

Тема 8. Графическая подсистема ПЭВМ, дисплейные устройства (мониторы) и проекторы, интерфейсы подключения дисплейных устройств

Лекция 13. Графическая подсистема

Конструкция и принцип действия графической карты.
Создание графического объекта. Этапы рендеринга. Шейдеры.
Потоковый процессор (на примере NVIDIA GeForce 8800).
Интегрированные графические устройства.
Встроенная графика. Графическое ядро, встроенное в процессор.

Графическая подсистема ПК

Графическая подсистема изначально входила в архитектуру ПК в виде отдельной платы расширения.

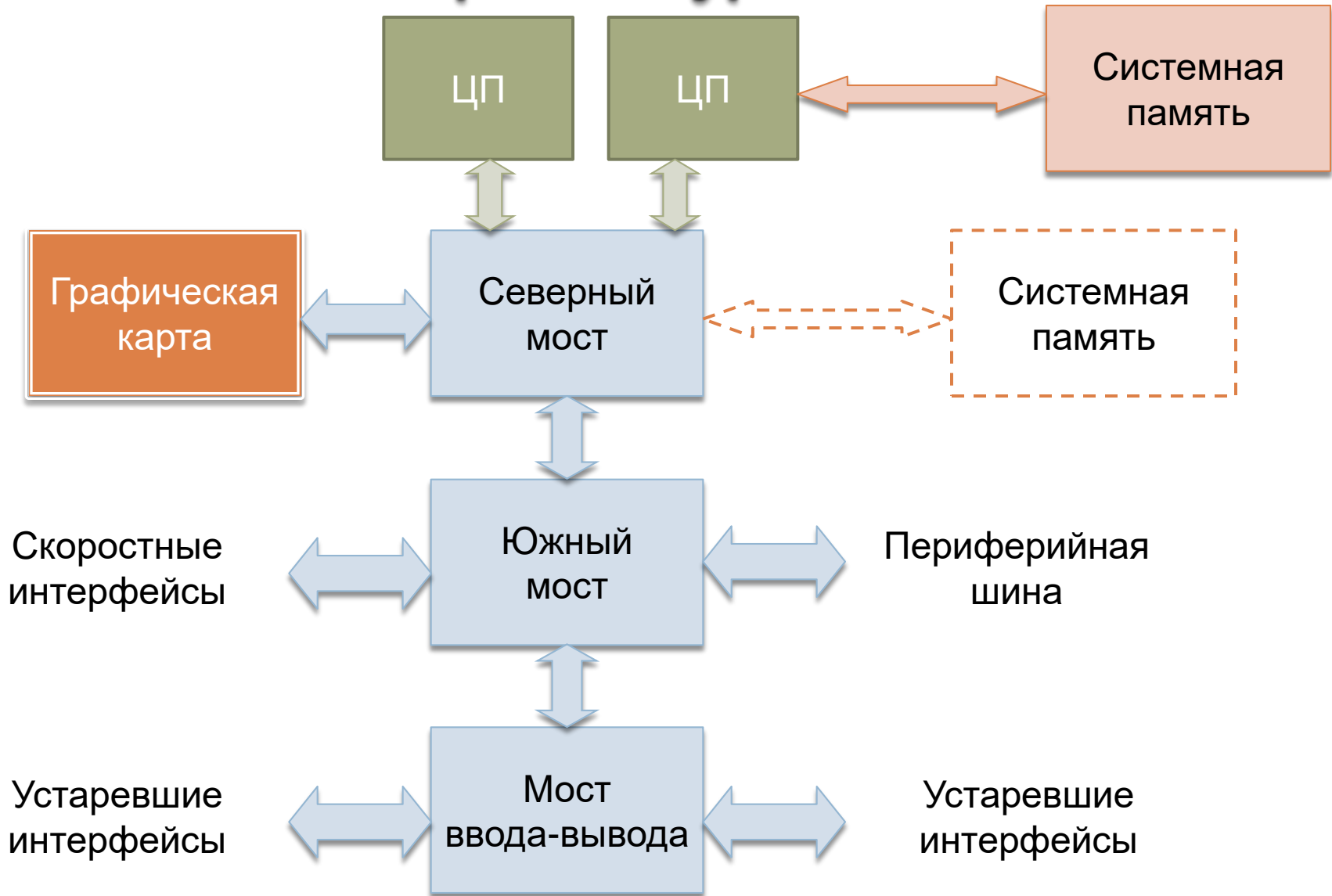
Впоследствии графическую подсистему удалось **интегрировать в состав микросхем системной логики**. Встроенная графическая карта обычно подключается к контроллеру памяти по внутреннему интерфейсу, но может быть реализована та или иная шина «на кристалле».

В связи с развитием технологий 3D-графики, понадобилось перенести это устройство на отдельную высокоскоростную шину, которая смогла бы обеспечить требуемую ширину канала между графическим процессором и системной памятью.

