Архитектура графических процессоров и GPGPU

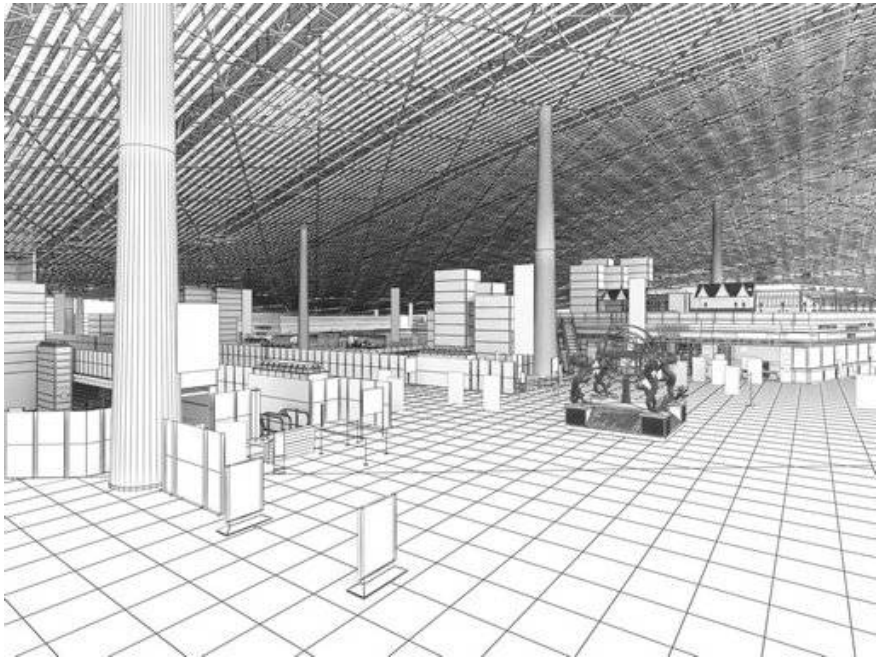
Алгоритмы машинной графики (потоковые,вычислительные, параллельные):

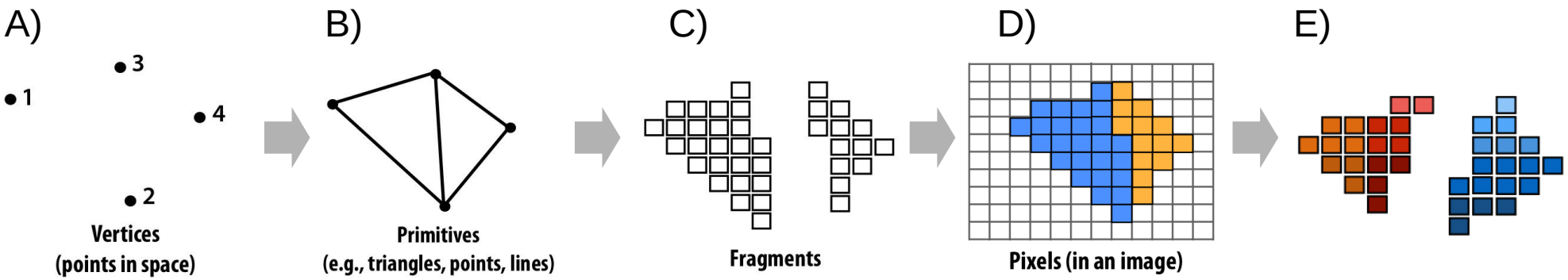
- преобразование систем координат;
- удаление невидимых поверхностей;
- отсечение невидимых областей;
- отрисовка базовых графических примитивов (точек, прямых, ломаных и т.п.);
- заливка / штриховка (растровая развертка сплошных областей);

Графический конвейер

- *Transform* – задание положения каждой вершины в сцене;
- *Clip* – отсечение срытых областей, в т.ч. за пределами области видимости;
- *Rasterize* – переход от векторного представления к пиксельному;
- *Shade* – вычисление цвета каждого пикселя;
- *Visibility/Blend* - рассчет наложений и цвета для прозрачных объектов.
- *Display* – окончательное формирование изображения в памяти.

Графический конвейер (пример)

