

3 Технология GPGPU

General-Purpose computation on Graphics Processing Units - высокопроизводительные общие вычисления на GPU.

Graphics Processing Unit (GPU) - высокопроизводительный многопроцессорный блок, использующийся в графическом конвейере видеокарты для ускорения графических операций.

Два основных производителя видеокарт, NVIDIA и AMD, разработали и анонсировали соответствующие платформы под названием **CUDA (Compute Unified Device Architecture)** и **CTM (Close To Metal или AMD Stream Computing)**

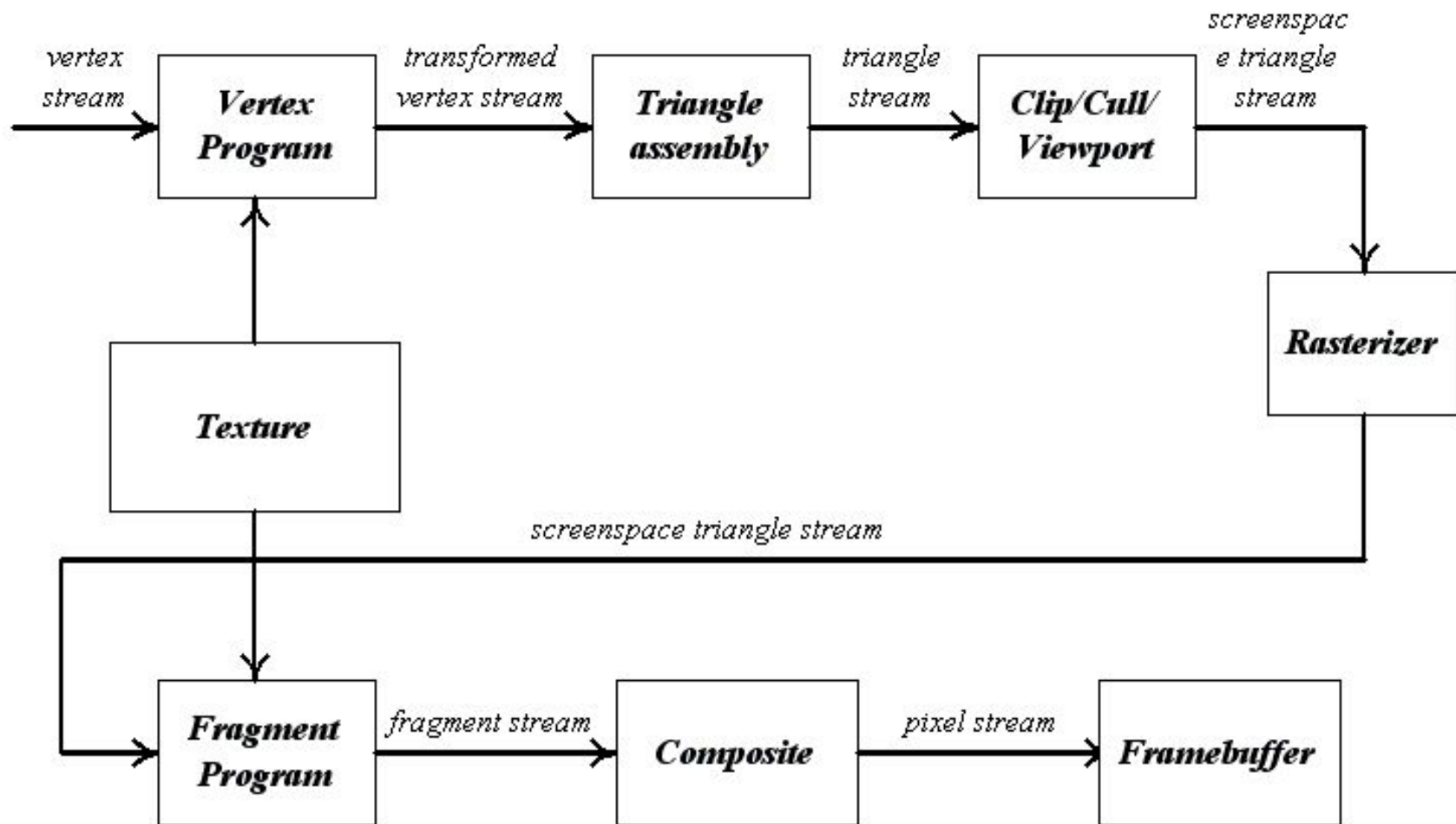
3.1 Графический конвейер

