Архитектура GPU

Первые компьютерные видеокарты имели лишь кадровый буфер: изображение формировалось центральным процессором компьютера и программным обеспечением, а карта отвечала за хранение кадров в буфере памяти и вывод их на монитор

Однако повышение требований к качеству изображения привело к созданию специализированного процессора, который занимается исключительно расчетом и формированием изображения, освобождая от этих обязанностей центральный процессор.

Современные ***графические процессоры*** по сложности не уступают центральным процессорам, и более того, во многих случаях в них используют технологии, опережающие применяемые в центральных процессорах. Как и центральные процессоры, графические процессоры характеризуются внутренней архитектурой, рабочей частотой графического ядра, технологическими нормами, по которым изготовлена микросхема.

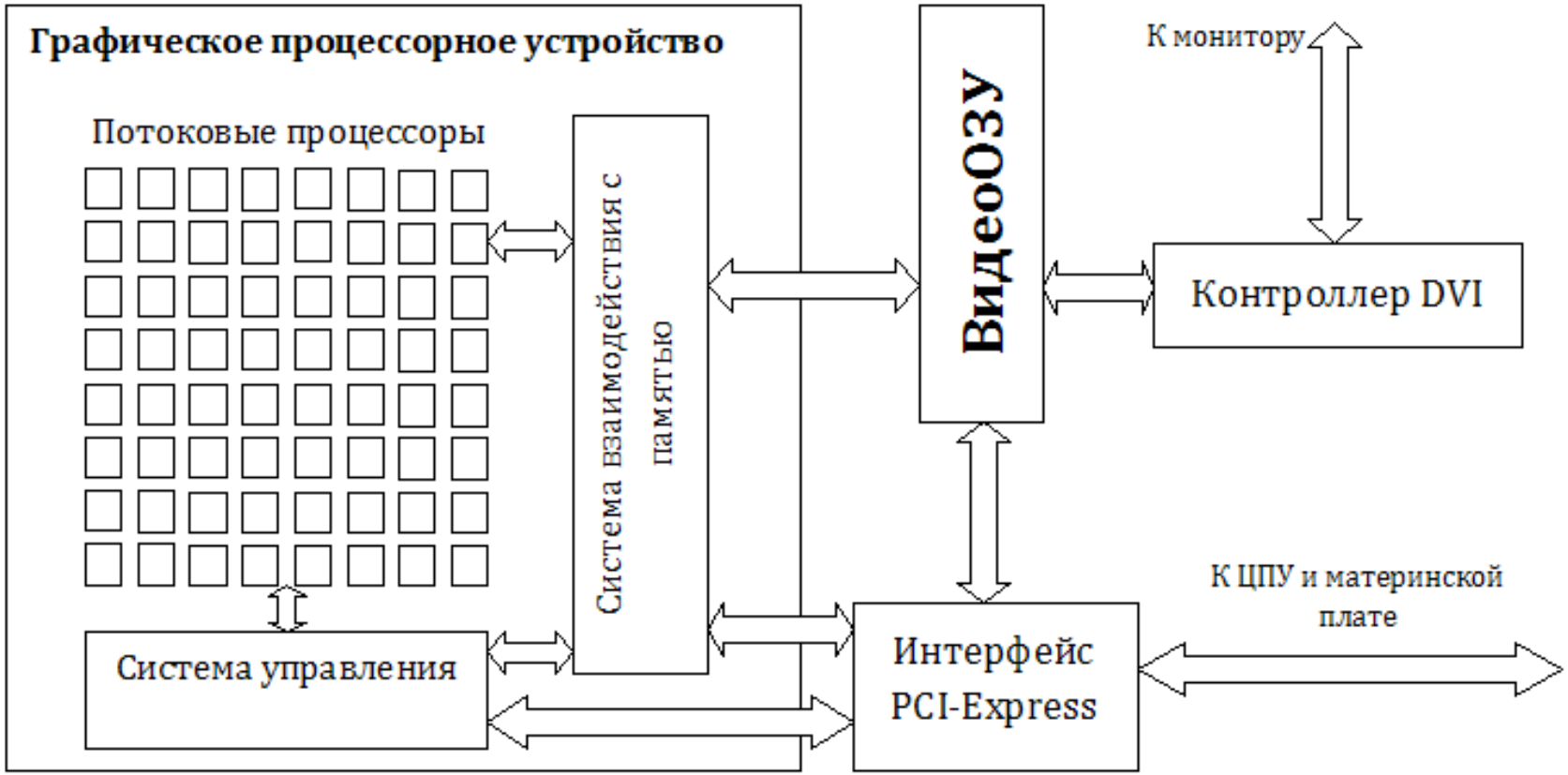
Тактовые частоты GPU ниже, чем у центрального процессора компьютера.

Однако благодаря большому количеству потоковых процессоров производительность GPU весьма значительна, а если еще учесть, что существует возможность установки на одну машину двух графических карт, то это позволяет получить пиковую производительность до нескольких ТФлопс.

Современные графические процессоры имеют довольно схожую архитектуру, они содержат набор одинаковых вычислительных устройств - потоковых процессоров, (**ПП, Thread Processor**), работающих с общей памятью графического процессора (**видеоОЗУ**)

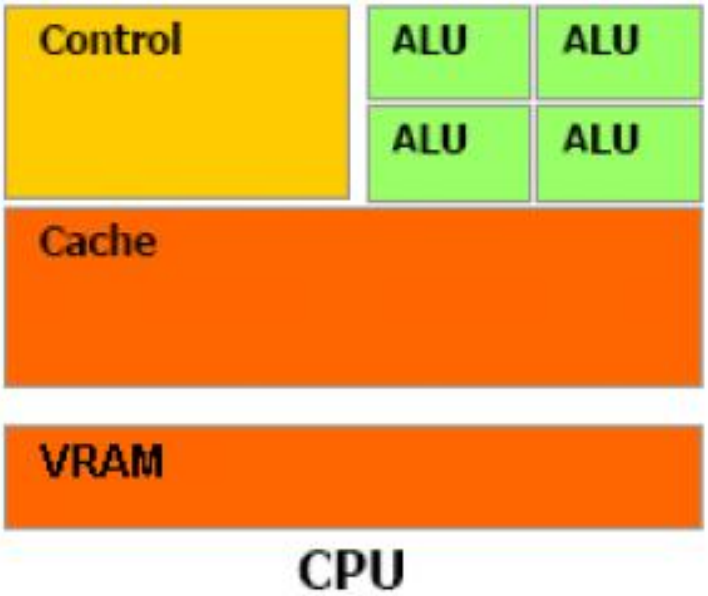
Число потоковых процессоров, а также размер видеоОЗУ может быть различным, в зависимости от модели GPU

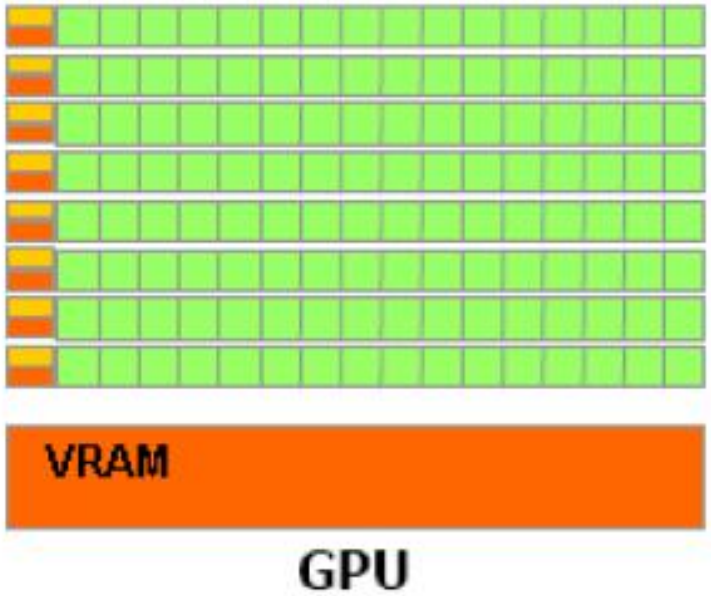
Все **ПП** синхронно исполняют одну и ту же команду, что позволяет отнести GPU к классу **SIMD** (англ. single instruction, multiple data — одиночный поток команд, множественный поток данных, **ОКМД** — принцип компьютерных вычислений, позволяющий обеспечить параллелизм на уровне данных).



Основные элементы видеоадаптера: видео память (VRAM) и графический процессор (GPU)

В отличие от обычных процессоров, имеющих одно устройство контроля исполнения, кэш и несколько арифметико-логических устройств, **графический процессор** состоит из набора мультипроцессоров, каждый из которых имеет своё устройство контроля исполнения (**Control**), область разделяемой (**Cache**) памяти, области регистров и значительно расширенный по сравнению с обычным процессором набор арифметико-логических устройств (**ALU**).

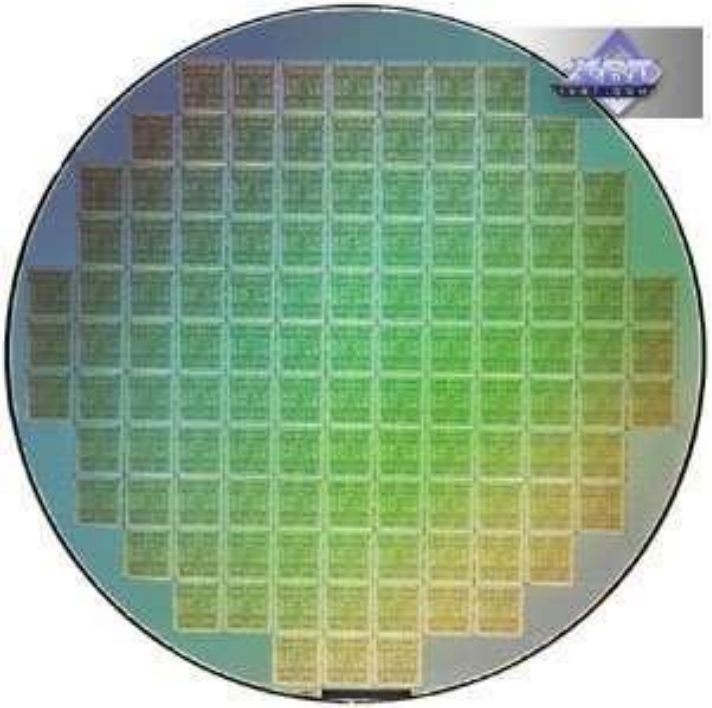




Такая структура позволяет добиться не только параллельного выполнения алгоритма на разных мультипроцессорах, но и параллельной обработки данных в рамках одного мультипроцессора.