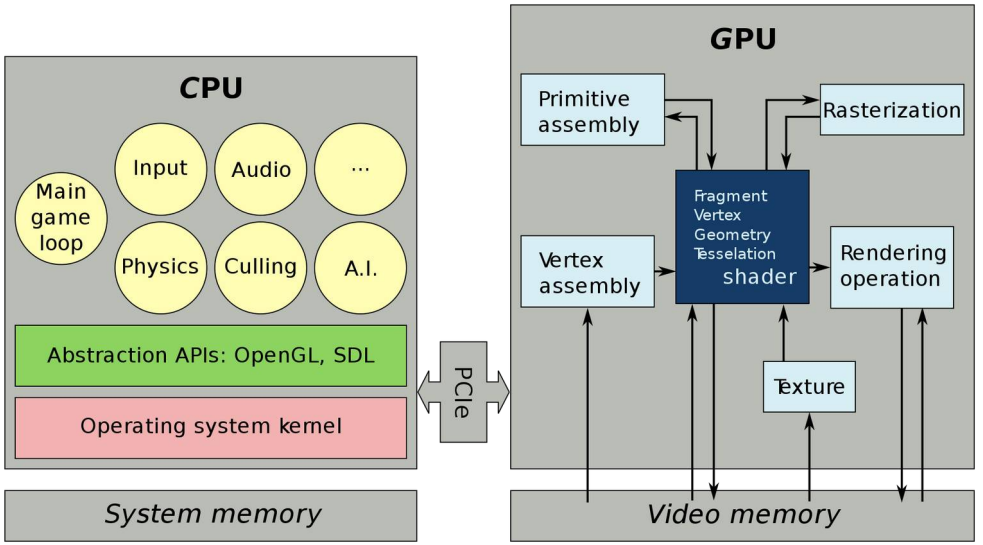




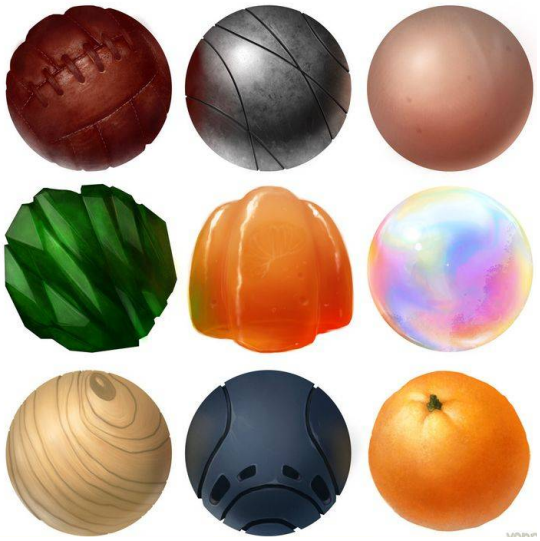
Шейдеры

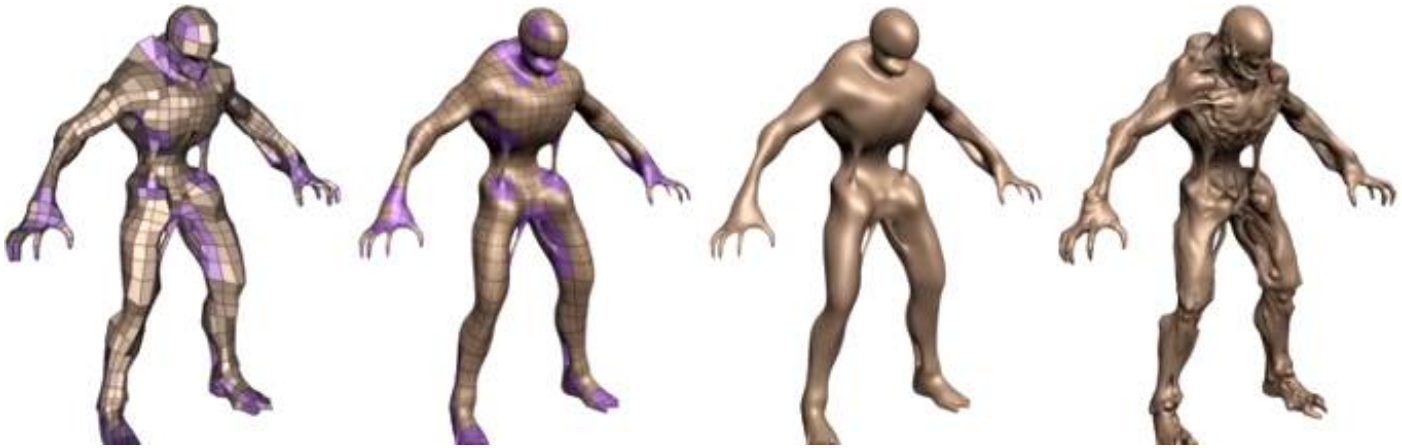
Шейдер – это программа, которая загружается в ускоритель, и конфигурирует его узлы для обработки соответствующих элементов. Шейдер позволяет снять ограничение на способ обработки эффектов.

- вершинные;
- геометрические;
- пиксельные или фрагментные.



Shader Forge это визуальный редактор шейдеров для Unity.





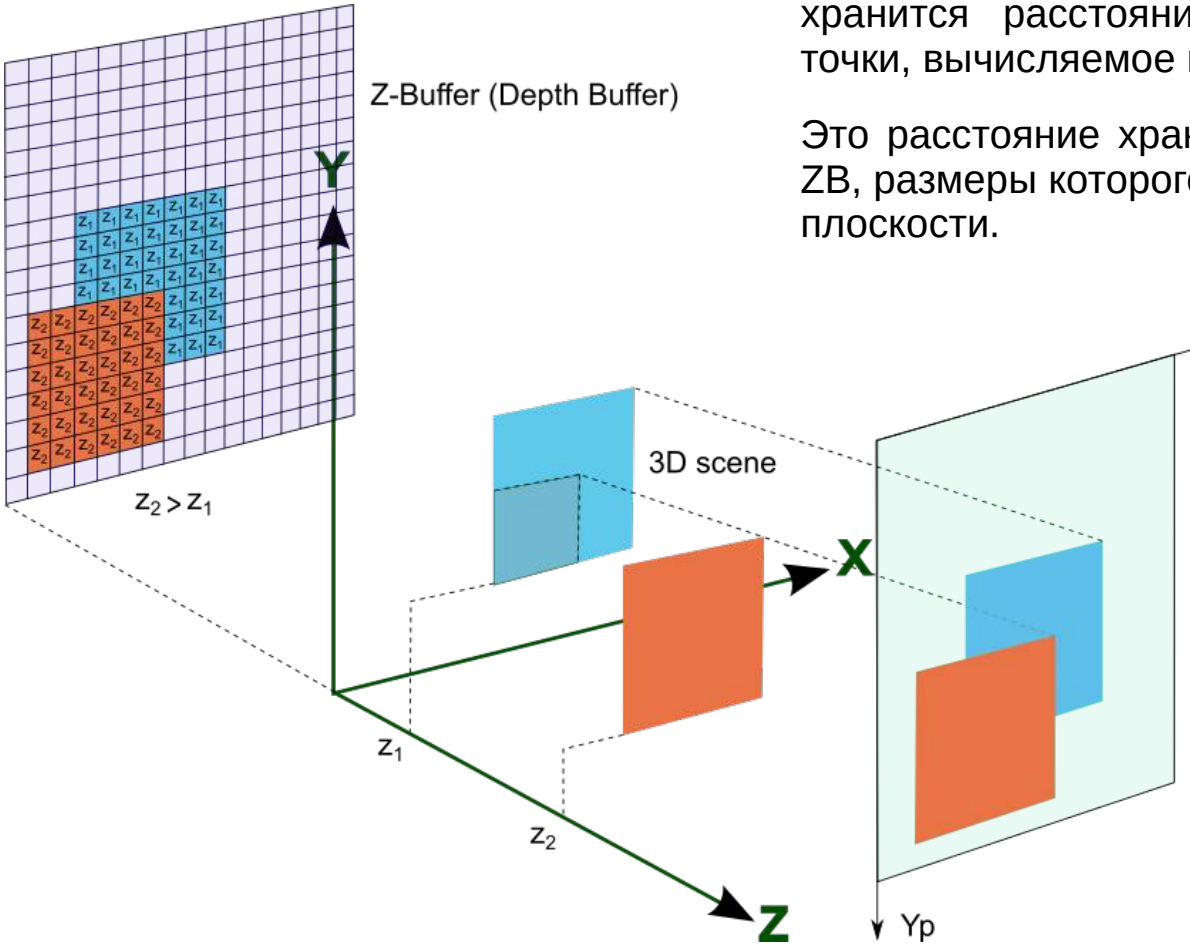
Тесселяция используется для увеличения геометрической сложности моделей, когда из низкополигональной получается более сложная.

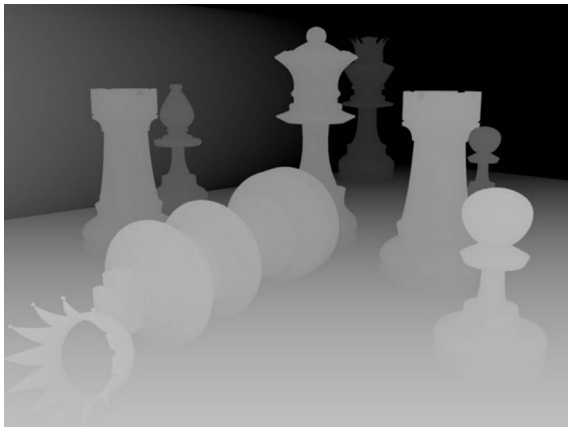
Z-буферизация

Z-буферизация —в компьютерной трёхмерной графике способ учёта удалённости элемента изображения.

При решении задачи загораживания методом Z-буфера при выводе текущего полигона формируется его растровое представление на картинной плоскости и для каждого пикселя картинной плоскости, кроме цвета, хранится расстояние проецируемой в этот пиксель точки, вычисляемое вдоль направления проецирования.

Это расстояние хранится в двумерном массиве глубин ZB, размеры которого совпадает в размерами картинной плоскости.



