" \t "_blank)).

Разъем(ы) питания.

Интерфейсные разъемы для подключения монитора/ов.

## Графический процессор (GPU)

Центральное место на печатной плате (ну или примерно центральное, но в любом случае заметное), занято графическим процессором. В окружении других компонентов.

Наверняка напрашивается аналогия с центральным процессором, и она вполне уместна. Тем более, что и название похожее – там CPU (Central Processing Unit), здесь GPU (Graphics Processing Unit). И то, и то – процессоры. Но в чем отличие?

Если смотреть на транзисторы, из которых состоят CPU и GPU, то у последних их больше. Например, на видеокарте установлен чип NVidia GA104, в котором 17.4 млрд транзисторов. А, скажем, AMD Ryzen 7 5800H обходится почти вдвое меньшим количеством.

Некогда большую часть работы по обсчету изображений (например, в играх), которые надо вывести на экран, выполнял CPU. Однако, постоянное усложнение этой работы и необходимость высокой вычислительной мощности привело к тому, что возможностей процессора стало не хватать.

CPU – универсальное устройство, выполняющее множество разных функций. С одной стороны это плюс, а с другой бутылочное горлышко. В данном случае узкоспециализированный GPU позволяет выполнять работу, связанную с подготовкой и выводом изображения на экран монитора, гораздо быстрее.

Связано это в первую очередь с многоядерной и многопоточной обработкой данных. Конечно, CPU тоже умеет выполнять операции параллельно, но приглядимся к ядрам и потокам. Ознакомимся со спецификациями: у AMD Ryzen 7 5800H 8 ядер и 16 потоков. Но что насчет чипа NVidia GA104?

Аналогом процессорным ядрам в графических чипах NVidia являются ядра CUDA, которых в данной модели 6 144. Эти ядра более простые и менее универсальные, но их на порядки больше.

У GPU AMD архитектура чипов несколько иная. Отличия для данной главы минимальны, но подробно об этом в другой главе. Если брать за пример сопоставимый графический процессор Radeon RX 6700 XT, то тоже видим более 17 млрд транзисторов и 2 560 процессорных ядер.

Какие же функции выполняет GPU? В первую очередь – преобразования данных, переданных центральным процессором в изображение, которое отображается на экране монитора. Эта работа выполняется очень быстро, ведь «картинка» постоянно изменяется и в секунду надо подготовить их десятки, а то и сотни, чтобы все изменения на экране происходили плавно, без рывков и замираний.

Поэтому GPU и получились такими сложными. Но было бы неразумно использовать такую мощь только для игр или специфических задач по визуализации. Тысячи простаивающих вычислительных ядер – это слишком расточительно. Поэтому видеокарты активно используются для математических операций в научных программах, инженерных и аналитических расчетах и т. п.

## Видеопамять

К сожалению, пользователи не имеют возможности устанавливать столько памяти, сколько им захочется, как это существует в материнских платах. Видеокарты изначально комплектуются определенным объемом VRAM, который зависит от используемой модели GPU. Изменить это значение почти невозможно.

На плате чипы видеопамяти обычно располагаются рядом с GPU, окружая его. На изображении видны 8 микросхем производства Micron, расположенные с трех сторон вокруг графического процессора. Видеокарты попроще могут обходиться меньшим количеством чипов, помощнее – большим.

Надо еще сказать, что сама видеопамять бывает разной. Сейчас можно найти видеокарты с чипами GDDR5, GDDR6 и GDDR6X. Первый тип уже считается устаревшим и применяется разве что в бюджетных или уходящих с рынка моделях видеокарт. В основном используется память 6-го поколения.

Есть еще память HBM2 (High Bandwidth Memory), но при всех своих преимуществах она не получила широко применения, в немалой степени из-за того, что она дороже, чем та же GDDR6(X).

Если говорить кратко, то в отличие от планарной GDDR6, технология HBM имеет многослойную структуру ячеек. Чем-то напоминает 3D NAND. Из-за этого у нее очень широкая шина доступа – 1024 бит и более. Физически чипы HBM располагаются в непосредственной близости от GPU.

Видеопамять непосредственно используется во время подготовки изображений, и чем больше объем, тем быстрее будет идет расчет, и тем меньше понадобится обращаться к процессору, ОЗУ и накопителю за новой порцией данных. И чем выше разрешение изображения, тем видеопамяти потребуется больше.

На данный момент необходимый минимум - 4 ГБ, который позволит получить приемлемую картинку на экране со средним количеством FPS.

## Система питания

Надежное электропитание – основа стабильной работы любого устройства, и видеокарта тут не исключение. Нагрузка на систему электропитания весьма высока, если учесть потребление электроэнергии современными GPU.

Организовано энергопотребление по тому же принципу, что и [питание процессора](https://andiriney.ru/fazy-pitaniya-protsessora-chto-eto/). Уже привычная связка ШИМ-контроллера и n-го количества фаз на силовых сборках. Контроллеров может быть и несколько.

Фазы питания располагаются либо с одной стороны графического процессора, либо с двух, если этих фаз много. А в топовых видеокартах их действительно много.

У взятой для примера видеокарты их 13 штук, 2 из которых питают видеопамять. В данном случае эти две фазы используют отдельные элементы для верхнего и нижнего плеча, в то время как для GPU применяются 50-амперные сборки.

## Система охлаждения

Именно этот компонент (огромный радиатор, несколько вентиляторов) и бросается в глаза при первом взгляде на видеокарту. Исключение составляют, пожалуй, только бюджетные маломощные графические адаптеры, которые обходятся скромными по размеру и весу системами охлаждения.

Радиатор представляет собой довольно замысловатую железяку с большим количеством тонких ламелей, состоящую из нескольких секций и теплорассеивающих пластин, контактирующих с графическим процессором, чипами памяти и силовыми элементами системы питания.

Практически все производители устанавливают несколько тепловых трубок для повышения эффективности работы радиатора.

Тепловая трубка – металлическая (часто медная, может быть покрыта никелем) герметичная, с легкоиспаряющейся жидкостью внутри. Эта жидкость испаряется в месте соприкосновения одного конца трубки с горячим элементом, после чего пар конденсируется на противоположном, холодном конце, после чего жидкость возвращается в горячий конец трубки.

Чаще всего применяются тепловые трубки с пористыми внутренними стенками, по которым жидкость под воздействием капиллярных сил перемещается в зону испарения вне зависимости от положения самой трубки в пространстве. Этой жидкостью может быть вода, аммиак, этанол и т. п.

Встречается и такой элемент, как испарительная камера. Это некая модернизация идеи тепловой трубки, использующая тот же принцип испарения жидкости, при которой отбирается тепло у охлаждаемого элемента. Разница в том, что плоская многослойная конструкция увеличивает эффективность этого процесса.