Графический процессор G80

NVIDIA Tesla (2006) (128 пп) GeForce 8800



GeForce — бренд семейства графических процессоров и чипсетов материнских плат компании NVIDIA, ориентированного на потребительский рынок. Графические процессоры GeForce используются преимущественно в видеоадаптерах для персональных и переносных компьютеров.

Quadro — серия видеокарт, разработанных фирмой NVIDIA для пользователей профессиональных приложений

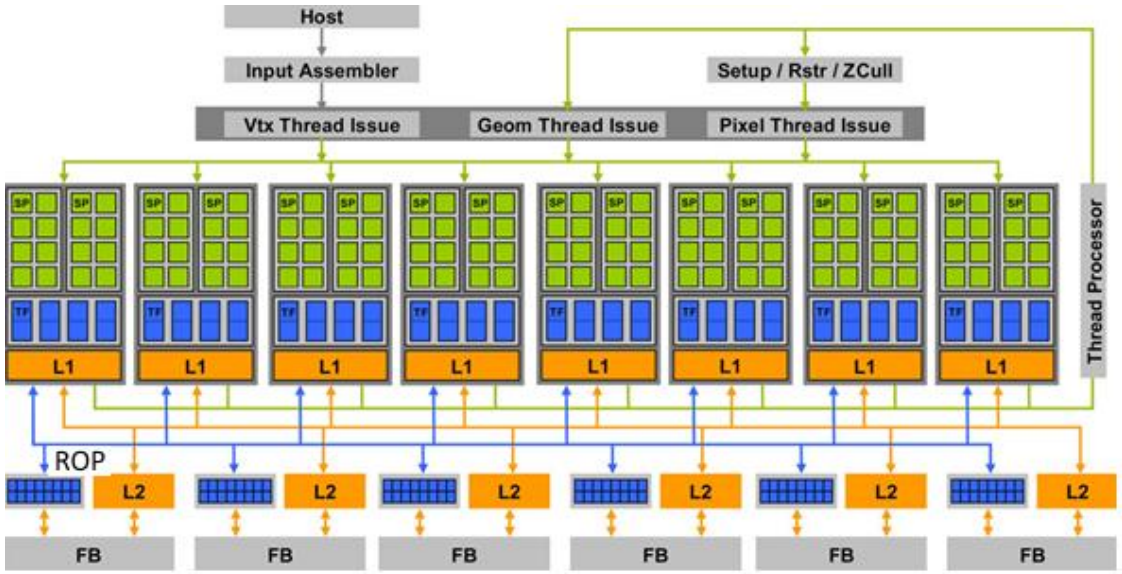
Tesla — название семейства вычислительных систем NVIDIA на основе графических процессоров с архитектурой CUDA, которые могут быть использованы для научных и технических вычислений общего назначения. Tesla строится на базе обычных графических процессоров, но, в отличие от видеоускорителей, не имеет средств вывода изображения на дисплей.

Tegra — семейство систем NVIDIA на кристалле (SoC — System-on-a-Chip), как платформа для производства мобильных интернет-устройств, таких как смартфоны, коммуникаторы, КПК и др. Кристалл Tegra объединяет в себе ARM-процессор, графический процессор, медиа- и DSP- процессоры, контроллеры памяти и периферийных устройств, имея при этом низкое энергопотребление.

Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
1.0	Tesla	G80	GeForce 8800 Ultra, GeForce 8800 GTX, GeForce 8800 GTS(G80)	Quadro FX 5600, Quadro FX 4600, Quadro Plex 2100 S4	Tesla C870, Tesla D870, Tesla S870

GeForce 8800 GTX

Структура графического процессора G80

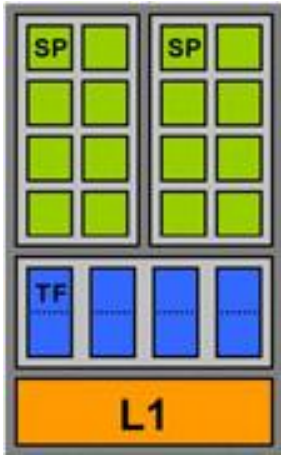


SP – Streaming Processor
TF – Texture Filtering Unit
L1/L2 – кэш
ROP - Raster Operations
FB – фрейм буфер

В состав процессора G80 входят 128 вычислительных ядер (SP - зеленые) общей *пиковой производительностью* 518 ГФлопс на тактовой частоте 1.35 ГГц.

Нитевые ядра SP объединены в 8 блоков по 16 в каждом, управляемых менеджером потоков и называемых **потоковыми мультипроцессорами**

Каждая *группа* имеет ***кэш-память*** первого (L1) и второго уровней (L2) (причем *кэш* второго уровня доступен для обращения всем остальным группам)



Потоковый *мультипроцессор* содержит 8 ядер *пиковой производительностью* 32 ГФлопс с поддержкой операций с числами в формате с плавающей точкой стандарта *IEEE 754,* что делает ее пригодной для серьезных, неигровых вычислений — например, научных, статистических или экономических.

Также в него входят:

- два блока специальных функций — чтение/запись в память, обработка текстур, выборки, переходы, вызовы процедур, операции синхронизации;
- блок разделяемой памяти 16 Кбайт;
- блок множественных инструкций.

Любой из 128 потоковых процессоров SP представляет собой обычное вычислительное устройство, способное работать с данными в формате с плавающей запятой. Следовательно, он не только способен обрабатывать шейдеры любого типа – вершинные, пиксельные или геометрические, но и использоваться для просчета физической модели или других расчетов, в рамках концепции (CUDA), причем, независимо от других процессоров.

Регистровый файл – единый банк регистров, на каждом SP имеется 32Кб. Самый быстрый тип памяти на графическом адаптере.

Разделяемая память – специальный тип памяти, предназначенный для совместного использования данных из одного блока. На каждом SM – 16Кб разделяемой памяти.

Основные *операции*, поддерживаемые вычислительными ядрами, следующие:

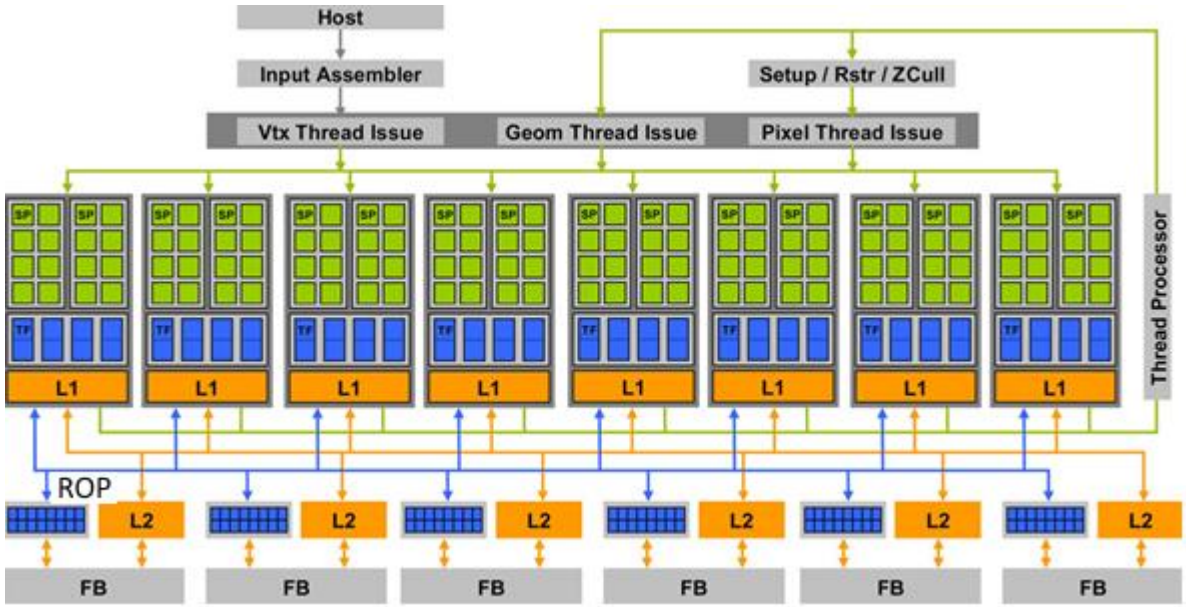
- FADD, FMUL, FMAD, FMIN, FMAX, FSET, F2I, I2F — округление до ближайшего четного, округление через 0, поддержка специальных чисел, сброс денормализованных операндов и результатов до 0;
- целочисленные инструкции: IADD, IMUL24, IMAD24, IMIN, IMAX, ISET, I2I SHR, *SHL*, AND, OR, XOR.

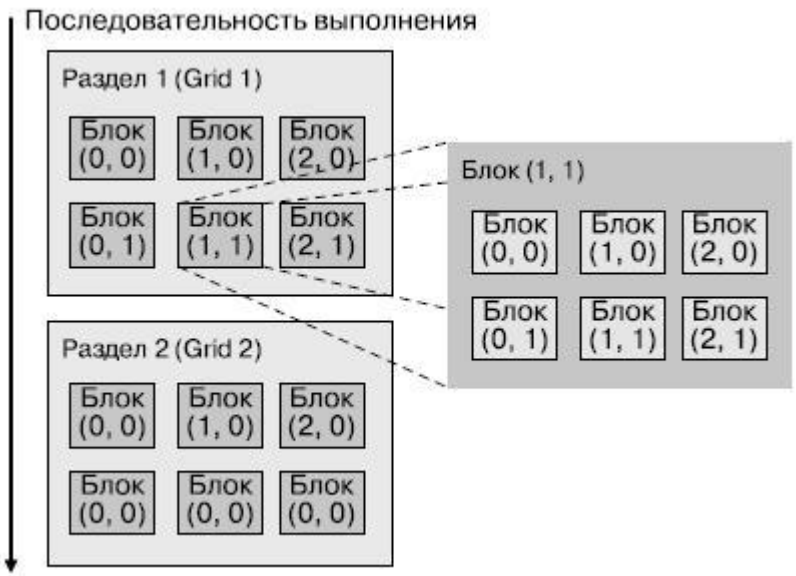
Выполнение инструкций конвейеризировано.