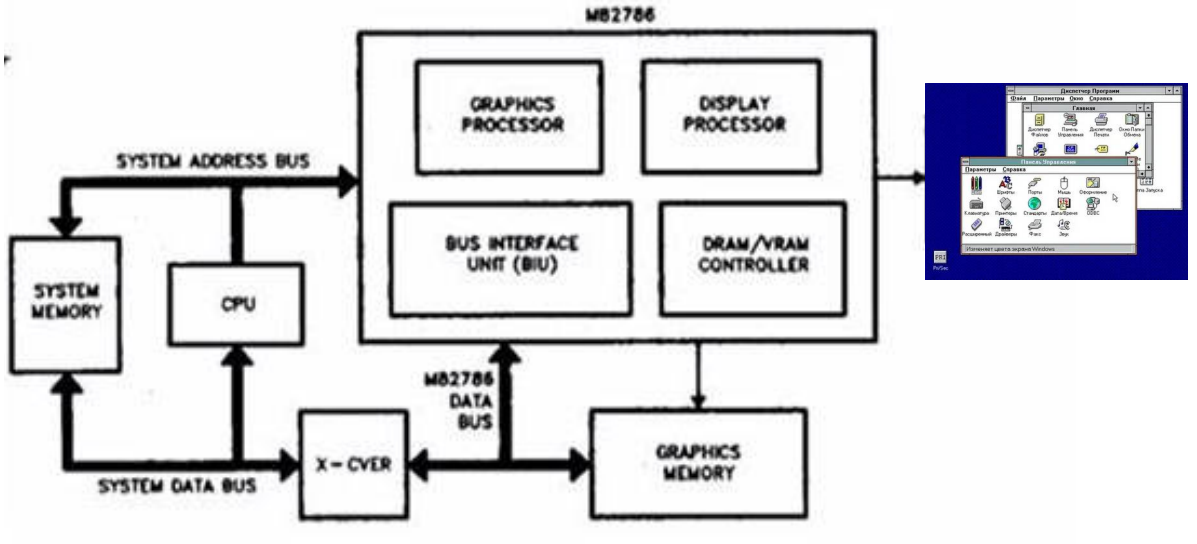


Различия в архитектурах GPU и CPU

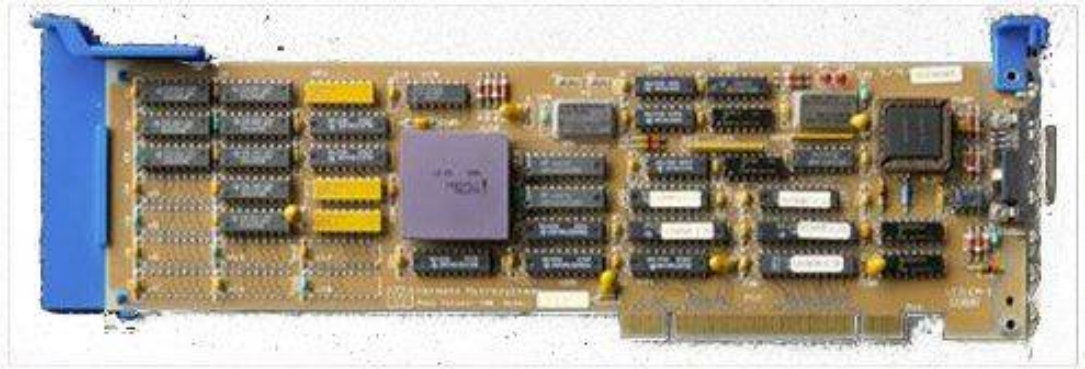
CPU	GPU
Ядра CPU проектируются для выполнения одного потока последовательных инструкций с максимальной производительностью	GPU предназначен для выполнения большого количества параллельных потоков команд.
В CPU доступ к памяти зависит от поступивших команд и часто происходит по случайным адресам. Требуется использовать кэш-память для усокрения доступа к памяти.	В GPU доступ к памяти преимущественно последовательный (пиксели и тексели читаются и пишутся последовательно). Большой кэш не требуется.
Доступ к памяти плохо распараллеливается, данные сосредоточены в сегментах и выборка осуществляется в небольшое количество модулей памяти (по крайней мере, для одной задачи.)	Доступ к памяти легко распараллеливается и пропускная сопособность каналов памяти велика.
В универсальных процессорах большую часть площади кристалла занимают различные блоки конвейера: декодеры, буферы, ROB, кэш-память и пр.	Аппаратная часть GPU оптимизирована под выполнение небольших и программ (шейдеров).
CPU хорошо справляется с зависимыми данными.	GPU предназначен для вычислений независимых данных (пикселей, текстур). При наличии зависимостей скорость вычислений существенно падает.

Первый дискретный графический сопроцессор Intel 82786

Графический процессор (GP) и дисплейный процессор (DP) были независимыми процессорами в 82786. Шинный интерфейсный блок (BIU) с контроллером DRAM / VRAM обрабатывал запросы шины между графическим процессором, процессором дисплея и внешним ЦП или шиной.



Графический процессор выполнял команды, размещенные в системной памяти и формирует изображения в битовых картах видеопамяти для дисплейного процессора во взаимодействии с контроллером видеопамяти и интерфейсным устройством шины.



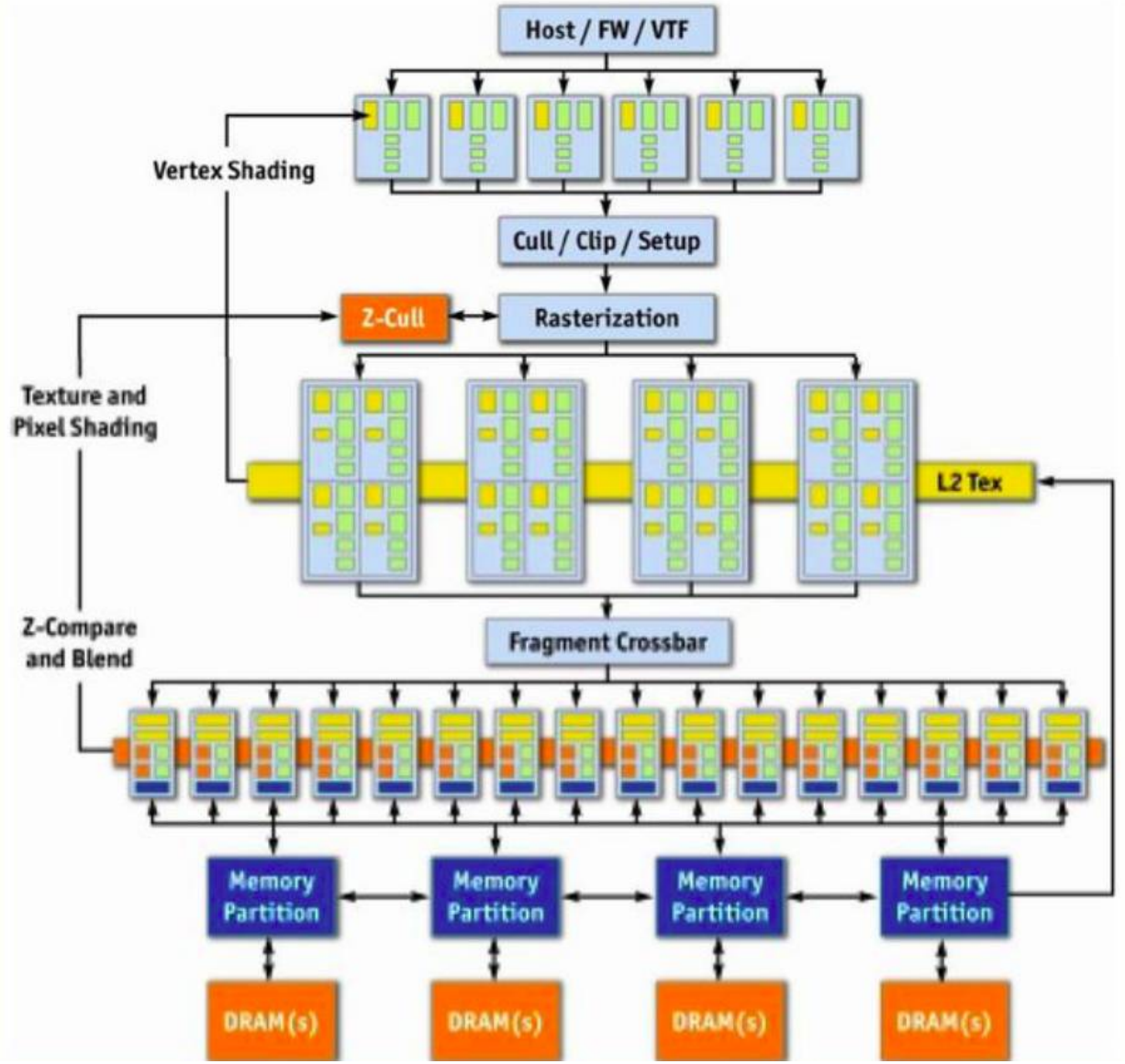
Дисплейный процессор преобразует битовые карты, создаваемые графическим процессором в растровые последовательности для видеоконтрольного устройства, которое отображает их в виде отдельных окон на экране графического монитора.