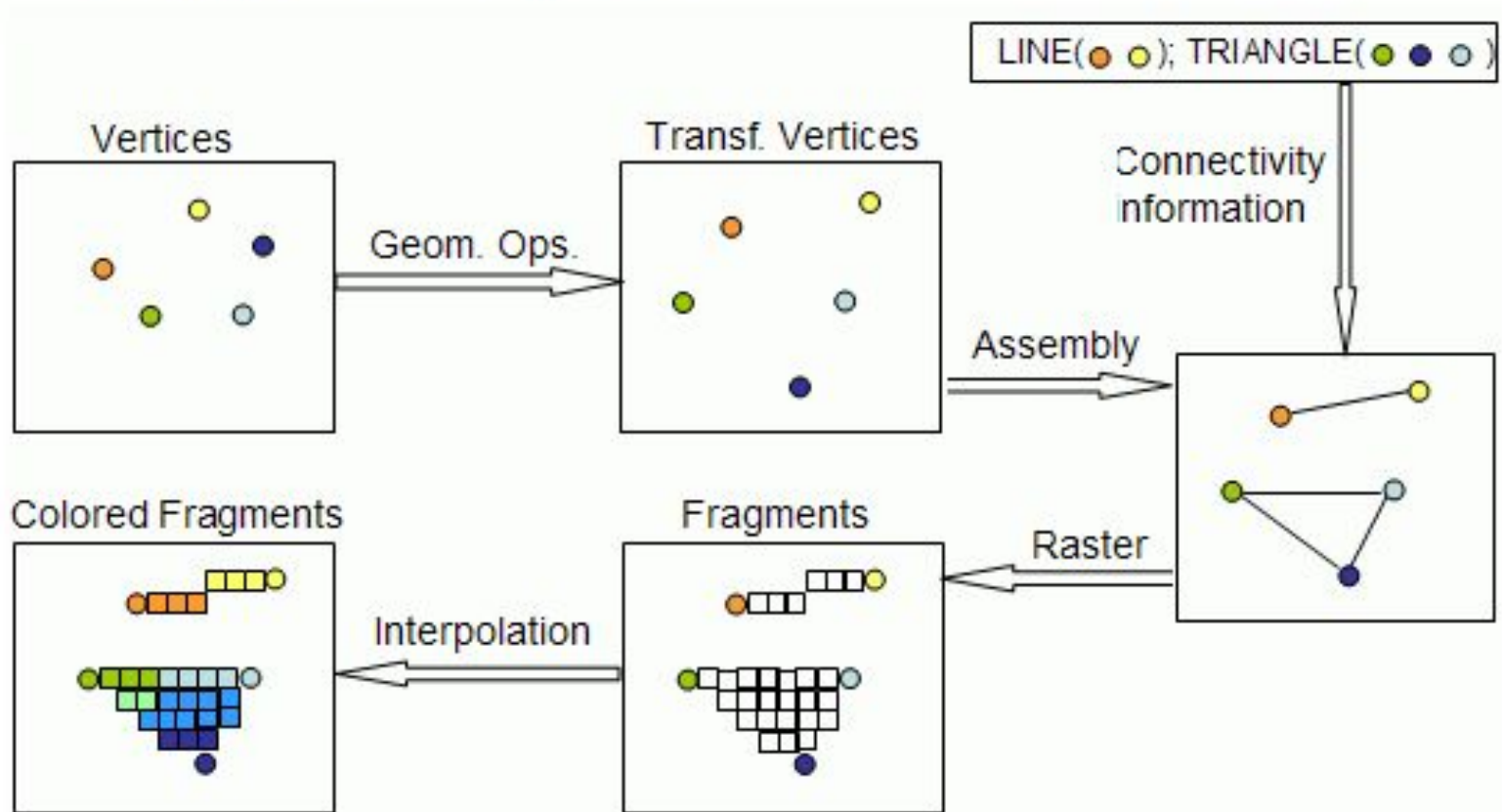
1. *Построение геометрической модели*
2. *Деление поверхности объекта на плоские простейшие элементы (тесселяция, tessellation)*
3. *Трансформация*
4. *Освещенность (lighting) и затенение (shading)*
5. *Проецирование*
6. *Обработка координат вершин*

7. *Удаление скрытых поверхностей*
8. *Наложение текстур*
9. *Эффекты прозрачности и полупрозрачности*
10. *Коррекция дефектов*
11. *Интерполяция цветов (disering)*

Выходная информация 3D-конвейера - это комплекс рассчитанных данных о каждом пикселе, которые помещаются в видеопамять.