В состав ядер также входит блок специальных функций, который осуществляет поддержку выполнения трансцендентных функций: *RCP*, RSQRT, EXP2, LOG2, SIN, COS.

Блок специальных функций позволяет выполнять аппроксимацию функций при помощи квадратичной аппроксимации. *Точность* вычислений — порядка 22,5-24,0 *бит*.

Результаты вычислений группы *потоковых процессоров* записываются в *кэш* второго уровня и становятся доступны всем остальным группам. Таким образом, данные циркулируют внутри процессора и покидают его, только когда все вычисления завершены. Это устраняет задержки, возникающие при обращении к видеопамяти.

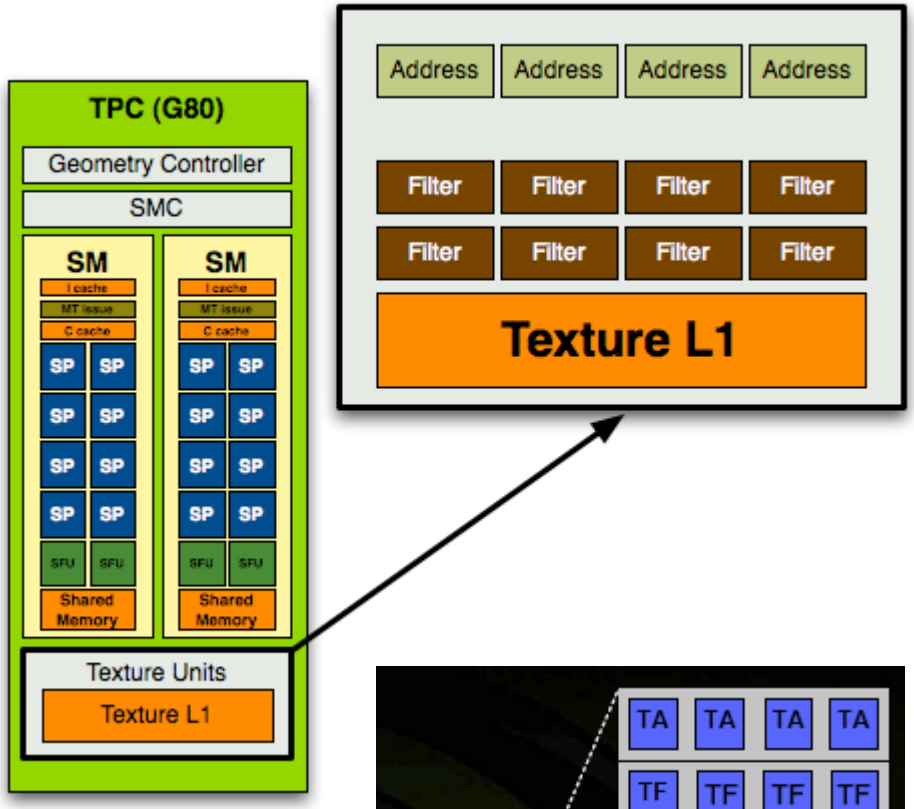




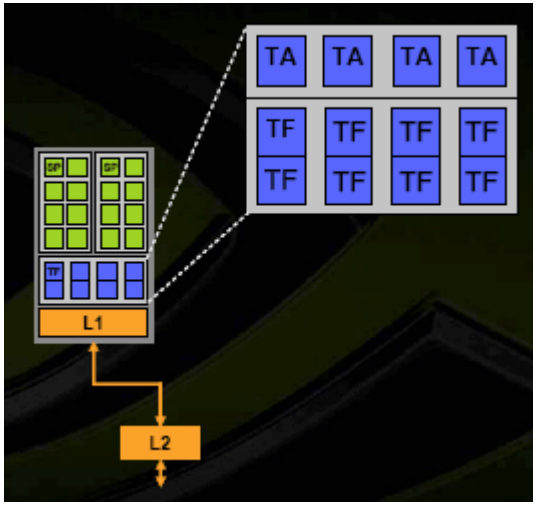
При решении прикладных задач на графическом процессоре разработчик должен предварительно разбить задачу на **разделы (grids**), причем один раздел представляет собой один шаг из последовательности шагов решения задачи.

В разделе задачи выделяются блоки, результаты выполнения которых могут быть получены параллельно — раздел распараллеливается.

В блоках выделяются элементы/нити, которые выполняются параллельно в режиме взаимодействия, — эти элементы будут соответствовать нитям, выполняющимся на отдельных ядрах



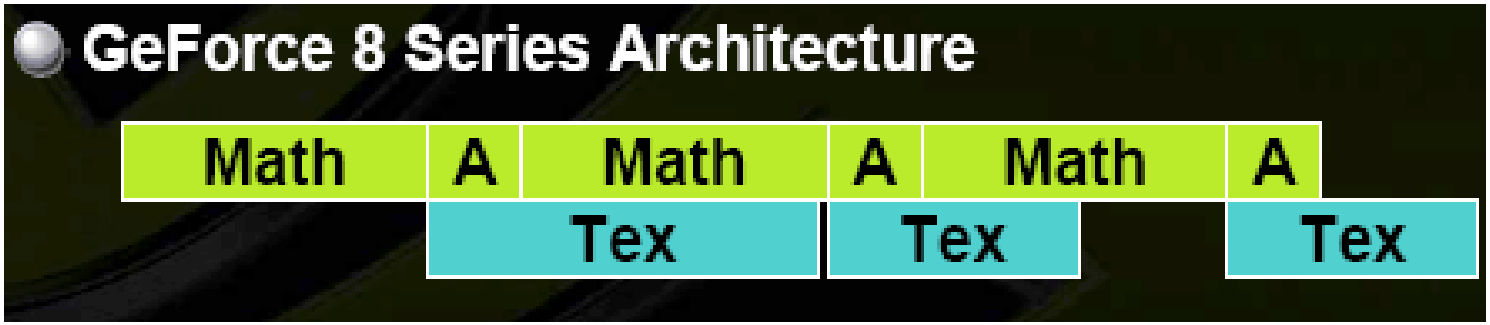
Вот так выглядел текстурный блок (Texture Unit) в G80.
На 8 блоков фильтрации приходилось только 4 блока адресации.



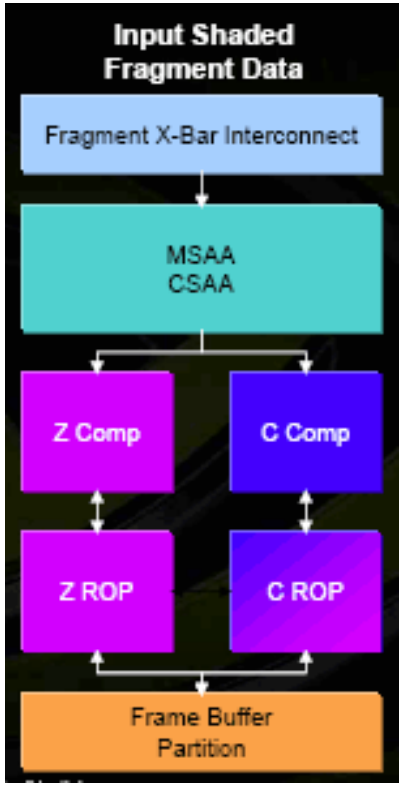
Каждый из блоков имеет четыре модуля адресации текстур (определения по координатам точного адреса для выборки - TA) и ровно такое же количество модулей билинейной фильтрации (TF).

Операция выборки и фильтрации текстур не требует ресурсов ALU и может теперь производиться полностью параллельно математическим вычислениям.

Генерация текстурных координат (на схеме — **A**) по-прежнему отнимает часть времени ALU. Это логично, ведь генерация текстурных координат требует стандартных плавающих операций, и заводить для нее отдельные ALU было бы непредусмотрительно.



Кроме шейдерных блоков, в наличии 6 блоков ROP (Raster Operations).



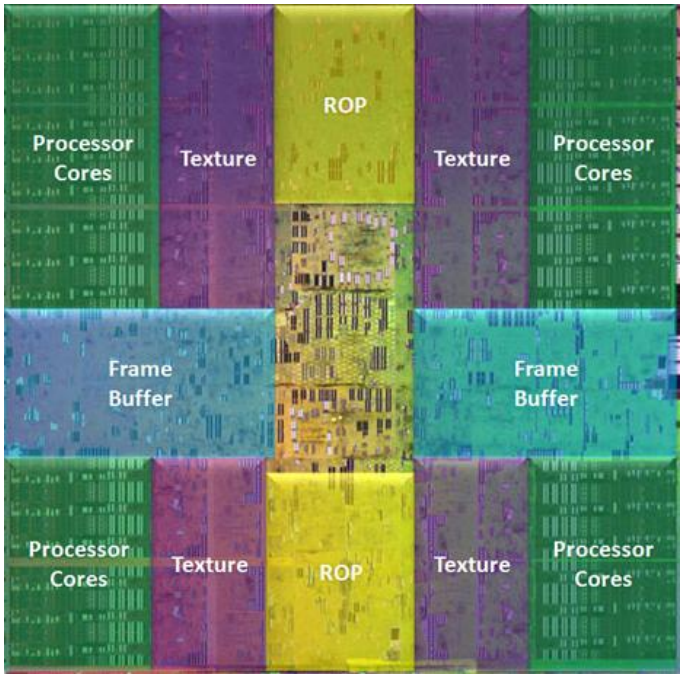
На диаграмме показаны два отдельных пути для Z и C, однако реально это один набор ALU, которые делятся на две группы при обработке пикселей с цветом, либо действуют как одна группа при обработке в режиме Z-Only, увеличивая, таким образом, пропускную способность вдвое.

В наше время нет смысла считать отдельные пиксели — важнее посчитать, сколько MSAA сэмплов может быть обработано за такт.