Современные графические процессоры GPU

Nvidia GForce6 (NV40), 2004

Этапы обработки:

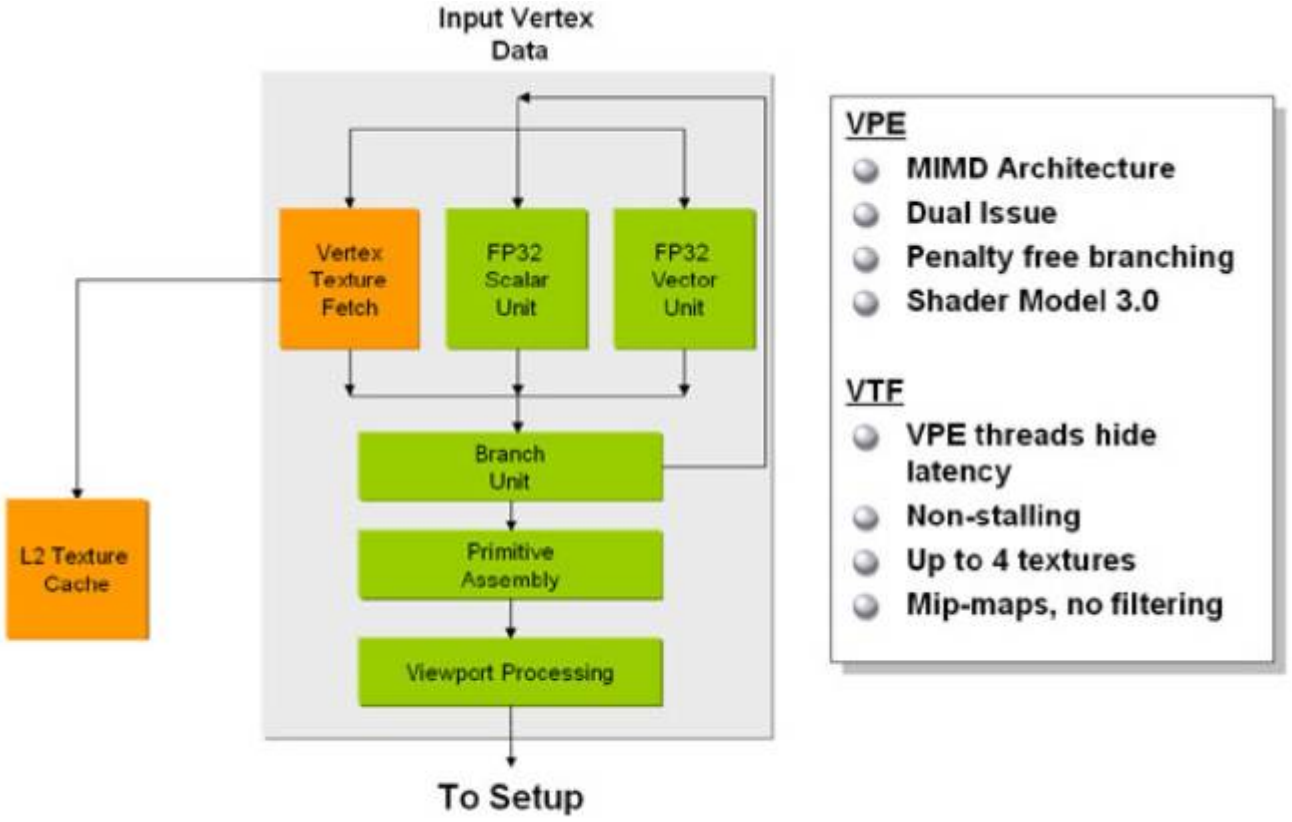
- Загрузка в чип ускорителя вершин из памяти акселератора
- Обработка в вершинных процессорах.
- Установка треугольников, отсечение невидимых поверхностей.
- Треугольники растеризуются и производится отсечение невидимых частей с использование Z-буфера (Hidden Surface Removal, HSR).
- Формируются квады 2x2 пикселя, подлежащие закраске. На квады накладываются текстурные фрагменты и производится интерполяция.
- Закраска фрагментов.
- Производится блендинг (смещение)
- Значения квадов записываются в буфер кадра
- Вывод изображения на экран.



Современные графические процессоры GPU

Nvidia GForce6

Вершинные процессоры

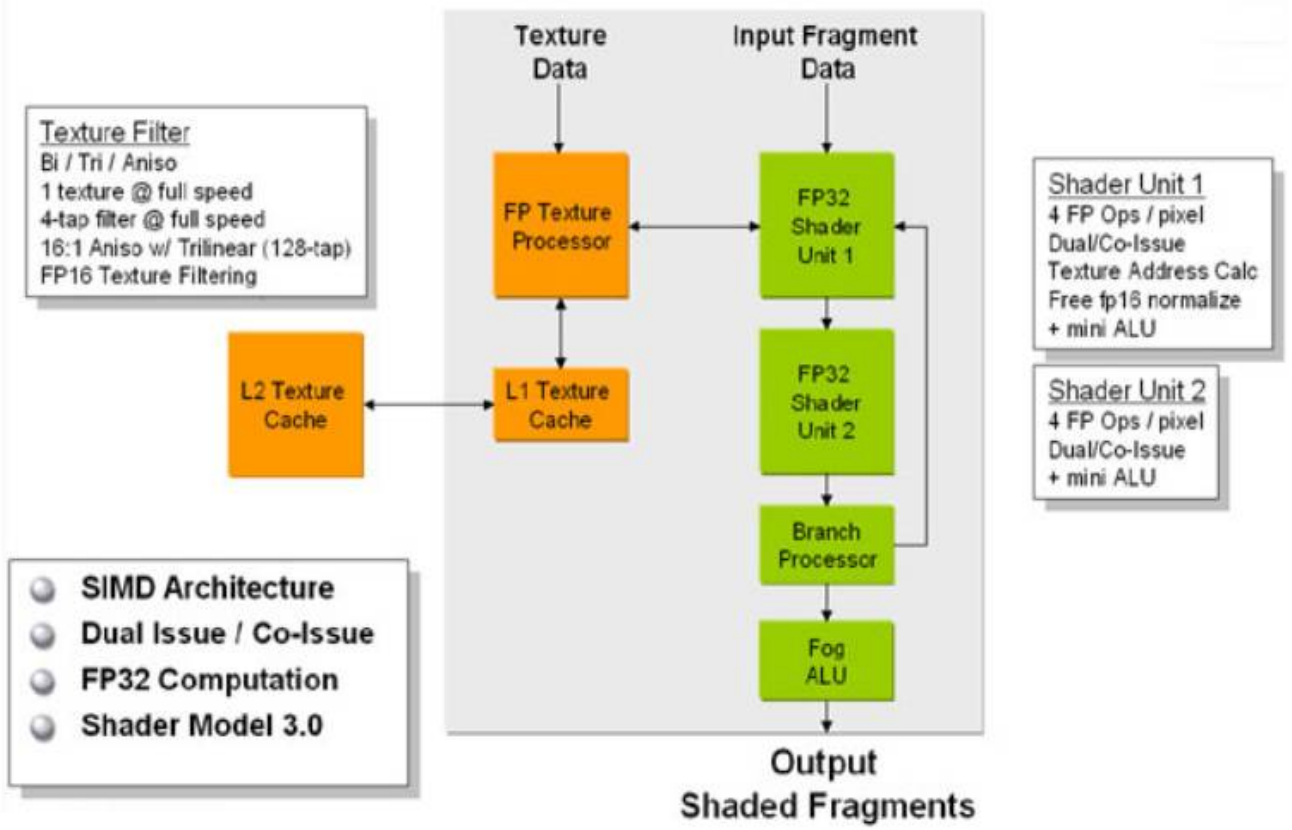


Благодаря расширению динамического выполнения (больше вариантов по циклам/ветвлениям и новые функции подпрограмм), можно создавать более эффективный код и реализовывать новые возможности для эффектов.