У разных производителей количество и конфигурация трубок различается. Часть их может использоваться для контакта с GPU, а для элементов системы питания устанавливается индивидуальная трубка. Может использоваться технология «прямого контакта», при которой тепловые трубки непосредственно контактируют с графическим процессором. Это позволяет улучшить теплообмен.

Несколько лет назад референсные видеокарты NVidia и AMD использовали систему охлаждения с помощью турбины. Суть ее в том, что устанавливается только один радиальный вентилятор, который забирает воздух из корпуса и прогоняет его через всю видеокарту, выбрасывая наружу.

Под кожухом видеокарты установлен радиатор с большим количеством ребер. Главным недостатком такой системы охлаждения является шумность. Единственный вентилятор имеет небольшой размер, а для обеспечения нужного потока воздуха ему приходится вращаться на большой скорости. А чем она выше, тем выше и шум.

Оба производителя отказались от турбинной системы, начиная с семейств видеокарт NVidia RTX 2000 и AMD RX 6000. Предпочтение отдано более привычным осевым вентиляторам, которых может быть от одного до нескольких в зависимости от размера видеокарты и «горячности» ее нрава.

Некоторые производители используют противоположное направление вращения вентиляторов. Если их два, то один вращается по часовой стрелке, другой против. При трех вентиляторах обычно средний имеет противоположное направление вращения по отношению к крайним. Производители утверждают, что таким образом удается бороться с турбулентностью воздушного потока и улучшать обдув радиатора.

Альтернативой штатным системам охлаждения может быть жидкостная, которую можно установить самостоятельно, или производитель изначально предлагает модификацию видеокарты со смонтированным радиатором СЖО. Принципиально она не отличается от той, что используется для CPU, но есть некоторые различия. Проблема в том, что помимо графического чипа необходимо охлаждать еще микросхемы видеопамяти и цепи VRM, что делает теплосъемник довольно сложным конструктивно.

Полезной вещью может оказаться и защитная панель, которая закрывает обратную сторону печатной платы. Она защищает плату от возможных повреждений, улучшает механическую прочность видеокарты, а часто еще и участвует в охлаждении, для чего через термопрокладки контактирует с печатной платой в месте установки, скажем, силовых элементов системы питания.

## Интерфейсный разъем для подключения к материнской плате

Все видеокарты последних поколений используют интерфейс PCI-Express. Последние поколения перешли на 4-ю версию, хотя без проблем работают и в более старых версиях, например, PCIe 3.0.

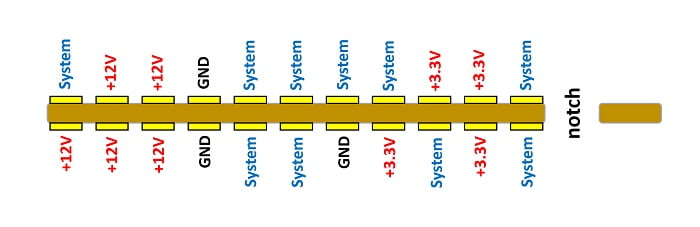
Видеокарты получают от процессора 16 интерфейсных линий. В зависимости от CPU может отличаться количество этих линий и версия интерфейса. Процессоры AMD Ryzen последних поколений и интеловские «камни» могут полноценно обеспечить видеокарту этими «дорожками» для доставки данных.

Если же по какой-то причине нет возможности обеспечить видеокарту всеми 16-ю линиями (например, к процессору подключен второй слот PCIe на материнской плате, и он задействован под какой-либо контроллер) и ей осталась только половина, то ничего не произойдет. Видеокарта будет прекрасно работать, и маловероятно, что вы вообще заметите какие-либо изменения.

## Разъем(ы) питания

Но не только передачей данных занимается разъем PCIe. Через него подается питание на видеокарту. Но есть проблема – он не в состоянии обеспечить нормальное питание даже младшей в линейке моделей видеокарты.

Взглянем на иллюстрацию, показывающую часть разъема PCIe. Здесь 22 контакта, часть занята для служебных нужд, а часть используется для подачи напряжений 12 В и 3.3 В.



Всего у нас есть пять 12-вольтовых контактов и три с напряжением 3.3 В. Согласно спецификациям на PCIe, максимальный ток для напряжения 3.3 В составляет 3 А, и 5.5 А для напряжения 12 В. Высчитать максимальную мощность теперь просто:

(5.5 (А) \* 12 (В)) + (3 (А) \* 3.3 (В)) = 75.9 Вт

Теперь вспомним, что для RTX 3060 максимальное потребление энергии указано равным 170 Вт, что более чем вдвое превышает возможности интерфейса PCIe. Поэтому дополнительную мощность принято брать непосредственно от блока питания.

Для этого на видеокартах есть от одного до трех разъемов (с шестью или восемью контактами) для подключения шлейфа/ов от БП.

# Устройство современных видеокарт

По моему мнению – AMD пытается угнаться за Nvidia. «Зеленый лагерь» (так называют Nvidia) владеет большей частью рынка, так как помимо игровых решений предоставляет профессиональные в виде видеокарт Tesla для решения задач по обучению искусственного интеллекта.

Именно Nvidia владеет большей частью рынка видеокарт, но в чем причина того, что видеокарты от Nvidia популярнее видеокарт от AMD? Попробуем разобраться.

## CUDA

Разве видеокарта со своими вычислительными мощностями необходима только для того, чтобы играть в игры? На самом деле нет. Они уже давно используются и для профессиональных задач.

Но стало это возможно только в 2006 году, когда вышла карточка GeForce 8800, в которой впервые появились ядра CUDA — Compute Unified Device Architecture. Это унифицированные ядра, которые впервые позволяли использовать видеокарту не только для игр, но и для любых массивных параллельных вычислений: типа рендеринга видео, симуляции воды, дыма, ну или майнинга крипты, если вдруг это еще актуально.

А в 2018 году, произошла другая революция — появилась архитектура Turing, а вместе с ней новые типы ядер и, конечно же, технологии трассировки лучей. Поговорим сначала о ней.

## Трассировка лучей

Что такое трассировка лучей? Несмотря на то, что видеокарты за годы своей эволюции обросли поддержкой кучи эффектов и игры действительно стали выглядеть впечатляюще круто, но всё равно остались выглядеть как игры. Почему?