Соответственно, при MSAA 16x чип может выдавать 6 полноценных пикселей за такт, при 8x — 12 и т.д. Интересно, что масштабируемость работы с буфером кадров на высоте — как мы помним, каждый блок ROP работает с собственным контроллером памяти и не мешает соседним.

Графический процессор GT200

NVIDIA (2009) (240 пп)



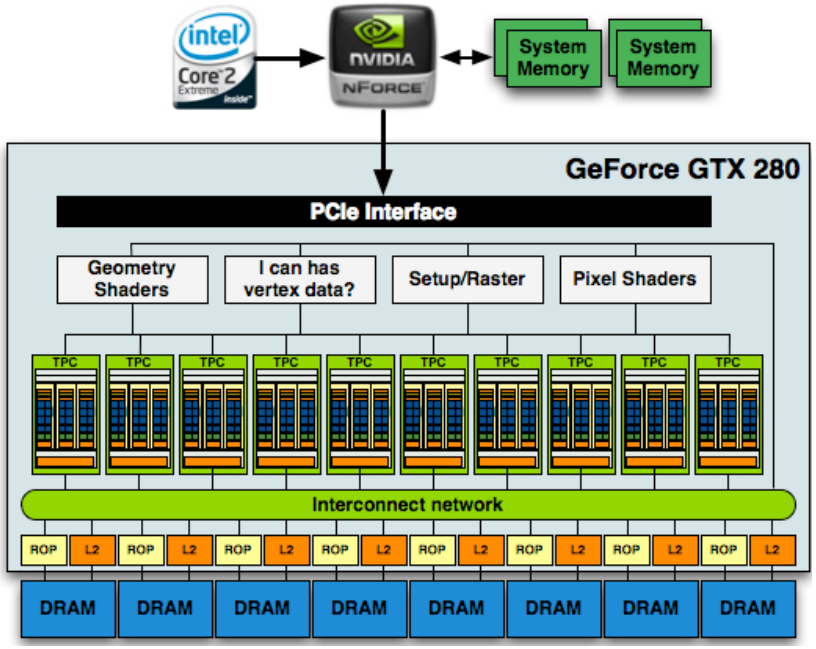
Расположение различных блоков в кристалле GT200

Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
1.3	Tesla	GT200, GT200b	GeForce GTX 295, GTX 285, GTX 280, GeForce GTX 275, GeForce GTX 260	Quadro FX 5800, Quadro FX 4800, Quadro FX 4800 for Mac, Quadro FX 3800, Quadro CX, Quadro Plex 2200 D2	Tesla C1060, Tesla S1070, Tesla M1060

GeForce GTX 260

Структура графического процессора GT200

В основе чипа GT200 лежит базовая архитектура G80. Главное преимущество нового GPU заключается в значительном увеличении количества вычислительных блоков, что вылилось в почти двукратное увеличение числа транзисторов – 1,4 млрд.



Новый чип состоит уже из десяти Texture Processing Clusters (TPC), которых в прошлом поколении графических процессоров было всего восемь.

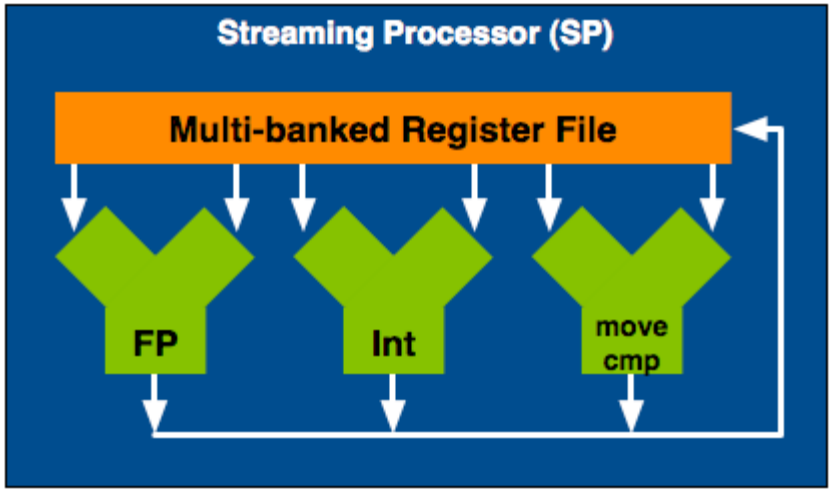
Кроме того, количество потоковых процессоров в TPC было увеличено с 16 до 24.

8 блоков ROP

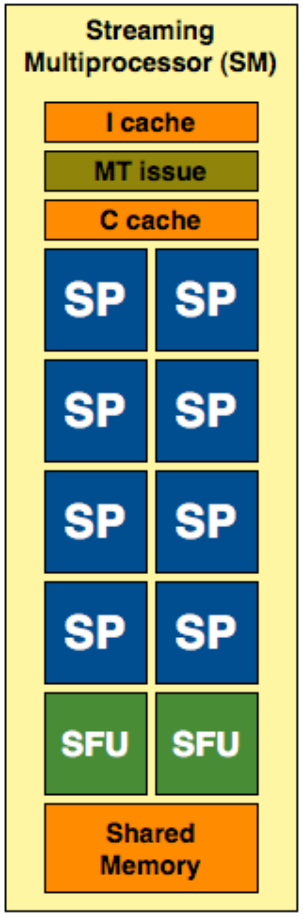
Итого общее количество потоковых процессоров равно 240 (24x10), текстурных блоков – 80, блоков ROP – 32.

Помимо увеличения количества вычислительных блоков и их организации они были доработаны, а их производительность улучшена.

Также было увеличено количество одновременно обрабатываемых потоков, а изменения в регистровом файле повысили эффективность обработки сложных шейдерных программ. Точность вычислений с плавающей запятой теперь была увеличена **до 64 бит**.



Самый низкий уровень составляют потоковые процессоры (SP – Streaming Processor). Каждый SP является самым настоящим микропроцессором с очередным типом исполнения команд, обладающим полноценным конвейером, парой ALU и FPU.



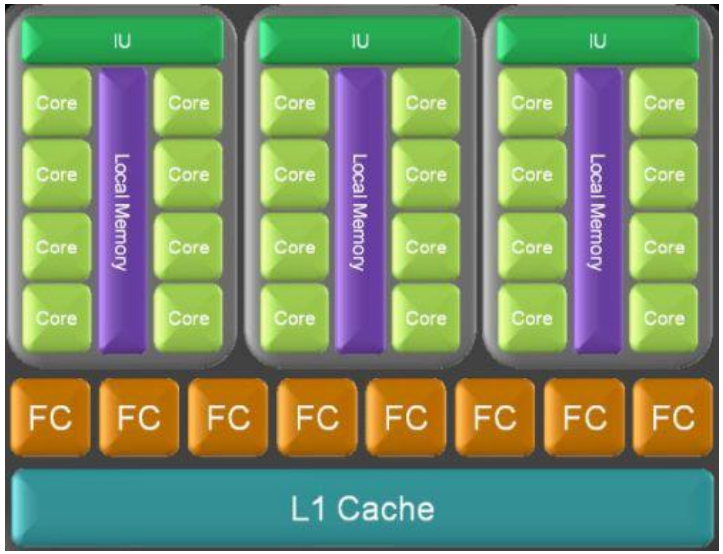
Первым уровнем объединения является потоковый мультипроцессор, или SM – Streaming Multiprocessor.

Потоковый мультипроцессор представляет собой массив из восьми SP, которые находятся в группе с двумя модулями специальных функций SFU – Special Function Units.

Каждый SFU содержит в своем составе четыре FPU, заточенные для трансцендентных операций (sin, cos и т.д.) и интерполяции, которые часто используются при расчетах, связанных, например, с анизотропной фильтрацией. SFU является в отдельности таким же полноценным микропроцессором, как и SP.

В SM также входит диспетчер исполнения команд MT, который занимается распределением нагрузки по SP и SFU внутри группы.

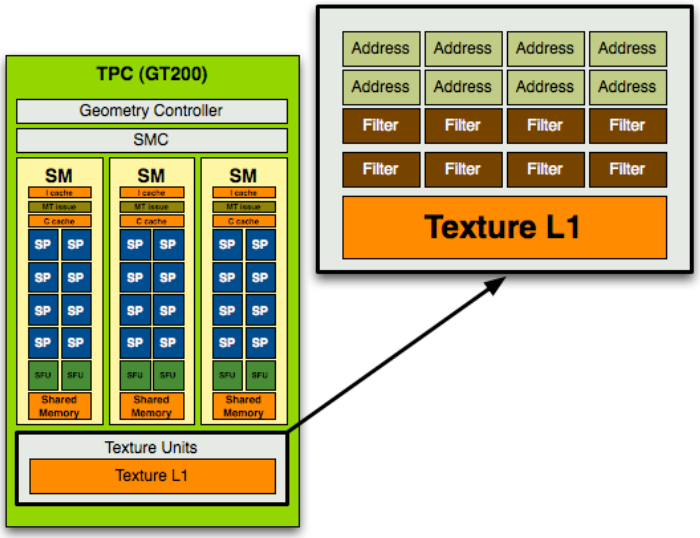
Вдобавок к SP, SFU и MT в мультипроцессоре содержится и небольшой объем памяти (16 Кб - сделано это специально, так как каждый SP работает со своим пикселем и общие объемы информации невелики), общий для всех процессоров.



Следующим уровнем объединения является кластер SM, называемый Texture/Processor Cluster (TPC).