- полная поддержка вершинных программ 3.0;
- 216 (65,535) длинных вершинных программ;
- вершинная обработка с учётом текстуры - карты смещения (displacement mapping);
- динамический контроль выполнения (flow control) - циклы и ветвления, вызов подпрограмм и возврат;
- Geometry Instancing (vertex stream divider).

Современные графические процессоры GPU

Nvidia GForce6
Пиксельные конвейеры

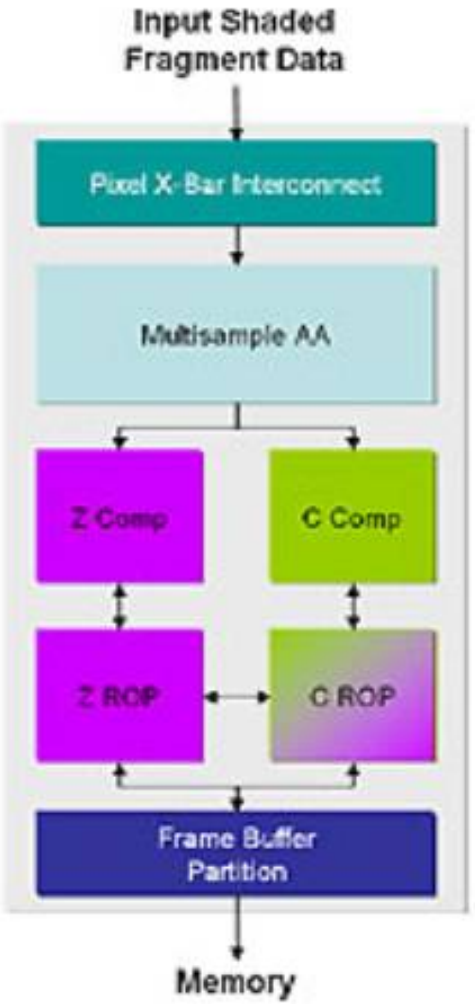


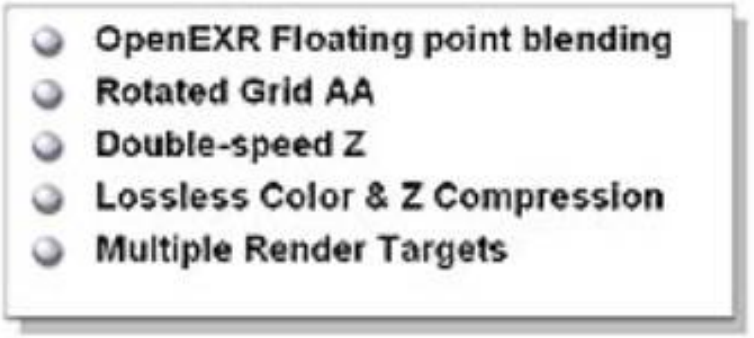
- Каждый из 16 пиксельных конвейеров имеет два блока пиксельных программ (суперскалярный дизайн) и один текстурный процессор с плавающей запятой.
- NV40 также оснащён четырьмя кэшами текстур L1, каждый из которых обслуживает четыре конвейера.
- Разгрузить интерфейс памяти помогает и массивный кэш L2.
- Архитектура блоков пиксельных программ-шейдеров имеет настоящий дизайн SIMD (одна инструкция - много данных).
- Если первый блок пиксельных программ на каждом конвейере может выполнять как арифметические операции, так и чтение текстур и нормализацию, то второй блок ограничен только арифметикой.
- Если блок не занимается текстурированием, то он может выполнять (в данный проход) пиксельное затенение. Блок 2 всегда доступен для пиксельного затенения.

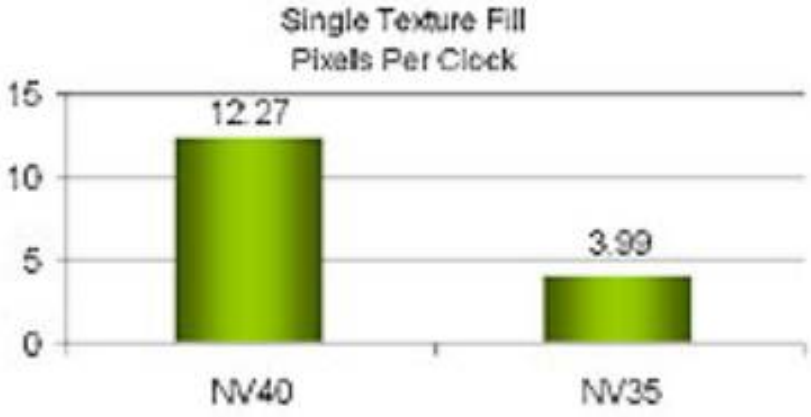
Современные графические процессоры GPU

Nvidia GForce6

Пиксельные конвейеры ROP (растровые операции)







Подсистема ROP карты GeForce 6800 имеет следующие характеристики:

- 16 пикселей за такт, цвет и Z;
- 32 пикселей за такт, только Z;
- 64-bit FP Frame Buffer Blending;
- цветовое и Z-сжатие без потерь;
- качественное сглаживание - поворот сетки (Rotated Grid);
- полная поддержка MRT;
- ускоренный рендеринг теней.

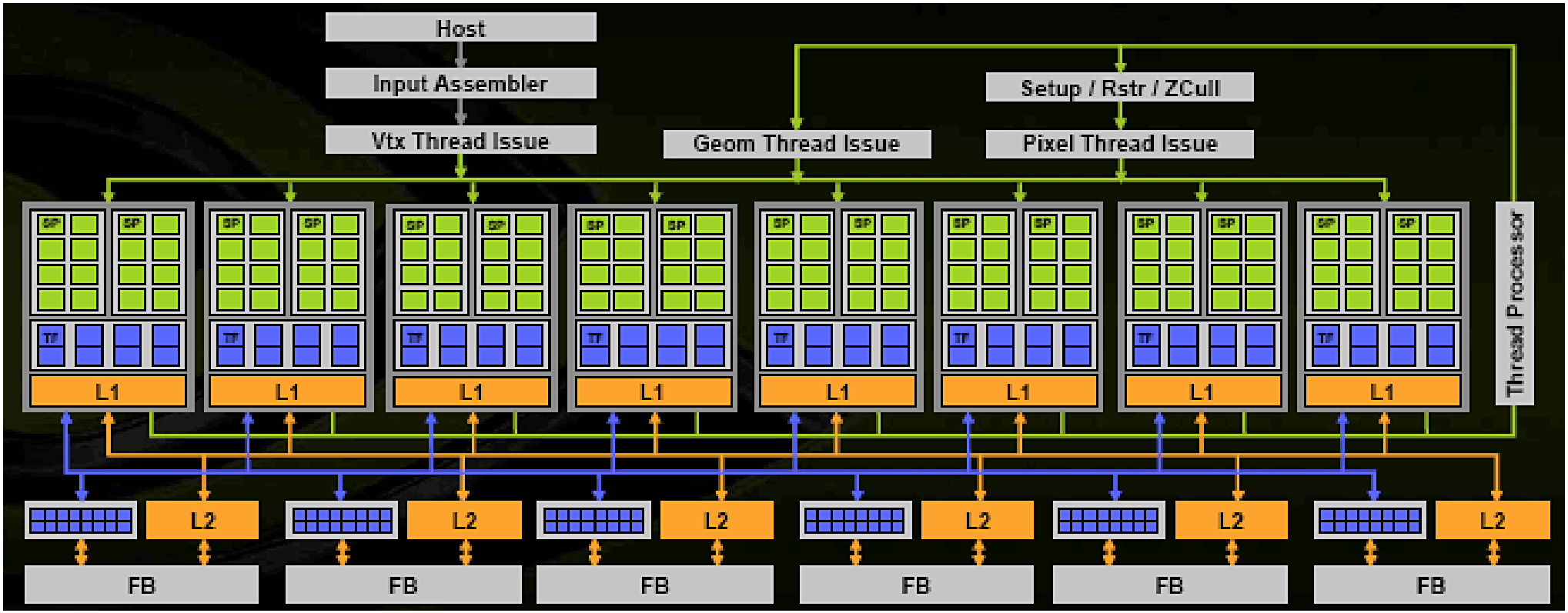
Современные графические процессоры GPGPU

NVidia G80

Чип состоит из 8 универсальных вычислительных блоков (шейдерных процессоров) по 4 TMU и 16 ALU в каждом. Всего, таким образом, имеется 128 ALU (называются потоковыми процессорами (SM, Stream Processors) и 32 TMU. Все ветвления, переходы, условия и т.д. применяются целиком к одному блоку и таким образом логичнее всего, его и называть шейдерным процессором, пускай и очень широким.