Дело в том, что до появления технологии трассировки лучей, в играх не было настоящего глобального освещения. Оно рассчитывалось для каждого объекта по отдельности, причем поочередно. Соответственно, мы не могли рассчитать, как объекты влияют друг на друга, как преломляется и отражается свет между разными объектами. А всё глобальное освещение сцены просто заранее «запекалось» в текстуру в графическом редакторе игры. Есть еще зонды и куча других техник, которые позволяют получить грубую имитацию глобального освещения с кучей недостатков — протеканием света сквозь объекты, повышенными требованиями к объему видеопамяти, отсутствием физики, так как если изменится положение объектов в сцене, то и освещение придется заново считать, а с использованием заранее подготовленных ресурсов — это невозможно.

А вот трассировка лучей впервые позволила, построить освещение по законам природы и сняла кучу ограничений. Ну, практически. Как это работает?

Вместо того, чтобы поочередно считать освещение для каждого объекта сначала выводится вся трехмерная сцена и упаковывается в BVH коробки для ускорения трассировки. После чего из камеры в упакованную 3D-сцену запускается луч, и мы смотрим с какой поверхностью он пересечется. А дальше от этой точки строится по одному лучу до каждого источника освещения. Так мы понимаем где свет, а где тень.

А если луч попал на отражающий объект, то строится еще один отраженный луч и так мы можем считать переотражения. Чем больше переотражений мы считаем, тем сложнее просчет, но реалистичнее результат.

Всё практически как в жизни, но для экономии ресурсов, лучи запускаются не от источника света, а из камеры. Иначе бы пришлось просчитывать много лучей, которые не попадают в поле зрения игрока, то есть делать бесполезные, отнимающие ресурсы GPU вычисления.

## Новые ядра

Для реализации трассировки лучей, помимо ядер CUDA пришлось придумать ядра нового типа. Это RT-ядра, что, собственно, и значит ядра трассировки лучей и тензорные ядра.

RT-ядра выполняют тот самый поиск пересечений между лучом и полигонами сцены. И делают это очень эффективно благодаря алгоритму BVH — Bounding Volume Hierarchy.

Суть алгоритма: Каждый полигон вкладывается в несколько коробок разного размера, как в матрешку. И вместо того, чтобы проверять пересечения с каждым полигоном сцены, коих миллионы, сначала проверяется попал ли луч в небольшое количесчтво коробок, в которые упакованы треугольники сцены. На последнем уровне BVH матрешки содержится коробка с несколькими треугольниками сцены. Коробок намного меньше, чем треугольников, поэтому на тестирование сцены уходит намного меньше времени, чем если бы мы перебирали каждый треугольник сцены.

Тензорные ядра — это вообще необычная вещь для видеокарты. Такие ядра используется для операций матричного перемножения. То есть могут умножать много чисел одновременно. Это очень полезно для обучения глубоких нейронных сетей. Поэтому как правило, нейросети сейчас обучают и используют именно на видеокартах.

Но и в играх тензорные ядра имеют высокий вес. Во-первых, очищение рейтрейснутой картинки от шума в профессиональных пакетах с поддержкой OptiX.

Но в первую очередь, для реализации фирменной технологии DLSS — Deep Learning Super Sampling. Это технология сглаживания и апскейлинга картинки при помощи нейросетей. Например, у вас 4K-монитор, но видеокарта не тянет 4К на ультрах. Что в этом случае делает DLSS. Картинка рендерится в более низком разрешении 1440P или 1080P, а потом несколько соседних кадров объединяются нейросетью в новый кадр более высокого разрешения. Да так, что часто выглядит даже лучше, чем в нативном разрешении. При этом мы получаем огромный прирост фреймрейта.

# Общий принцип работы видеокарты

## Этап 1. Растеризация.

Трёхмерный объект изначально векторный. Он состоит из треугольников, которые можно описать вершинами. То есть, по сути, объект — это набор вершин, со своими координатами в трехмерном пространстве.

Но экран двумерный, да ещё и состоит из пикселей. Поэтому для отображения векторный 3D объекта в двухмерном пространстве необходимо его спроецировать.

Мы переносим координаты вершин на плоскую поверхность соединяем их при помощи уравнений прямых на плоскости и заполняем пикселями плоскости треугольников. На этом этапе мы получаем двухмерную проекцию объекта на экране.

## Этап 2. Текстурирование

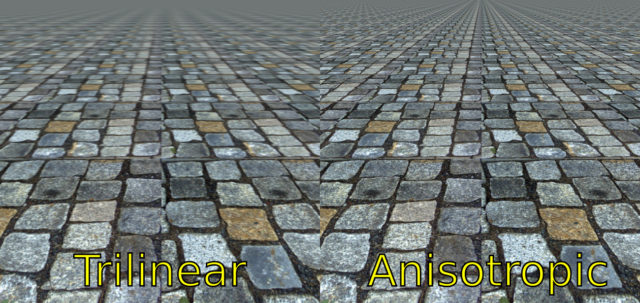
Дальше необходимо как-то раскрасить модельку. Поэтому на растрированный объект по текстурным координатам натягивается текстура.

Но просто натянуть текстуру недостаточно, ее нужно как-то сгладить. Иначе при приближении к объекту вы просто будите видеть сетку текселей. Прям как в первых 3D-играх типа DOOM. Поэтому дальше к текстуре применяются различные алгоритмы фильтрации.

На первых видеокартах применялась билинейная интерполяция. Всё, что она делает — это линейно интерполирует промежуточные цвета между четырьмя текселями, а полученное значение становится цветом пикселя на экране.

Такую интерполяцию активно используют и сейчас. Но в дополнении к ней также делают трилинейную интерполяцию — это еще более продвинутая версия сглаживания, но используют её только на стыках разных уровней детализации текстур, чтобы их замаскировать.

А вот чтобы вернуть четкости текстурам под углами к камере используют анизотропную интерполяцию. Чем выше её коэффициент, тем четче получается картинка, так как для получения цвета каждого пикселя делается до 16 выборок. Особенно это заметно на поверхностях, находящихся под острым углом к камере. То есть, к примеру, на полу.



Окей, теперь мы получили цветное изображение. Но этого всё ещё недостаточно, потому как в сцене нет освещения. Поэтому переходим к следующему этапу с названием пиксельный шейдер или затенение пикселей.

## Этап 3. Пиксельный шейдер.

Вообще шейдер — это штука, которая позволяет программисту что-то делать с вершинами, треугольниками и пикселями на программном уровне. В случае Пиксельных Шейдеров — даёт разработчикам разработчиком возможность динамически менять цвет каждого пикселя экрана по программе.

Кстати, впервые поддержка простых шейдеров появились в 2001 году, когда появилась NVIDIA GeForce 3. До этого освещение тоже делалось, но аппаратными средствами и разработчики особо не могли влиять на результат. Так вот сегодня это самый ресурсоемкий этап, на котором для каждого пикселя нужно просчитать как он отражает, рассеивает и пропускает свет. Как ложатся тени по поверхности модели и прочее. То есть иными словами рассчитывается финальный цвет пикселя.