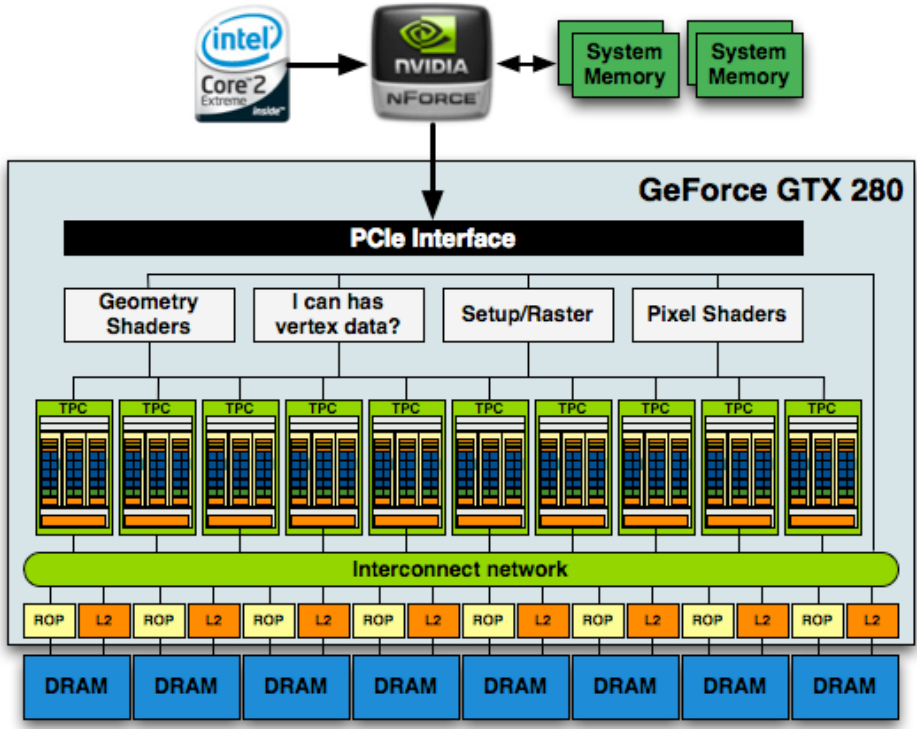
Внутри одного кластера шейдерные процессоры объединены в три блока-юнита со своей выделенной областью локальной памяти.

Также на каждый кластер приходится по 8 текстурных блоков



Вот так выглядел текстурный блок в GT200

Уровнем выше располагается общая управляющая логика чипа, распределяющая и планирующая нагрузки на различные кластеры, контроллер PCI-Express 2.0 и шина Interconnect Network, соединяющая процессорную мощь SPA с уровнем L2 кэша текстур и блоками обработки растровой графики (Raster Operation Unit - ROP), которые в свою очередь уже имеют прямой доступ к фрейм буферу.



Графический процессор GF100

NVIDIA Fermi (2010) **(512 пп)** **GeForce GTX 580**
Quadro 600 - 6000

Quadro 5000/6000

Quadro 4000

Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
2.0	Fermi	GF100 GF110	GeForce GTX 590, GeForce GTX 580, GeForce GTX 570, GeForce GTX 480, GeForce GTX 470, GeForce GTX 465, GeForce GTX 480M	Quadro 6000, Quadro 5000, Quadro 4000, Quadro 4000 for Mac, Quadro Plex 7000, Quadro 5010M, Quadro 5000M	Tesla C2075, Tesla C2050/C2070, Tesla M2050/M2070/M2075/M2090

GeForce GTX 590

Quadro 5000/6000

Quadro 4000

Структура графического процессора GF100

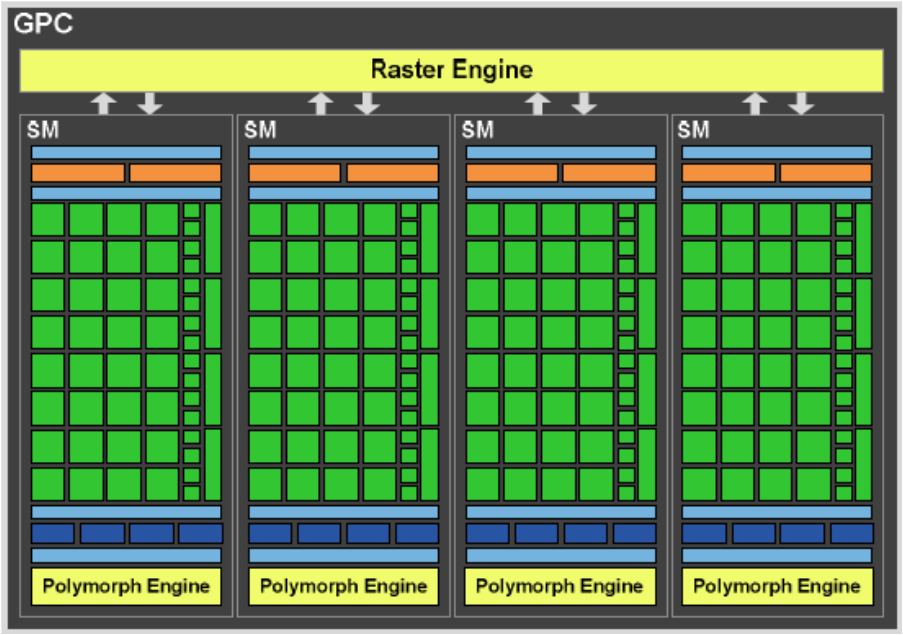
GF100 (Fermi) это:

- 512 CUDA-процессоров
- 16 геометрических блоков
  SM - мультипроцессор
- 4 блока растеризации
  GPC-Graphics Processing Cluster
- 64 текстурных блока
- 48 модулей ROP
- 384-битный интерфейс памяти GDDR5

В GF100 внесено множество усовершенствований, направленных на обработку сложной графической информации.

Графический процессор GF100 построен на базе масштабируемой архитектуры, в основе которой лежит применение объединенных в кластеры GPC (Graphic Processing Cluster – кластеры обработки графики) потоковых мультипроцессоров (SM, Streaming Multiprocessor).

GPC (Graphic Processing Cluster – кластеры обработки графики)

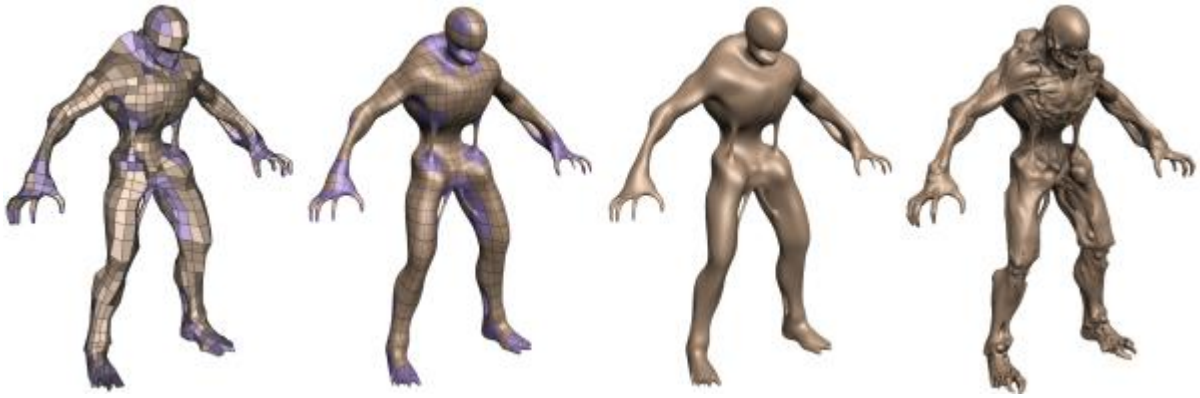


Каждый такой кластер содержит четыре мультипроцессора (SM), а также все необходимые блоки для обработки геометрических данных и текстурирования.

Фактически, каждый GPC представляет собой самостоятельный графический процессор, не имеющий лишь собственной подсистемы памяти.

GF100 состоит из четырех таких кластеров, совместно использующих шесть контроллеров памяти, шесть модулей ROP (по 8 блоков ROP в каждом) и L2-кеш.

В настоящее время все объекты и персонажи игр разрабатываются в программах трехмерного моделирования. Дизайнеры должны вручную создавать несколько моделей с разным уровнем деталей (LOD, Level Of Details), используя ту или иную в зависимости от удаленности объекта от переднего края игровой сцены.



Учитывая, что каждый объект передается в GPU заново для каждого кадра, требуются достаточно сложные алгоритмы для использования модели с оптимальным в данный момент уровнем детализации. Причем существенное ограничение накладывает не только предельная производительность GPU при обработке геометрической информации, но и пропускная способность интерфейса PCI Express.

Метод тесселяции, основанный на картах смещения, позволяет в значительной степени обойти эту проблему.

Напомним, что карта смещения – это монохромная текстура, используемая не для закрашивания полигона, а для изменения его геометрических свойств.

Яркость каждой точки на этой текстуре определяет отклонение (высоту) это точки над исходной поверхностью.

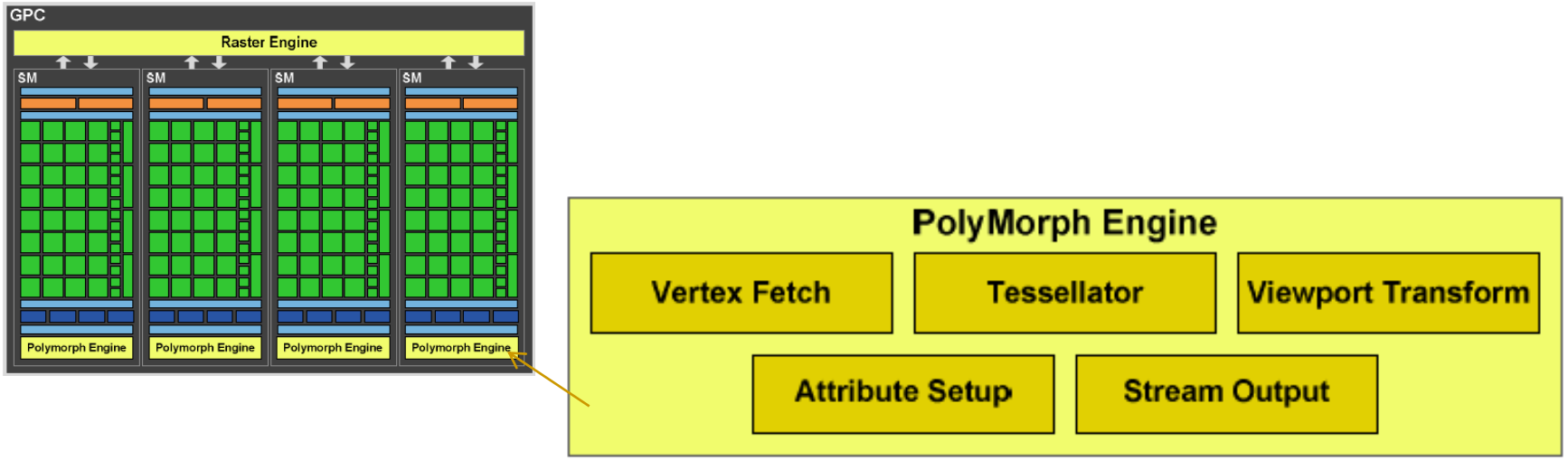
В отличие от традиционных методов, когда объем имитируется обычными плоскими текстурами, тесселяция позволяет получить гораздо более сложные и естественно выглядящие объекты, правильно рассчитывать тени и т.п.