Каждый такой процессор снабжен собственным кэшем первого уровня, в котором теперь хранятся не только текстуры, но и другие данные, которые могут быть запрошены шейдерным процессором.

Кроме управляющего блока и 8 вычислительных шейдерных процессоров в наличии 6 блоков ROP, исполняющих определение видимости, запись в буфер кадра и MSAA (синие, рядом с блоками кэша L2) сгруппированные с контроллерами памяти, очередями записи и кэшем второго уровня.



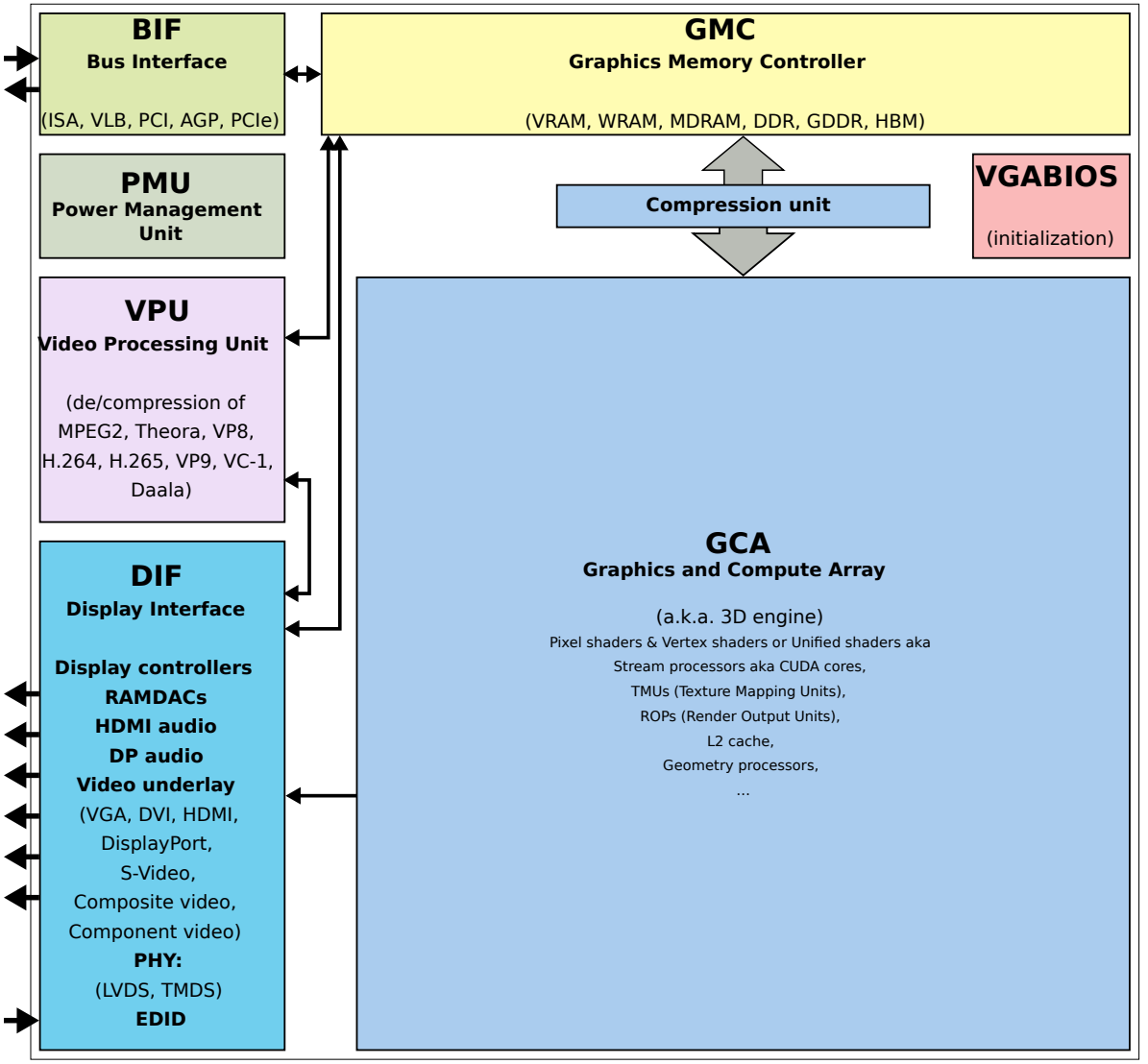
Обобщенная схема GPGPU

Отличительными особенностями по сравнению с ЦП являются:

- архитектура, максимально нацеленная на увеличение скорости расчёта текстур и сложных графических объектов;
- ограниченный набор команд.
-

Высокая вычислительная мощность GPU объясняется особенностями архитектуры. Современные CPU содержат несколько ядер, тогда как графический процессор изначально создавался как многопоточная структура с множеством ядер. Разница в архитектуре обусловливает и разницу в принципах работы. Если архитектура CPU предполагает последовательную обработку информации, то GPU исторически предназначался для обработки компьютерной графики, поэтому рассчитан на массивно параллельные вычисления.

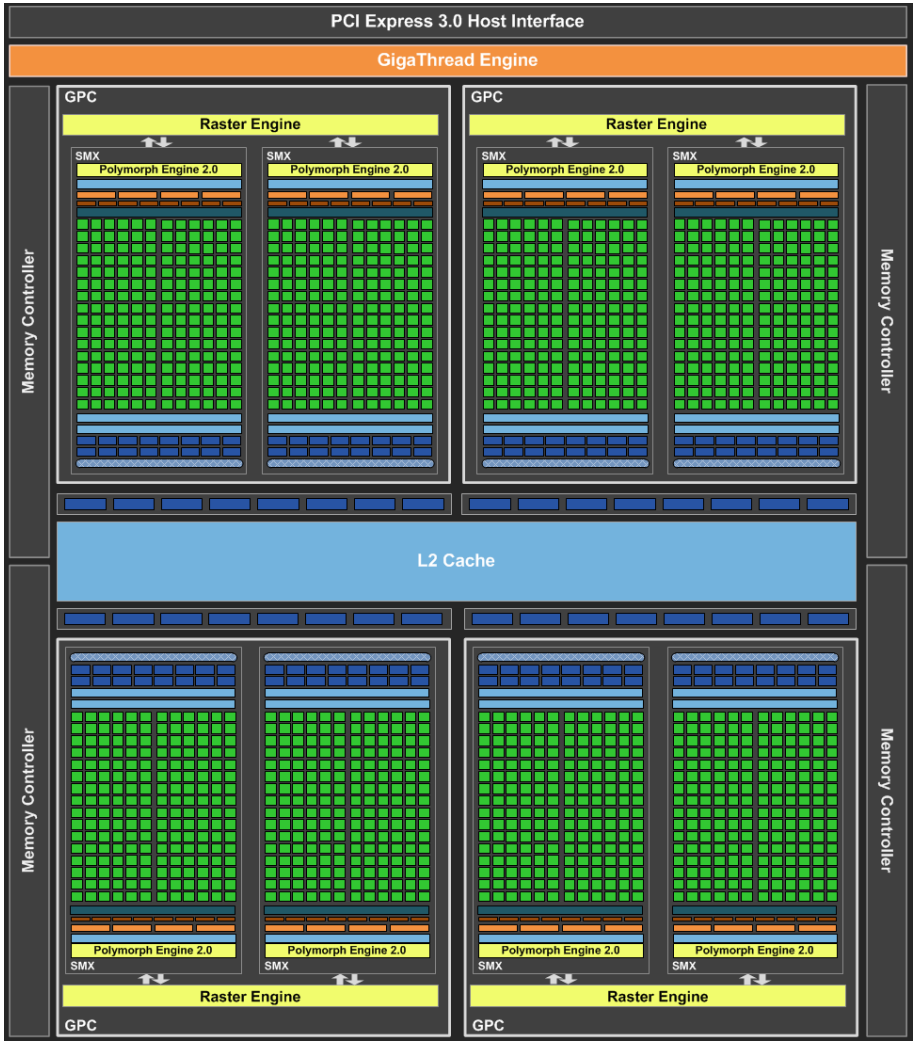
Каждая из этих двух архитектур имеет свои достоинства. CPU лучше работает с последовательными задачами. При большом объёме обрабатываемой информации очевидное преимущество имеет GPU. Условие только одно — в задаче должен наблюдаться параллелизм.