Каждый объект сцены описывается при помощи нескольких текстур:

* Карта нормалей, текстура, в которой хранятся векторы нормалей для каждой точки поверхности. При помощи этих векторов рассчитывается попиксельное освещение.
* Карта зеркальности, которая описывает сколько света отражается от поверхности.
* Карта шероховатостей (roughness mapbump map), которая описывает микрорельеф поверхности или то, как поверхность будет рассеивать свет.
* Альбедо карта, то есть карта диффузии или естественный цвет объекта.
* И прочие.

## Этап 4. Сохранение

После кучи вычислений при помощи информации из всех вышеперечисленных текстур наступает последний этап. Мы получили финальный цвет пикселя и сохраняем его в видеопамять. А после обработки всей сцены мы уже можем выводить картинку на экран.

# AMD vs Nvidia – сравнение архитектур

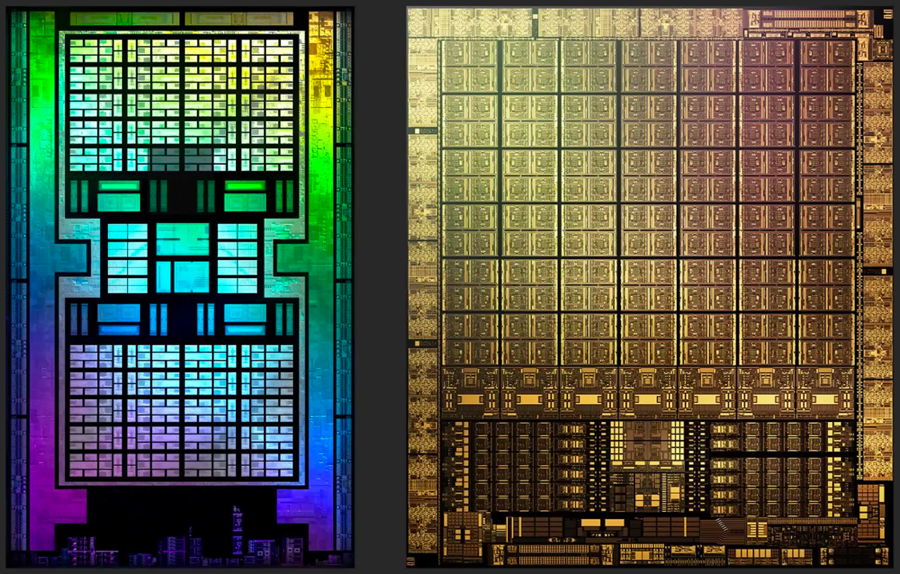
Спустя два года после запуска Turing в сентябре 2020 года NVIDIA сменила архитектуру своих видеокарт на [Ampere](https://www.techspot.com/review/2099-geforce-rtx-3080/). AMD не осталась в стороне и вскоре после этого тоже [обновила архитектуру](https://www.techspot.com/review/2144-amd-radeon-6800-xt/) RDNA до второй версии. Так какая из них лучше?

## Внутри кристаллов

При проектировании крупномасштабного процессора обычно требуется, чтобы общие ресурсы, такие как контроллеры и кэш, находились в центре, чтобы гарантировать, что все компоненты имеют одинаковый путь к ним.

Интерфейсные системы вроде контроллеров локальной памяти или видеовыходов должны располагаться по краям микросхемы, чтобы упростить их подключение к тысячам отдельных проводов, соединяющих графический процессор с остальной частью карты.

Ниже приведены изображения кристаллов AMD Navi 21 и NVIDIA GA102 в искусственных цветах. Оба изображения были подчищены и показывают только один слой внутри чипа, однако при этом дают хорошее представление о внутренностях современного графического процессора



Наиболее очевидное различие между конструкциями заключается в том, что NVIDIA не следует централизованному подходу к компоновке микросхем: все системные контроллеры и основной кэш находятся внизу, а логические блоки расположены в длинных столбцах. Они проделывали это и раньше, но только с моделями среднего и нижнего ценового сегмента.

Например, Pascal GP106 (используемый в [GeForce GTX 1060](https://www.techspot.com/review/1209-nvidia-geforce-gtx-1060/)) был буквально вдвое меньше GP104 (из GeForce [GTX 1070](https://www.techspot.com/review/1182-nvidia-geforce-gtx-1070/)). В более ранней версии размер кристалла был больше, а кэш-память и контроллер располагались посередине. У младшего брата они переместились в сторону.

Для всех предыдущих топовых ГП NVIDIA использовала классическую централизованную компоновку. Зачем же было менять подход? Интерфейсы здесь ни при чем, ведь контроллеры памяти и PCI Express работают на краю кристалла. С тепловыми проблемами это тоже не связано, ведь даже если кэш-часть или контроллер кристалла будут нагреваться сильнее, чем логические секции, вам наверняка захочется, чтобы посередине схемы было больше теплопоглощающего кремния. Хотя причина этого изменения не вполне понятна, есть подозрение, что она связана с реализацией блоков вывода рендеринга (ROP).

Позже мы рассмотрим их более подробно, а пока просто скажем, что, хотя изменение макета выглядит странно, оно не оказывает существенного влияния на производительность. Это связано с тем, что 3D-рендеринг сопровождается большим количеством длительных задержек — как правило, из-за необходимости ожидания данных. Таким образом, дополнительные наносекунды, добавленные за счет того, что некоторые логические блоки находятся дальше от кэша, скрываются в общей схеме чипа.

Прежде чем мы продолжим, стоит отметить инженерные изменения, реализованные AMD в компоновке Navi 21 по сравнению с Navi 10, установленном в [Radeon RX 5700 XT](https://www.techspot.com/review/2010-geforce-rtx-2060-super-vs-radeon-5700-xt/). Несмотря на то, что новый чип в два раза больше предыдущего как по площади, так и по количеству транзисторов, разработчикам удалось улучшить тактовые частоты без значительного увеличения энергопотребления. Например, [Radeon RX 6800 XT](https://www.techspot.com/review/2144-amd-radeon-6800-xt/) имеет базовую частоту и частоту разгона 1825 и 2250 МГц, соответственно, при TDP, равном 300 Вт. Те же показатели для Radeon RX 5700 XT: 1605 МГц, 1905 МГц и 225 Вт.

## Управление ГП - Как все устроено внутри чипов

Драйверы, которые AMD и NVIDIA создают для своих чипов, по сути, работают как трансляторы: они преобразуют процедуры, выданные через API, в последовательность операций, понятную графическим процессорам. Затем все зависит от аппаратного обеспечения: какие инструкции выполняются в первую очередь, какая часть микросхемы их выполняет и так далее.