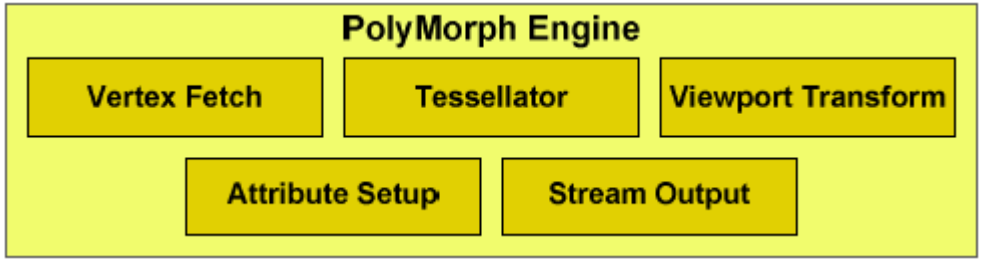
Огромное преимущество карт смещения в том, что они позволяют создать универсальную модель, уровень детализации которой определяется лишь используемой картой смещения. Важно также, что по сути карта смещения – это обычная текстура, методы оптимизации и компрессии которой уже давно и успешно отработаны. В графическом процессоре GF100 реализованы средства аппаратной поддержки теселляции, причем инженеры NVIDIA уделили этому аспекту максимум внимания.

Использование тесселляции фундаментальным образом изменило распределение нагрузки внутри графического процессора и вынудило инженеров NVIDIA несколько изменить компоновку вычислительных блоков и ввести новый тип блока – PolyMorph Engine.

Каждый графический кластер (GPC) оснащен четырьмя такими блоками – по одному на каждый мультипроцессор (SM).

Каждый PolyMorph Engine выполняет пять стадий: выбор вершин, тесселяция, преобразование координат, преобразование атрибутов, потоковый вывод.





На первом этапе вершины выбираются из глобального буфера, после этого вершина отправляется в мультипроцессор, где ее координаты преобразуются в координаты сцены и определяется уровень тесселяции (аналог уровня детализации, LOD).

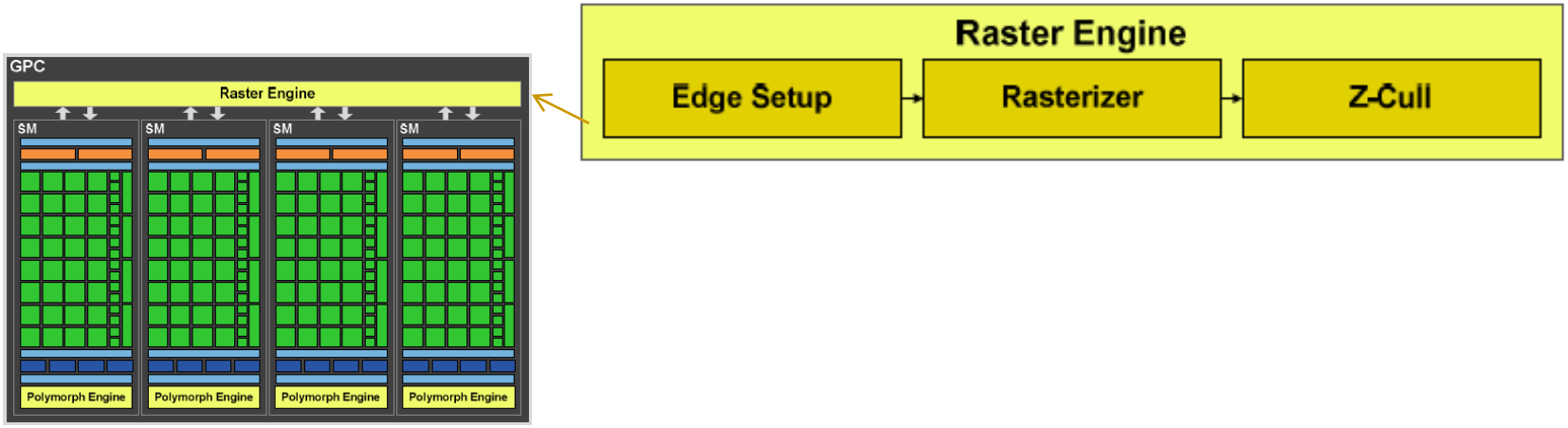
После этого вершина передается на второй этап – тесселяцию. На этом этапе полигон разбивается на несколько новых, более мелких, по карте смещения определяются их координаты.

Полученные новые вершины вновь обрабатываются в мультипроцессоре и передаются через потоковый вывод в память для дальнейшей обработки.

После того, как геометрические данные обработаны в PolyMorph Engine они передаются для растеризации в Raster Engine.

В этом блоке отфильтровываются невидимые примитивы (невидимые поверхности), затем геометрические данные преобразуются в экранные точки, которые в свою очередь сортируются и фильтруются по Z-координате.

Каждый кластер (GPC) оснащен одним блоком растеризации, обрабатывающим до 8 точек за такт, то есть суммарная производительность GF100 составляет 32 точки за такт – это в 8 раз больше, чем обеспечивал GT200.



Каждый мультипроцессор состоит из 32 вычислительных блоков CUDA. Каждый процессор CUDA оснащен одним логическим блоком ALU и одним FPU.

Кроме того, каждый мультипроцессор оснащен 16 Load/Store-блоками, позволяющими определить адреса данных в кеше или памяти для 16 потоков за каждый такт.

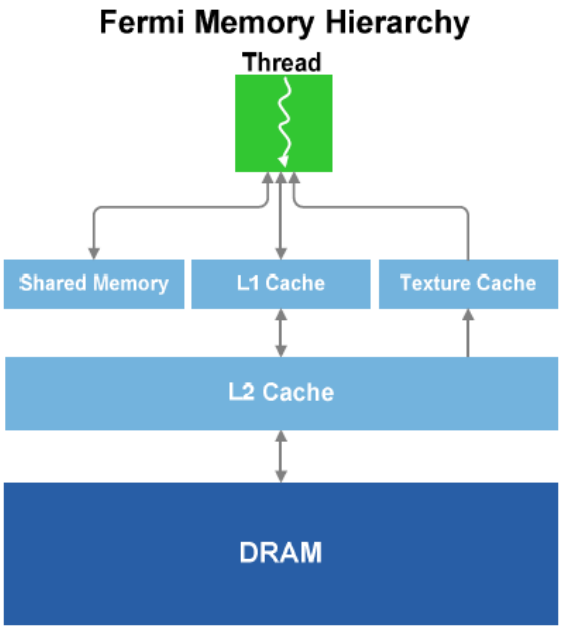
Предусмотрены и четыре блока специальных функций (SFU - Special Function Unit), выполняющих такие операции как синус, косинус, квадратный корень. Каждый SFU выполняет одну операцию на поток за такт, так что ветвь (warp, 32 потока) выполняется за 8 тактов.

Мультипроцессор организует потоки в ветви по 32 потока, для управления этими ветвями используется два планировщика ветвей – две ветви могут выполняться на одном мультипроцессоре одновременно.

Планировщики GF100 передают по одной инструкции от каждой ветви группе из 16 ядер CUDA, 16 блоков LD/ST или четырех SFU.

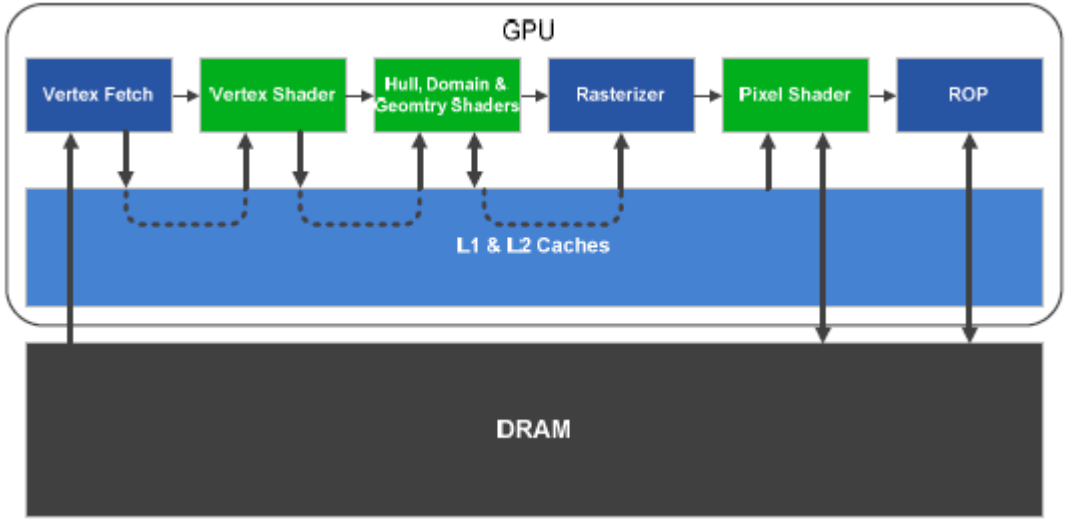
Кроме того, каждый SM оснащен четырьмя текстурными блоками – каждый из них отбирает до четырех текстурных семплов за такт, результат может быть сразу же отфильтрован – предусмотрена билинейная, трилинейная и анизотропная фильтрация.

Общая память и кеш. Общая память – это быстрая, программируемая, расположенная в микросхеме память, позволяющая максимально оптимизировать обмен данными внутри потока.