Современные графические процессоры GPU

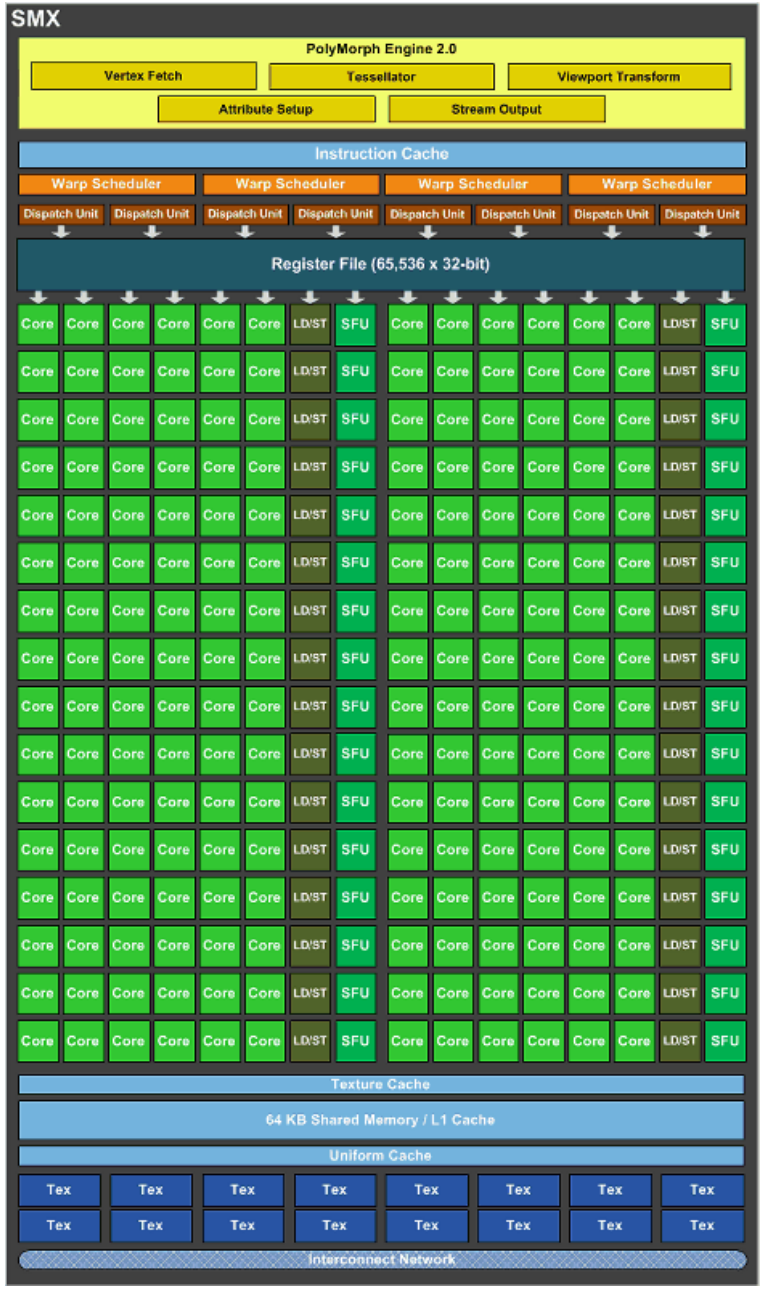
NVidia GeForce GTX 600

- Технология производства 28 нм;
- 3.54 миллиардов транзисторов;
- Площадь ядра 294 мм²;
- Унифицированная архитектура с массивом процессоров для потоковой обработки различных видов данных: вершин, пикселей и др.;
- Аппаратная поддержка DirectX 11 API, в том числе шейдерной модели Shader Model 5.0, геометрических и вычислительных шейдеров, а также тесселяции;
- 256-битная шина памяти, четыре независимых контроллера шириной по 64 бита каждый, с поддержкой GDDR5 памяти;
- Базовая частота ядра 1006 МГц;
- Средняя турбо-частота ядра 1058 МГц;
- 8 потоковых мультипроцессоров, включающих 1536 скалярных ALU для расчётов с плавающей запятой (поддержка вычислений в целочисленном формате, с плавающей запятой, с FP32 и FP64 точностью в рамках стандарта IEEE 754-2008);
- 128 блоков текстурной адресации и фильтрации с поддержкой FP16 и FP32 компонент в текстурах и поддержкой трилинейной и анизотропной фильтрации для всех текстурных форматов;
- 4 широких блока ROP (32 пикселя) с поддержкой режимов антиалиасинга до 32 выборок на пиксель, в том числе при FP16 или FP32 формате буфера кадра. Каждый блок состоит из массива конфигурируемых ALU и отвечает за генерацию и сравнение Z, MSAA, блендинг;
- Интегрированная поддержка RAMDAC, двух портов Dual Link DVI, а также HDMI и DisplayPort.
- Интегрированная поддержка четырёх мониторов, включая два порта Dual Link DVI, а также HDMI 1.4a и DisplayPort 1.2
- Поддержка шины PCI Express 3.0



Современные графические процессоры GPU

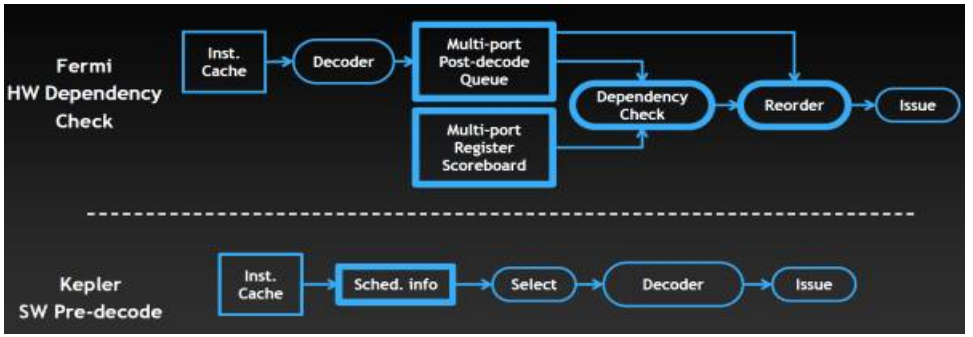
NVidia GeForce GTX 600



Мультипроцессоры — это основная составная часть GPU компании NVIDIA. По сравнению с предыдущими SM, новые SMX обеспечивают более высокую производительность, что видно по количеству функциональных устройств в составе SMX, но при этом потребляют значительно меньше энергии. А уменьшенное количество мультипроцессоров на GPU (8 в отличие от 16 в GF100/GF110) было продиктовано установленными рамками по площади ядра.