Этот начальный этап управления инструкциями обрабатывается набором модулей в микросхеме. В RDNA 2 графические и вычислительные шейдеры маршрутизируются через отдельные конвейеры, которые планируют и отправляют инструкции остальной части микросхемы: первый называется *Graphics Command Processor*, второй — асинхронными вычислительными блоками (ACE).

Графические процессоры достигают высокой производительности за счет параллельного выполнения задач, поэтому следующий уровень организации дублируется в чипе. Если проводить аналогию с реальным заводом, это будет похоже на компанию с центральным офисом в одном месте, но производством в нескольких других местах.

AMD называет это Shader Engine (SE), тогда как в NVIDIA они имеют название графических кластеров (GPC): названия разные, но суть одна.

Причина такого разделения проста: блоки обработки графики просто не могут справиться со всем и сразу. Так что имеет смысл продвинуть некоторые обязанности по планированию и организации дальше. Это также означает, что каждый блок может делать что-то независимо от других: скажем, один — обрабатывать графические, другой — вычислительные шейдеры.

В случае RDNA 2 каждый SE содержит собственный набор фиксированных функциональных блоков — схем, предназначенных для выполнения одной конкретной задачи:

Блок Primitive Setup — подготавливает вершины к обработке, а также генерирует больше вершин (тесселяция) и отбраковывает их;

Растеризатор — преобразует трехмерный мир треугольников в двухмерную сетку пикселей;

Блоки вывода рендеринга (ROP) — считывают, записывают и смешивают пиксели.

Блок Primitive Setup работает с частотой 1 треугольник за такт. Параметр может показаться не очень большим, но следует не забывать, что эти чипы работают на частотах между 1,8 и 2,2 ГГц, и эта настройка не должна оказываться узким местом ГП. Для Ampere этот блок находится на следующем уровне организации.

Ни AMD, ни NVIDIA не особенно распространяются о своих растеризаторах. NVIDIA называют их Raster Engines. Мы знаем, что они обрабатывают 1 треугольник за такт, но больше никакой информации о них нет.

Каждый SE в чипе Navi 21 содержит 128 ROP; GA102 от NVIDIA включает в себя 112 ROP. Может показаться, что у AMD здесь есть преимущество, ведь большее количество ROP означает, что за такт может обрабатываться больше пикселей. Но такие устройства нуждаются в хорошем доступе к кэш-памяти и локальной памяти, и мы поговорим об этом позже в этой статье. А пока давайте дальше рассмотрим на разделение SE/GPC.

Shader Engines AMD разделены на то, что они сами называют двойными вычислительными блоками (DCU), при этом чип Navi 21 использует десять DCU для каждого SE. В случае Ampere и GA102 они называются кластерами обработки текстур (TPC), причем каждый графический процессор содержит 6 TPC. Они также работают со скоростью 1 треугольник за такт, и хотя графические процессоры NVIDIA работают на меньшей частоте, чем AMD, и у них намного больше TPC, чем у Navi 21 — SE. Таким образом, при той же тактовой частоте GA102 имеет здесь заметное преимущество, поскольку весь чип содержит 42 блока Primitive Setup, тогда как новый RDNA 2 от AMD — только 4. Но поскольку на один Raster Engine приходится шесть TPC, GA102 фактически имеет 7 систем примитивов, в то время как Navi 21 — четыре. Кажется, что NVIDIA имеет здесь явное лидерство.

Последний уровень организации чипов — вычислительные блоки (CU) в RDNA 2 и потоковые мультипроцессоры (SM) в Ampere — производственные линии в ГП-«заводах».

В значительной степени они составляют основную начинку ГП, поскольку содержат все программируемые блоки, используемые для обработки графики, вычислений, а теперь еще и трассировки лучей. Каждый из них занимает очень небольшую часть общего пространства кристалла, но они по-прежнему чрезвычайно сложны и очень важны для общей производительности чипа.

 Итак, номенклатура у чипов разная, но функции во многом схожи. И поскольку многое из того, что они делают, ограничивается программируемостью и гибкостью, любые преимущества одного из них по сравнению с другим сводятся к простому ощущению масштаба.

 Но в случае с CU и SM - AMD и NVIDIA используют разные подходы к обработке шейдеров. И пусть в некоторых областях у них много общего, но много и различий.

## Подсчет ядер по методу NVIDIA

Если Turing имела множество существенных отличий от Pascal, то Ampere кажется довольно мягким обновлением предыдущей архитектуры — по крайней мере, на первый взгляд. Точно известно, что по сравнению с Turing новая архитектура имеет более чем в два раза большее количество ядер CUDA в каждом SM.

В Turing потоковые мультипроцессоры содержат четыре раздела (иногда называемых блоками обработки), каждый из которых содержит логические блоки 16x INT32 и 16x FP32. Эти схемы предназначены для выполнения очень специфических математических операций с 32-битными значениями данных: блоки INT обрабатывают целые числа, а FP — числа с плавающей запятой.

NVIDIA заявляет, что SM Ampere имеет в общей сложности 128 ядер CUDA, но, строго говоря, это неправда — или с таким же успехом можно считать, что у Turing их было столько же. Блоки INT32 действительно могли обрабатывать значения с плавающей запятой, но только в очень небольшом количестве простых операций. Для Ampere NVIDIA увеличила поддерживаемый диапазон математических операций с плавающей запятой, чтобы соответствовать другим модулям FP32. Это означает, что общее количество ядер CUDA на SM действительно не изменилось, просто половина из них теперь имеет больше возможностей.

Поскольку блоки INT/FP могут работать независимо, SM Ampere может обрабатывать до 128 вычислений FP32 за цикл или 64 операций FP32 и 64 операций INT32 одновременно. Turing же умела делать только последнее.

Таким образом, новый графический процессор может потенциально вдвое увеличить производительность FP32 по сравнению с предшественником. Для вычислительных рабочих нагрузок это большой шаг вперед, но для игр польза окажется гораздо меньшей. Это стало очевидно после тестирования GeForce RTX 3080, в которой используется чип GA102 с 68 включенными SM.

Несмотря на то, что пиковая пропускная способность FP32 составляет 121% по сравнению с GeForce 2080 Ti, в среднем она увеличивает частоту кадров только на 31%. Так почему же вся эта вычислительная мощность тратится зря?

Простой ответ: зря она не тратится, просто игры не всегда запускают инструкции FP32. Когда NVIDIA выпустила Turing в 2018 году, компания отметила, что в среднем около 36% инструкций, обрабатываемых графическим процессором, связаны с процедурами INT32. Эти вычисления обычно выполняются для определения адресов памяти, сравнения двух значений и логического управления.