В GF100 помимо общей памяти используется также L1-кеш, собственный внутри каждого мультипроцессора (**SM**).
L1-кеш работает в паре с общей памятью, в то время как общая память предназначена для алгоритмов с упорядоченным доступом к памяти, L1-кеш ускоряет те алгоритмы, где адреса данных не известны заранее.

В GF100 каждый мультипроцессор оборудован 64 Кб памяти, которая может быть поделена на 48 Кб общей памяти и 16 Кб L1-кеша или наоборот. Кроме того, предусмотрен унифицированный L2-кеш объемом 768 Кб. Он обеспечивает максимально быстрый обмен данными между различными блоками



Блоки ROP (Растерные операции**)**

Блоки ROP в GF100 организованы порциями по восемь блоков в каждом. Каждый блок может вывести 32-бит целое значение за такт, либо FP16 за два такта, либо FP32 – за четыре. Благодаря улучшенным алгоритмам сжатия и увеличенному количеству ROP заметно ускорено сглаживание 4x и 8x MSAA – оно выполняется соответственно в 1,6 и 2,3 раза быстрее, чем GT200.

Стоит отметить, что GF100 выполняет сглаживание 8xAA лишь на 9% медленнее, чем 4xAA. Кроме того, в GF100 реализован новый режим сглаживания – 32xCSAA (Coverage Sample Antialiasing).

В 2010 году NVIDIA представила **GF100** с архитектурой Fermi, который лег в основу топовой видеокарты GeForce GTX 480. GF100 производился по 40-нм техпроцессу и получил 512 потоковых процессоров, или, как их называли в NVIDIA, ядер CUDA.

Видеокарта GeForce GTX 480

Однако видеокарта GTX 480 не задействовала все процессоры: из 512 работали только 480 штук.

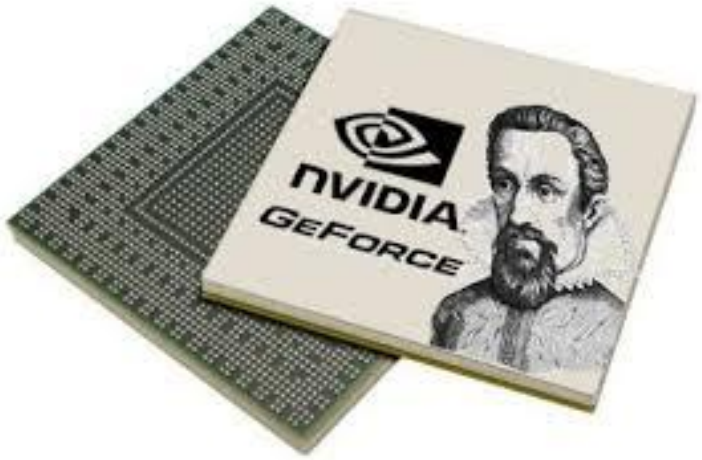
В конце 2010 г. была представлена видеокарта GeForce GTX 580 на базе обновленного **GF110**. Адаптер работал со всеми 512 ядрами CUDA и 64 текстурными модулями, что дало прибавку в производительности.

Видеокарта GeForce GTX 580

Начиная с 2011 года NVIDIA выпустила по два новых поколения графических ускорителей.

Сначала NVIDIA выпустила 28-нм чип GK104 с абсолютно новой архитектурой Kepler, которая была разработана с целью увеличения производительности на ватт. Kepler рассчитан на энергоэффективность.

NVIDIA Kepler (2011-2012) (1024 пп)



GeForce GTX 680

Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
3.0	Kepler	GK104, GK106, GK107	GeForce GTX 770, GeForce GTX 760, GeForce GT 740, GeForce GTX 690, GeForce GTX 680, GeForce GTX 670, GeForce GTX 660 Ti, GeForce GTX 660, GeForce GTX 650 Ti BOOST, GeForce GTX 650 Ti, GeForce GTX 650, GeForce GTX 880M, GeForce GTX 780M, GeForce GTX 770M, GeForce GTX 765M, GeForce GTX 760M, GeForce GTX 680MX, GeForce GTX 680M, GeForce GTX 675MX, GeForce GTX 670MX, GeForce GTX 660M, GeForce GT 750M, GeForce GT 650M, GeForce GT 745M, GeForce GT 645M, GeForce GT 740M, GeForce GT 730M, GeForce GT 640M, GeForce GT 640M LE, GeForce GT 735M, GeForce GT 730M	Quadro K5000, Quadro K4200, Quadro K4000, Quadro K2200, Quadro K2000, Quadro K2000D, Quadro K600, Quadro K420, Quadro K500M, Quadro K510M, Quadro K610M, Quadro K1000M, Quadro K2000M, Quadro K1100M, Quadro K2100M, Quadro K3000M, Quadro K3100M, Quadro K4000M, Quadro K5000M, Quadro K4100M, Quadro K5100M, NVS 510	Tesla K10, GRID K340, GRID K520

Весной 2013 года NVIDIA представила чип GK110, на котором основываются флагманские видеокарты GeForce GTX 780 и GeForce GTX Titan.

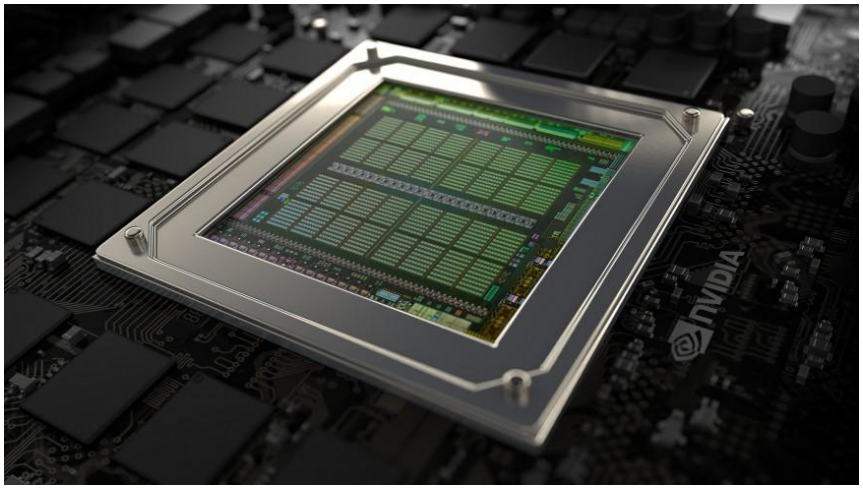
NVIDIA Maxwell (2013) (2880 пп) GK110

GeForce GTX 780

Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
3.5	Kepler	GK110, GK208	GeForce GTX Titan Z, GeForce GTX Titan Black, GeForce GTX Titan, GeForce GTX 780 Ti, GeForce GTX 780, GeForce GT 640 (GDDR5), GeForce GT 630 v2, GeForce GT 730, GeForce GT 720, GeForce GT 710,GeForce GT 740M (64-bit, DDR3)	Quadro K6000, Quadro K5200	Tesla K40, Tesla K20x, Tesla K20

В 2014 году NVIDIA представила GM204 с архитектурой (Maxwell)

Это более сложный чип. Он имеет большую полезную площадь и, как следствие, большее число транзисторов. GM204 состоит из 5,2 миллиарда кремниевых затворов.



GM204

Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
3.5	Maxwell	GK110, GK208	GeForce GTX Titan X, GeForce GTX 980 Ti, GeForce GTX 980, GeForce GTX 970, GeForce GTX 960, GeForce GTX 950, GeForce GTX 750 SE, GeForce GTX 980M, GeForce GTX 970M, GeForce GTX 965M	Quadro M6000 24GB, Quadro M6000, Quadro M5000, Quadro M4000, Quadro M2000, Quadro M5500, Quadro M5000M, Quadro M4000M, Quadro M3000M	Tesla M4, Tesla M40, Tesla M6, Tesla M60

GeForce GTX 980

Структура графического процессора GM204

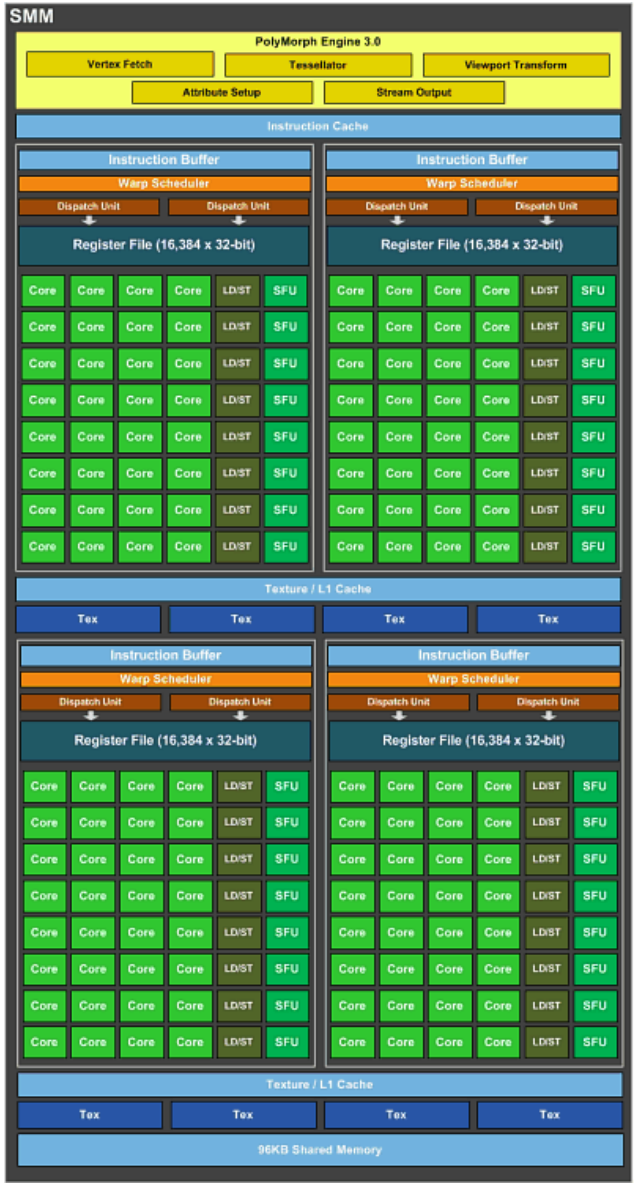
Как Fermi и Kepler, чип GM204 (Maxwell) состоит из кластеров графической обработки Graphics Processing Cluster (GPC), которые содержат по несколько потоковых мультипроцессоров (в данном случае — SMM).