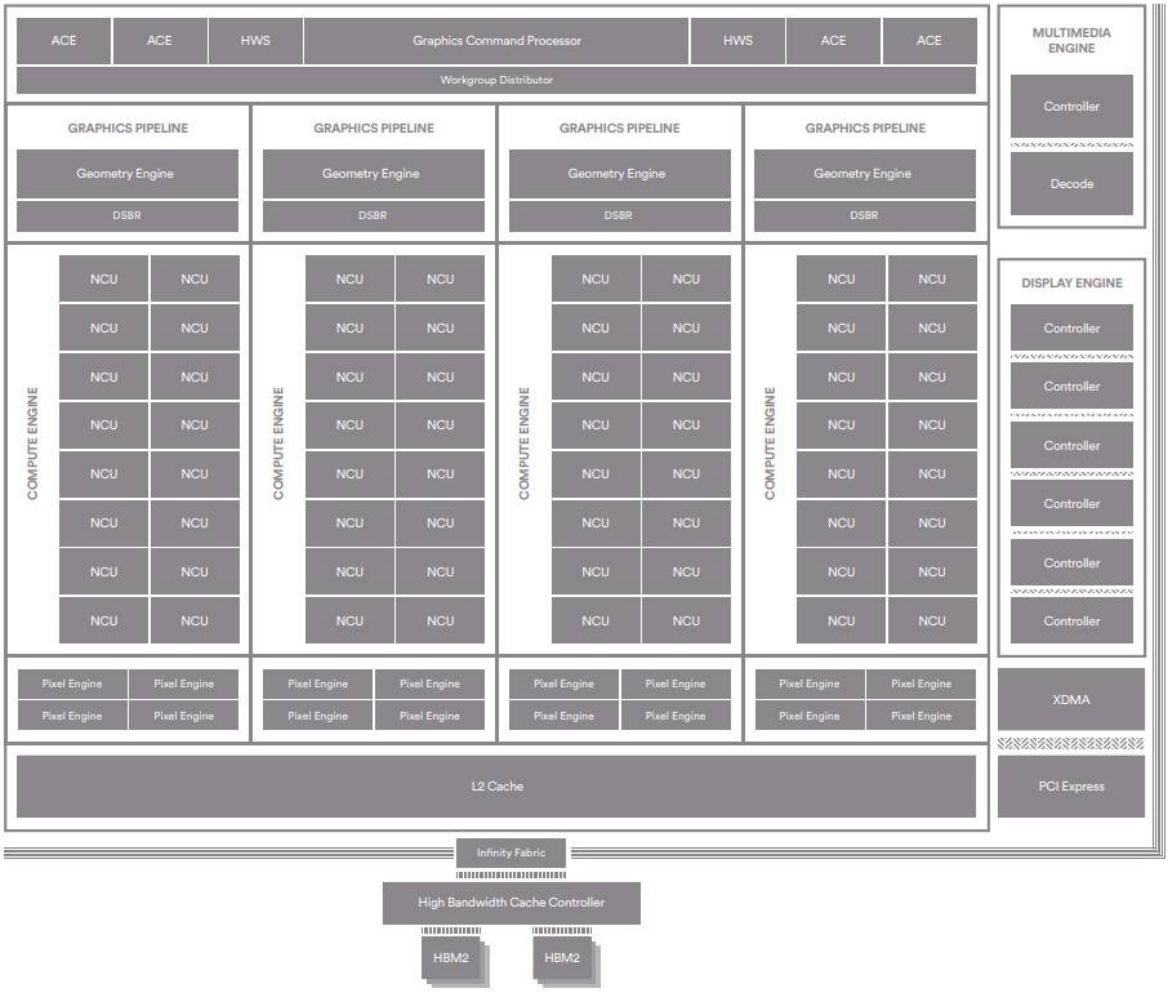
Большая часть ключевых блоков GPU включена в состав SMX: потоковые процессоры (CUDA Cores) выполняют все математические операции над пикселями, вершинами и занимаются неграфическими вычислениями, текстурные модули (TMU) фильтруют текстурные данные, загружают и записывают их из/в видеопамять, блоки специальных функций (Special Function Units, SFU) выполняют сложные операции (вычисление синуса, косинуса, квадратного корня и т.п.) и интерполяции графических атрибутов. Ну а движок PolyMorph обеспечивает выборку вершин, занимается тесселяцией, преобразованием в экранные координаты, установкой атрибутов и потоковым выводом (stream output).

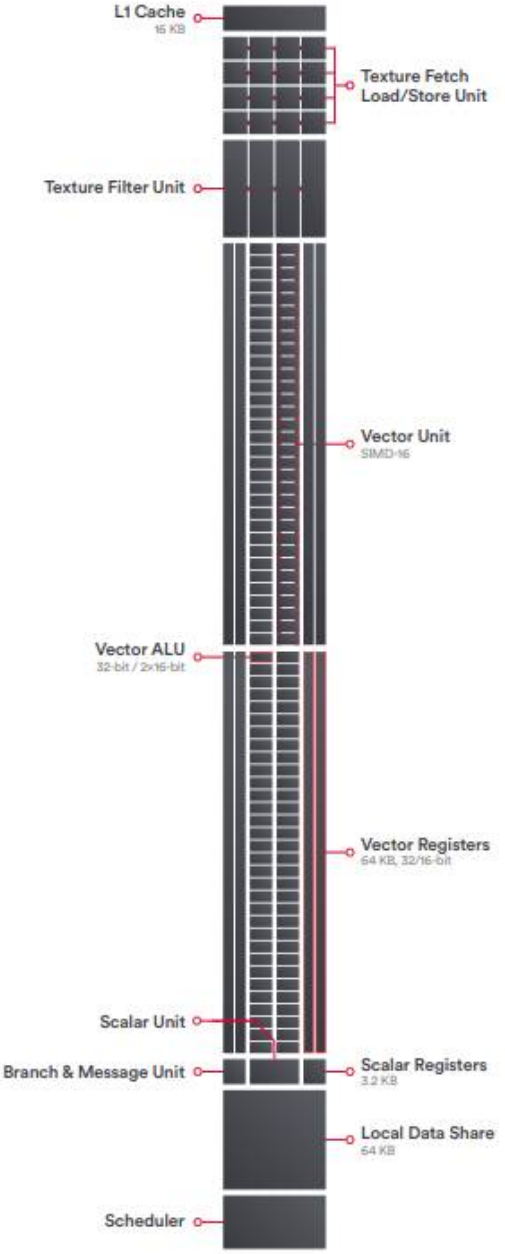
Процессор содержит сложную аппаратную стадию, служащую для предотвращения конфликтов доступа к данным. Специальная таблица регистров (multi-port register scoreboard) отслеживает регистры, данные в которых ещё не готовы, а блок проверки зависимостей (dependency check) анализирует их использование, проверяя зависимости команд.



Современные графические процессоры GPU

AMD Radeon VII





Количество универсальных процессоров – 3840; Количество текстурных блоков – 240; Количество блоков блендинга – 64; Эффективная частота памяти - 2000 Мгц; Тип памяти – HBM2; Шина памяти – 4096-бит; Объем памяти - 16 ГБ; Пропускная способность памяти - 1 ТБ/с; Количество транзисторов – 13.2 млрд.

Полезные ссылки

1) DX Current: Настоящее аппаратного ускорения графики: https://www.ixbt.com/video2/dx-current.shtml
2) Современная терминология 3D графики: https://www.ixbt.com/video2/terms2k5.shtml#dm