Таким образом, для этих операций функция двойной скорости FP32 не используется, поскольку блоки с двумя путями данных могут работать только с целыми числами или с плавающей запятой. SM переключится в этот режим только в том случае, если все выстроенные в очередь 32 потока, обрабатываемые им в данный момент, выполняют одну и ту же операцию FP32. Во всех остальных случаях SM в Ampere работают так же, как и в Turing.

Это означает, что [GeForce RTX 3080](https://www.techspot.com/review/2099-geforce-rtx-3080/) имеет только 11-процентное преимущество FP32 над 2080 Ti при работе в режиме INT+FP. Вот почему реальный прирост производительности в играх не так высок, как предполагают исходные данные.

Какие тут еще улучшения? На каждый SM приходится меньше тензорных ядер, но каждое из них оказывается намного более мощным, чем в Turing. Эти схемы выполняют очень специфические вычисления (например, умножают два значения FP16 и складывают ответ с другим FP16), и теперь каждое ядро ​​выполняет 32 таких операции за цикл.

Также есть поддержка новой функции [Fine-Grained Structured Sparsity](https://blogs.nvidia.com/blog/2020/05/14/sparsity-ai-inference/). Если кратко, то с ее помощью математическая скорость может быть удвоена путем удаления данных, которые не влияют на ответ. Опять же, это хорошая новость для профессионалов, работающих с нейронными сетями и искусственным интеллектом, но на данный момент в этом нет никаких значительных преимуществ для игровых разработчиков.

Ядра трассировки лучей также претерпели доработки: теперь они могут работать независимо от ядер CUDA, поэтому, пока они выполняют обход BVH или математику пересечения примитивов лучей, остальная часть SM все еще может обрабатывать шейдеры. Часть ядер трассировки лучей, отвечающая за проверку пересечений, также имеет вдвое большую производительность.

Ядра трассировки лучей также оснащены дополнительным оборудованием, которое помогает применять трассировку лучей к размытию движения, но эта функция в настоящее время доступна только через собственный [Optix API](https://developer.nvidia.com/optix) от NVIDIA.

Есть и другие хитрости, но в целом подход основан на разумной, но неуклонной эволюции, а не на совершенно новом дизайне. Но учитывая, что в исходных возможностях Turing с самого начала не было ничего особенно плохого, это и неудивительно.

Что же насчет AMD — что они сделали с вычислительными модулями в RDNA 2?

## Трассировка лучей по-особенному

На первый взгляд, AMD не сильно изменила вычислительные блоки: они по-прежнему содержат два набора векторных блоков SIMD32, скалярный блок SISD, блоки наложения текстур и стек различных кэшей. Произошли некоторые изменения в отношении того, какие типы данных и связанные с ними математические операции они могут выполнять. Но наиболее заметным изменением для обычного потребителя является то, что AMD теперь предлагает аппаратное ускорение для определенных процедур трассировки лучей.

Эта часть вычислительных блоков выполняет проверки пересечения лучевого бокса или лучевого треугольника — то же самое, что и ядра трассировки лучей в Ampere. Однако последние также ускоряют алгоритмы обхода BVH, тогда как в RDNA 2 это делается с помощью вычислительных шейдеров с использованием модулей SIMD 32.

Независимо от того, сколько шейдерных ядер или насколько высоки их тактовые частоты, использование специализированных схем, предназначенных для выполнения только одной задачи, всегда будет предпочтительнее, чем обобщенный подход. Именно поэтому и были изобретены графические процессоры: все в мире рендеринга можно сделать с помощью ЦП, но общий характер делает их непригодными для графических применений.

 Блоки Ray Accelerator находятся рядом с текстурными процессорами, поскольку они фактически являются частью одной и той же структуры. Хотя эта система действительно предлагает большую гибкость и устраняет необходимость в том, чтобы части кристалла занимались только трассировкой лучей и ничем другим одновременно с ней, первая реализация ее имеет некоторые недостатки. Наиболее примечательным из них является то, что текстурные процессоры могут обрабатывать только операции, связанные с текстурами или пересечениями примитивов лучей. Учитывая, что ядра трассировки лучей NVIDIA теперь работают полностью независимо от остальной части SM, это дает Ampere явное преимущество по сравнению с RDNA 2 в проработке структур ускорения и тестах пересечений, необходимых для трассировки лучей.

 Пока производительность трассировки лучей в новейших видеокартах AMD [была исследована лишь вкратце](https://www.techspot.com/review/2146-amd-radeon-6800/), но уже заметно, что влияние трассировки лучей очень зависит от игры.

 Например, в Gears 5 видеокарта Radeon RX 6800 (в которой используется вариант графического процессора Navi 21 с 60 CU) при включении трассировки лучей снизила частоту кадров только на 17%, тогда как в Shadow of the Tomb Raider — аж на 52%. Для сравнения, у NVIDIA RTX 3080 (с использованием 68 SM GA102) средняя потеря частоты кадров в этих двух играх составила 23% и 40% соответственно.

 Чтобы больше рассказать о реализации AMD, необходим более подробный анализ трассировки лучей, однако в качестве первой итерации технологии она кажется вполне конкурентоспособной, но чувствительной к тому, какое приложение выполняет трассировку.

 Как упоминалось ранее, вычислительные блоки в RDNA 2 теперь поддерживают больше типов данных: наиболее заметными из них являются типы данных с низкой точностью, такие как INT4 и INT8. Они используются для тензорных операций в алгоритмах машинного обучения, и хотя AMD имеет отдельную архитектуру (CDNA) для ИИ и центров обработки данных, это обновление предназначено для использования с [DirectML](https://www.techspot.com/article/2137-next-gen-directx-12/).

Этот API является недавним дополнением к семейству Microsoft [DirectX 12](https://www.techspot.com/article/2137-next-gen-directx-12/). Комбинация аппаратного и программного обеспечения обеспечивает лучшее ускорение шумоподавления в алгоритмах трассировки лучей и временного масштабирования. В случае с последним у NVIDIA, конечно же, есть своя технология под названием DLSS. Она использует тензорные ядра в SM для выполнения части вычислений — но учитывая, что аналогичный процесс может быть построен и через DirectML, может показаться, что эти компоненты в некоторой степени избыточны. Однако и в Turing, и в Ampere тензорные ядра также обрабатывают все математические операции, связанные с форматами данных FP16.

В RDNA 2 такие вычисления выполняются с использованием шейдерных блоков и упакованных форматов — то есть, каждый 32-битный векторный регистр содержит два 16-битных.