Схема графического конвейера





3.2 Сравнение архитектуры центрального процессора и графического процессора



Пример: Характеристики Nvidia GeForce GTX 590

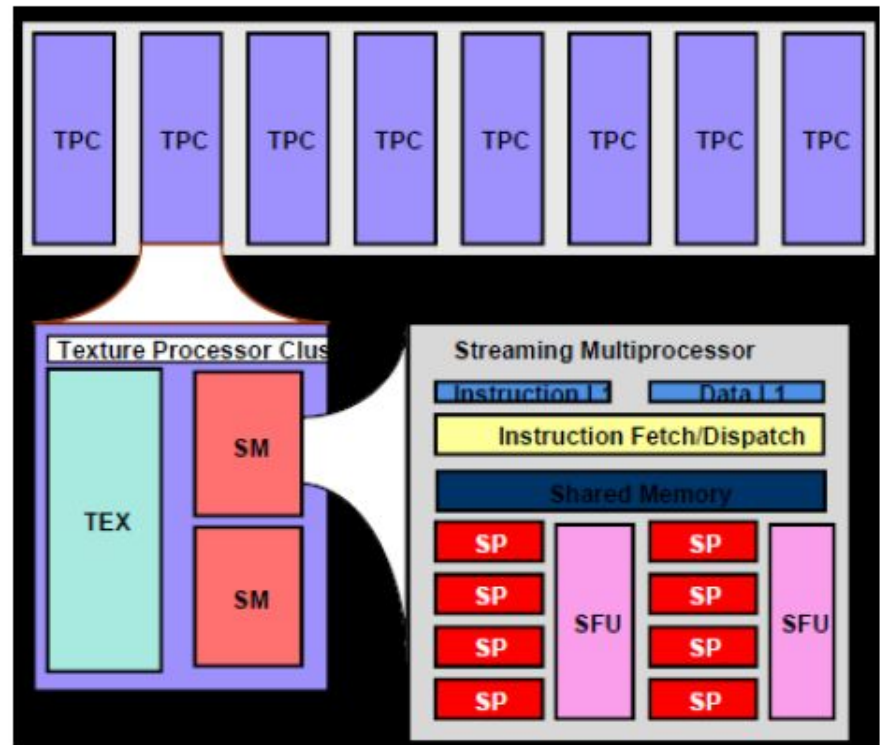
- Технология производства 40 нм;
- 2 чипа по 3 миллиарда транзисторов каждый;
- Унифицированная архитектура с массивом процессоров для потоковой обработки различных видов данных: вершин, пикселей и др.;
- Аппаратная поддержка DirectX 11;
- Двойная 384-битная шина памяти: дважды по шесть независимых контроллеров шириной по 64 бита каждый, с поддержкой GDDR5 памяти;
- Частота ядра 607 МГц;
- Удвоенная частота АЛУ 1215 МГц;
- 2×16 потоковых мультипроцессоров, включающих в общем 1024 скалярных АЛУ для расчётов с плавающей точкой (целочисленные и плавающие форматы, поддержка одинарной и двойной точности в рамках стандарта IEEE 754-2008);
- 2×64 блока текстурной адресации и фильтрации с поддержкой FP16- и FP32-компонент в текстурах и поддержкой трилинейной и анизотропной фильтрации для всех текстурных форматов;
- 2×6 широких блоков ROP (всего 96 пикселей) 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, MSAА, блендинг;
- Для каждого GPU интегрированная поддержка RAMDAC, двух портов Dual Link DVI, а также HDMI и DisplayPort.



Аппаратная архитектура GPU

GPU - набор мультипроцессоров

- ❑ SPA - Streaming Processor Array
- ❑ TPC - Texture Processor Cluster
- ❑ SM - Streaming Multiprocessor
 - ❑ Multi-threaded processor core
 - ❑ Fundamental processing unit for CUDA thread block
- ❑ SP - Streaming Processor
 - ❑ Scalar ALU for a single CUDA thread



Аппаратная реализация :

Модель исполнения

□ Каждый блок потоков состоит из warp'ов

- Warp - SIMD-группа потоков фиксированного размера, состоящая из скалярных потоков с последовательными координатами.

□ Блок потоков всегда выполняется только на одном мультипроцессоре

- Разделяемые переменные блока потоков хранятся в on-chip модуле разделяемой памяти мультипроцессора
- Файл локальных регистров делится между всеми потоками, обрабатываемыми мультипроцессором
 - ✓ Слишком большого размер блока потоков для ядра, использующего слишком много локальных регистров, вызывает сбой этапа исполнения

□ Мультипроцессор может обрабатывать несколько блоков потоков одновременно

- Локальный регистровый файл и разделяемая память делятся между всеми потоками / блоками потоков, работающими одновременно
- Тем самым, с уменьшением числа используемых локальных регистров (на поток) и размера используемой разделяемой памяти (на блок) увеличивается число потоков / блоков потоков, способных одновременно находиться в обработке
- Не влияет на логику работы, только на производительность
 - ✓ На логическом уровне блоки потоков всегда изолированы