Каждый кластер GPC содержит выделенный движок растеризации и по четыре мультипроцессора SMM.

В свою очередь, каждый мультипроцессор состоит из 128 вычислительных блоков CUDA (потоковых процессора), блока обработки геометрии PolyMorph и восьми блоков текстурирования (TMU).

Модель GeForce GTX 980 является верхним представителем линейки, содержащем четыре кластера GPC и 16 мультипроцессоров SMM. В целом это составляет 2048 потоковых процессоров CUDA и 128 блоков TMU.

GeForce GTX 970 же содержит те же четыре GPC, но в которых отключены три из имеющихся шестнадцати мультипроцессоров. То есть, в итоге «урезанный» чип имеет 1664 потоковых процессора и 104 блока TMU.



Обе модели линейки GeForce GTX 900 имеют по четыре 64-битных контроллера памяти, что в сумме составляет 256-битную шину памяти.

К каждому контроллеру памяти «привязаны» по 16 блоков ROP и 512 КБ кэш-памяти второго уровня, что в целом составляет 64 блока ROP и 2048 КБ кэша.

В архитектуре Maxwell применяются абсолютно новые потоковые мультипроцессоры (Streaming Multiprocessor — SM), которые имеют как лучшую энергоэффективность, так и производительность по отношению к площади чипа.

Организация мультипроцессоров серьезно изменилась. Каждый мультипроцессор GM204 разделен еще на четыре отдельных логических вычислительных раздела, каждый из которых имеет свой буфер инструкций, планировщик варпов (инструкция над группой потоков) и состоит из 32 вычислительных ядер.

В чипах архитектуры Kepler каждый мультипроцессор SMX содержит управляющую логику, которая распределяет и планирует работу и обмен данными для 192 вычислительных ядер, поэтому управляющий блок довольно сложен сам по себе.

В архитектуре Maxwell было принято решение разделить мультипроцессор SMM на вычислительные блоки, каждый из которых содержит собственный блок управления, обслуживающий лишь 32 ядра и поэтому более простой.

В апреле 2016 года Nvidia представила GPU GP100 поколения Pascal, содержащий 3840 ядер CUDA.

Видеокарта MSI GeForce GTX 1080 — это производительный видеоадаптер топового уровня. 1
8 Гб видеопамяти стандарта GDDR5X с высокой пропускной способностью. Максимальное разрешение экрана, поддерживаемое устройством, составляет 7680 на 4320 пикселей.

Данная модель позволяет использовать сразу два аналогичных адаптера благодаря поддержке технологии SLI.

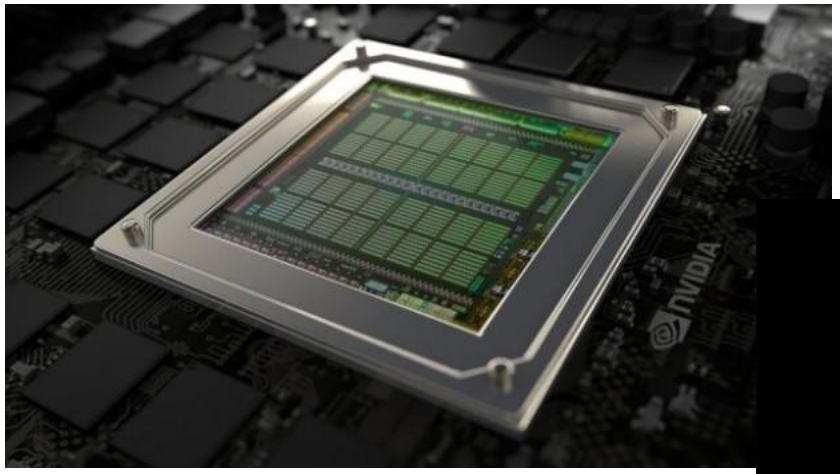
Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
6.1	Pascal	GP102, GP104, GP106, GP107, GP108	Nvidia TITAN Xp, Titan X, GeForce GTX 1080 Ti, GTX 1080, GTX 1070, GTX 1060, GTX 1050 Ti, GTX 1050, GT 1030	Quadro P6000, Quadro P5000, Quadro P4000, Quadro P2000, Quadro P1000, Quadro P600, Quadro P400, Quadro P5000(Mobile), Quadro P4000(Mobile), Quadro P3000(Mobile)	Tesla P40, Tesla P6, Tesla P4

NVIDIA представила в 2017 году первые карты с микроархитектурой **Volta**. GV100 — новый флагманский GPU Nvidia поколения **Volta**. Он состоит из 5376 ядер CUDA.

Версия	Архитектура	GPU	GeForce	Quadro, NVS	Tesla
7.0	Volta	GV100			Tesla V100

GV100 является одним из самых мощных графических процессоров компании и демонстрирует производительность в 7,5 терафлопс в двойной точности и 15 терафлопс в обычной точности.

Процессоры Tesla V100 используются в суперкомпьютерах и системах искусственного интеллекта



Серия GTX 11 действительно должна обеспечить резкое увеличение графического выхода. Флагманский диапазон 11-й серии обещает 4K-игры, которые станут гораздо ближе к гладким 60 кадрам в секунду

Volta гораздо мощнее предыдущих графических процессоров на базе Pascal. Но NVIDIA сдерживала выпуск ускорителей на этом процессоре, чтобы не составить конкуренцию с серией ускорителей на Pascal

16 сентября 2018 г. состоялась презентация нового поколения ускорителей NVIDIA 2070

Основной особенностью новых карточек является применение новых графический процессоров поколения Turing, которые поддерживают аппаратное ускорение трассировки световых лучшей (RayTracing).

Обрабатывают лучи специальные новые ядра Nvidia RT - отсюда и пошло название RTX.

NVIDIA GeForce RTX 2080 Ti — 11 ГБ видеопамяти, 10 миллиардов лучей и 78 трлн RTX-операций в секунду;
NVIDIA GeForce RTX 2080 — 8 ГБ видеопамяти, 8 миллиардов лучей и 60 трлн RTX-операций в секунду;
NVIDIA GeForce RTX 2070 — 8 ГБ видеопамяти, 6 миллиардов лучей и 45 трлн RTX-операций в секунду.

GEFORCE RTX™ 2080 Ti	GEFORCE RTX™ 2080	GEFORCE RTX™ 2070
Founders Edition	Founders Edition	Founders Edition
$ 1,199.00	$ 799.00	$ 599.00

Nvidia GeForce RTX 2080 Ti - от 95990 рублей;
Nvidia GeForce RTX 2080 - от 63990 рублей;
Nvidia GeForce RTX 2070 - от 47990 рублей.

Модель GeForce RTX 2070 предлагает до 2304 ядер CUDA, работающих на частоте до 1410 МГц в стоке . Также у GeForce RTX 2070 имеется на борту до 8 ГБ памяти типа GDDR6 с 256-битной шиной. Пропускная способность видеокарты составляет 448 Гбайт/сек.

GeForce RTX 2080 имеет более высокие характеристики, а именно чип TU104 с 2944 ядрами CUDA. Минимальная частота видеокарты составляет 1515 МГц, в бусте 1710 МГц и 1800 МГц в Founders EditionУвидеоускорителя имеется те же 8 ГБ GDDR6 с 256-битной шиной и пропускной способностью до 448 ГБ/с.

GeForce RTX 2080 Ti самая продвинутая видеокарта имеет графический процессор TU102 с 4352 ядрами CUDA. Стоковая частота у них 1545 МГц, а пиковая 1635 МГц. Карточка имеется 11 ГБ памяти типа GDDR6 с 353-битной шиной и пропускной способностью до 616 ГБ/сек.

Соединять несколько видеокарт в пары теперь можно при помощи нового интерфейса под названием NVLink. Им оборудованы только две старшие видеокарты Nvidia GeForce RTX 2080 Ti и RTX 2080. Производитель обещает более высокую пропускную способность, возросшую в 50 раз по сравнению с SLI прошлого поколения.

Стоимость одного переходника для связки видеокарт немногим превышает 6000 рублей.

Ещё одним новым интерфейсом стал VirtualLink, выполненный в форме разъёма USB Type-C. Новый стандарт VirtualLink позволяет передавать видео, данные и питание к шлемам виртуальной реальности через один кабель.

Более ста суперкомпьютеров из Top 500 используют ускорители вычислений.

В общей сложности 104 системы в списке пятисот самых мощных суперкомпьютеров в мире используют вычислительные ускорители AMD FirePro S, NVIDIA Tesla или сопроцессоры Intel Xeon Phi.

66 систем из 104 использует NVIDIA Tesla, 27 используют Intel Xeon Phi и три используют AMD FirePro.

Самым мощным суперкомпьютером на планете по-прежнему является китайский Tianhe 2. Он способен выполнять 33,86 квадриллионов операций с плавающей запятой в секунду (33,86 петафлопс). Данная система базируется на процессорах Intel Xeon E5-2692v2 (двенадцать ядер, тактовая частота 2,2 ГГц), а также использует 48 тысяч сопроцессоров Xeon Phi Intel 31S1P. В общей сложности, Tianhe 2 располагает 3,12 миллионами x86-ядер.

На втором месте находится суперкомпьютер Titan, который установлен в национальной лаборатории Ок-Ридж (Oak Ridge) министерства энергетики США, и обладает мощностью 17,59 петафлопс. Данная система представляет собой кластер Cray XK7 с шестнадцатиядерными AMD Opteron и ускорителями NVIDIA Tesla K20X.

Tianhe 2

Ломоносов

Самым мощным суперкомпьютером в России является Ломоносов 2, который имеет мощность 1,849 петафлопс (36 место в рейтинге). Данная система была построена компанией T-Platforms для МГУ имени М.В. Ломоносова на базе процессоров Intel Xeon E5-2697v3 (14 ядер, 2,60 ГГц) и ускорителей NVIDIA Tesla K40m.