Какой же подход лучше?

AMD преподносит свои блоки SIMD32 как векторные процессоры, поскольку они выдают одну инструкцию для нескольких значений данных. Один векторный блок содержит 32 потоковых процессора, и поскольку каждый из них работает только с одним фрагментом данных, по факту операции носят скалярный характер. По сути, это то же самое, что и SM в Ampere, где каждый блок обработки также применяет одну инструкцию для 32 значений данных.

Но если у NVIDIA весь SM может обрабатывать до 128 вычислений FMA FP32 за цикл, один вычислительный блок RDNA 2 производит только 64 таких вычисления. Использование FP16 увеличивает это значение до 128 FMA за цикл, что совпадает с тем, что делают тензорные ядра в Ampere при стандартных вычислениях FP16.

SM NVIDIA могут выполнять инструкции для одновременной обработки целочисленных значений и значений с плавающей запятой (например, 64 FP32 и 64 INT32) и имеют независимые блоки для операций FP16, тензорной математики и процедур трассировки лучей. Блоки управления AMD выполняют большую часть рабочей нагрузки блоков SIMD32, хотя у них есть отдельные скалярные блоки, которые поддерживают простую целочисленную математику.

Таким образом, может показаться, что у Ampere здесь преимущество: у GA102 больше SM, чем у Navi 21, и у них больше возможностей, когда дело доходит до пиковой пропускной способности, гибкости и предлагаемых функций. Но у AMD есть свой туз в рукаве.

## Система памяти и многоуровневые кэши

Давайте сначала взглянем на Ampere. В целом, внутри произошли некоторые заметные изменения: объем кэша 2-го уровня увеличился на 50% (Turing TU102 имел 4096 КБ, соответственно), а кэши 1-го уровня в каждом SM увеличились вдвое.

Как и раньше, кэш-память L1 здесь настраивается с точки зрения того, сколько места в кэше можно выделить для данных, текстур или общих вычислений. Для графических шейдеров (например, вершинных или пиксельных) и асинхронных вычислений кэш фактически установлен на:

64 КБ для данных и текстур;

48 КБ для общей памяти;

16 КБ для конкретных операций.

Остальная часть внутренней памяти осталась прежней, но за пределами графического процессора ждет приятный сюрприз. NVIDIA стала работать с Micron и теперь использует модифицированную версию GDDR6 для своих потребностей в локальной памяти. По сути, это тот же GDDR6, но шина данных полностью заменена. Вместо того, чтобы использовать обычную настройку 1 бит на вывод, при которой сигнал очень быстро колеблется между двумя значениями напряжения ([PAM](https://en.wikipedia.org/wiki/Pulse-amplitude_modulation)), GDDR6X использует четыре значения напряжения:

Благодаря этому GDDR6X эффективно передает 2 бита данных на вывод за цикл, поэтому при той же тактовой частоте и количестве выводов полоса пропускания удваивается. GeForce RTX 3090 поддерживает 24 модуля GDDR6X, работающих в одноканальном режиме и рассчитанных на 19 Гбит/с, что дает пиковую пропускную способность 936 ГБ/с. Это на 52% больше, чем у GeForce RTX 2080 Ti. Таких показателей пропускной способности в прошлом можно было достигнуть только при помощи HBM2, реализация которого может быть куда более дорогостоящей, чем GDDR6.

Однако такую память производит только Micron, а использование PAM4 добавляет дополнительной сложности производственному процессу, требуя гораздо более жестких допусков при передаче сигналов. AMD пошла по другому пути: вместо того, чтобы обращаться за помощью к стороннему поставщику, они использовали свое подразделение ЦП, чтобы изобрести что-то новое. Общая система памяти в RDNA 2 не сильно преобразилась по сравнению с предшественницей — но есть два существенных изменения.

Каждый шейдерный движок теперь имеет два набора кэшей первого уровня. Но как можно втиснуть в графический процессор 128 МБ кэш-памяти третьего уровня? Используя конструкцию SRAM для кэша L3, AMD встроила в чип два набора кэш-памяти высокой плотности объемом 64 МБ. Транзакции данных обрабатываются 16 наборами интерфейсов, каждый из них сдвигает 64 байта за такт.

Так называемый Infinity Cache имеет свой собственный тактовый домен и может работать на частоте 1,94 ГГц, что дает пиковую внутреннюю пропускную способность 1986,6 ГБ/с. И поскольку это не внешняя DRAM, задержки здесь исключительно низкие. Такой кэш идеально подходит для хранения структур ускорения трассировки лучей, и поскольку обход BVH включает в себя множество проверок данных, Infinity Cache должен в этом особенно помочь.

На данный момент не ясно, работает ли кеш третьего уровня в RDNA 2 так же, как в ЦП Zen 2: то есть, как кэш жертвы (victim cache) второго уровня. Обычно, когда необходимо очистить последний уровень кэша, чтобы освободить место для новых данных, любые новые запросы этой информации должны поступать в DRAM.

В кэше жертвы хранятся данные, помеченные для удаления из следующего уровня памяти, и имея под рукой 128 МБ, Infinity Cache потенциально может хранить 32 полных набора кэша L2. Эта система снижает нагрузку на контроллеры GDDR6 и DRAM.

Старые конструкции графических процессоров AMD боролись с нехваткой внутренней пропускной способности — особенно после увеличения тактовой частоты, — но дополнительный кэш во многом поможет при решении этой проблемы.

## Так что же лучше?

Использование GDDR6X дает GA102 огромную полосу пропускания для локальной памяти, а большие кэши помогают уменьшить влияние промахов кэша. Массивная кэш-память 3-го уровня Navi 21 позволяет реже использовать DRAM, при этом графический процессор может работать на более высоких тактовых частотах без дефицита данных.

Решение AMD придерживаться GDDR6 означает, что сторонним поставщикам доступно больше источников памяти, в то время как любая компания, производящая GeForce RTX 3080 или 3090, будет вынуждена использовать Micron. И хотя GDDR6 поставляется с модулями различной плотности, GDDR6X в настоящее время ограничен 8 Гб.

Система кэширования в RDNA 2, возможно, является лучшим подходом, чем та, что используется в Ampere, поскольку использование нескольких уровней встроенной SRAM всегда обеспечивает более низкие задержки и лучшую производительность для заданного диапазона мощности, чем внешняя DRAM, независимо от пропускной способности последней.

## Мультимедиа-движок и видеовыход

Обе архитектуры обеспечивают вывод изображения через HDMI 2.1 и DisplayPort 1.4a. Первый предлагает более широкую полосу пропускания сигнала, но оба они рассчитаны на 4K при 240 Гц с HDR и 8K при 60 Гц. Это достигается при помощи либо цветовой субдискретизации 4:2:0, либо DSC 1.2a. Это алгоритмы сжатия видеосигнала, обеспечивающие значительное снижение требований к полосе пропускания без большой потери визуального качества. Без них даже пиковой пропускной способности HDMI 2.1 в 6 ГБ/с было бы недостаточно для передачи изображений 4K с частотой 60 Гц.

Ampere и RDNA 2 также поддерживают системы с [переменной частотой обновления](https://en.wikipedia.org/wiki/Variable_refresh_rate) (FreeSync для AMD, G-Sync для NVIDIA), в кодировании и декодировании видеосигналов между ними также нет заметной разницы.

Независимо от того, какой процессор рассматривать, в обоих есть поддержка декодирования 8K AV1, 4K H.264 и 8K H.265, хотя и еще не было тщательно изучено, насколько хорошо они оба работают в таких ситуациях. Ни одна из компаний не раскрывает подробностей о внутреннем устройстве в этих областях. Какими бы важными они ни были в наши дни, все внимание по-прежнему привлекают другие аспекты графических процессоров.

## Созданы для вычислений, созданы для игр

Раньше AMD и NVIDIA использовали разные подходы к выбору архитектуры и конфигурации. Но по мере того, как 3D-графика набирала все большую популярность, они становились все более похожими.

На данный момент у NVIDIA есть три чипа, использующих технологию Ampere: GA100, GA102 и GA104.

Последний — урезанная версия GA102. GA100 — совсем другое дело.

В нем нет ядер трассировки лучей и CUDA с поддержкой INT32+FP32 — вместо этого он содержит множество дополнительных модулей FP64, еще больше load/store систем и огромный объем кэш-памяти L1/L2. Все это объясняется тем, что он разработан для вычислений ИИ и анализа данных.

GA102/104, в свою очередь, должны охватывать все остальные рынки, на которые нацелена NVIDIA: геймеров, профессиональных графических художников и инженеров, а также маломасштабный ИИ и вычислительные системы. Ampere должен быть мастером на все руки — а это задача не из легких.

RDNA 2 была разработана только для игр на ПК и консолях, хотя могла с таким же успехом работать в тех же областях, что и Ampere. Однако AMD решила сохранить свою архитектуру GCN и обновить ее в соответствии с требованиями сегодняшних клиентов.

Там, где RDNA 2 породила Big Navi, можно сказать, что CDNA породила Big Vega: в Instinct MI100 находится их чип Arcturus — 50-миллиардный транзисторный графический процессор с 128 вычислительными блоками.

Хотя NVIDIA в значительной степени доминирует на профессиональном рынке с моделями Quadro и Tesla, Navi 21 просто не нацелена на конкуренцию с ними.

Скоро AMD выпустит Radeon RX 6900 XT, в котором используется полная версия Navi 21 (без отключенных CU), которая сможет работать так же хорошо, как [GeForce RTX 3090](https://www.techspot.com/review/2105-geforce-rtx-3090/) или даже лучше. Но GA102 на этой карте также не полностью включен, поэтому у NVIDIA всегда есть возможность обновить эту модель до «супер»-версии, как они это сделали с Turing в прошлом году.

Можно утверждать, что, поскольку RDNA 2 используется в [Xbox Series X](https://www.techspot.com/tag/xbox+series+x/)[/S](https://www.techspot.com/tag/xbox/) и [PlayStation 5](https://www.techspot.com/news/87693-first-two-weeks-playstation-5.html), разработчики игр будут отдавать предпочтение этой архитектуре для своих игровых движков. Но стоит просто вспомнить времена, когда GCN использовался в Xbox One и PlayStation 4, чтобы получить представление, как это, вероятно, будет происходить.

Первая версия, выпущенная в 2013 году, использовала графический процессор, построенный на архитектуре GCN 1.0, не появлявшийся в видеокартах для настольных ПК вплоть до следующего года. Xbox One X, выпущенный в 2017 году, использовал GCN 2.0, которому к тому времени было уже более 3 лет.

Оба графических процессора обладают исключительными возможностями и представляют собой чудо того, что может быть достигнуто в производстве полупроводников. NVIDIA и AMD предлагают разные инструменты, поскольку пытаются решать разные проблемы: Ampere стремится быть всем для всех, RDNA 2 — только для игр.

# Выводы

Начиная с 1980 года видеокарты прошли огромный путь эволюции – от простейшего устройства вывода символов на экран до кристаллов с миллиардами транзисторов и множеством прорывных технологий. Эти технологии обновляются, появляются и исчезают (как, например, SLI) очень быстро, и в наше время сложно следить за всем что происходит на этом рынке.

Чипы становятся сложнее и сложнее с каждым годом, и их создание сегодня можно назвать настоящим искусством. Сейчас над проектированием одного чипа работают сотни высококвалифицированных сотрудников, тратится множество времени и денег. И все для того, чтобы рядовой пользователь или кампания могла использовать видеокарты для своих потребностей и нужд.

# Список литературы

<https://vgamuseum.ru/>

виртуальный музей видеокарт

<https://www.accross.su/blog/view/108>

краткая история развития видеокарт

<https://habr.com/ru/post/522490/> и <https://habr.com/ru/post/565796/>

история видеокарт с 1950 по 2020 год

<https://games.mail.ru/pc/articles/hard/istorija_videokart_chast_1/>

история видеокарт

<https://www.nvidia.com/ru-ru/geforce/graphics-cards/30-series/>

GeForce видеокарты 30-й серии

<https://www.techspot.com/review/2099-geforce-rtx-3080/>

обзор видеокарт серии GeForce-30

<https://www.techspot.com/review/2144-amd-radeon-6800-xt/>

обзор видеокарты AMD RX 6800 XT

<https://www.techspot.com/article/1874-amd-navi-vs-nvidia-turing-architecture/>

Сравнение архитектур Turing и Navi

<https://www.techspot.com/review/2010-geforce-rtx-2060-super-vs-radeon-5700-xt/>

сравнение видеокарт от amd и nvidia

<https://blogs.nvidia.com/blog/2020/05/14/sparsity-ai-inference/>

технология оптимизации математических расчетов

<https://www.techspot.com/article/2137-next-gen-directx-12/>

Графическое API DirectX12

<https://www.nvidia.com/en-gb/geforce/news/rtx-io-gpu-accelerated-storage-technology/>

Технология обращения напрямую к накопителю минуя процессор

<https://andiriney.ru/anatomiya-videokarty/>

устройство видеокарты

<https://se7en.ws/vse-nu-pochti-o-videokartakh-chast-1/>

краткое устройство видеокарты