Для задач, требовательных к быстродействию в 3D и видео, предлагаются **отдельные карты расширения**. Более того, выпускаются «двойные» карты, реализованы возможности объединения карт в единый конвейер и поочередного использования двух карт. **Внешние видеокарты**.

Изначально к ПК можно было подключать одно устройство отображения. Возможность нарастить количество дисплеев была нестандартной, но допустимой. Сегодня графическая подсистема поддерживает до двух независимых дисплеев на один адаптер.

Место графической подсистемы в архитектуре ПК



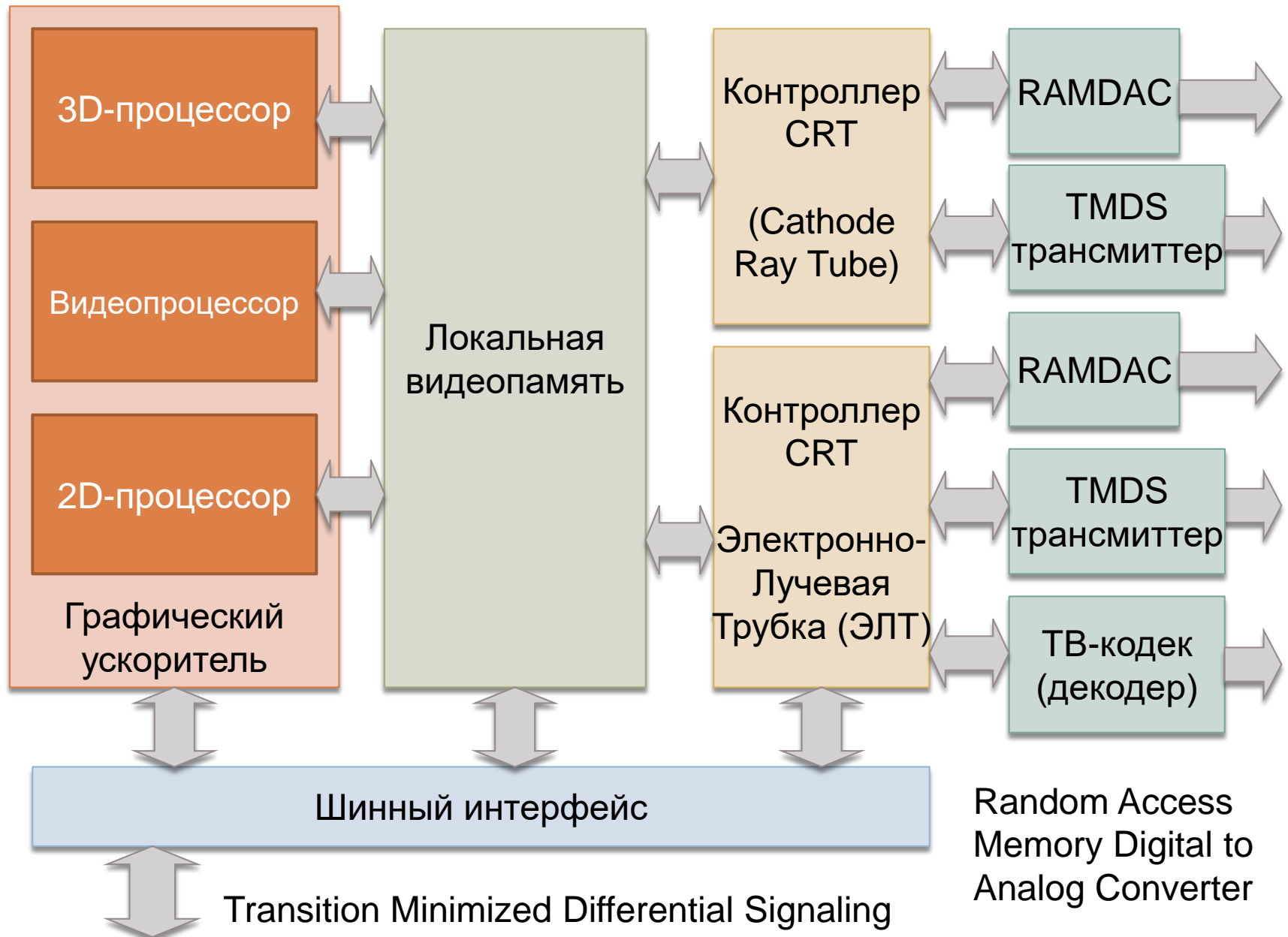
Текстовый режим

На заре появления ПК подсистема вывода изображения не была полностью графической. Растровое изображение в видеопамати строилось с помощью знакогенератора и Video BIOS. Таблицы кодировок хранились в локальной памяти или в ПЗУ видео-адаптера. Вывод текста выполнялся с помощью вызовов Video BIOS, которые обращались к таблицам знакогенератора и регистрам курсора.

Сегодня такой механизм вывода текста используется на этапе POST (Power-On Self-Test — самотестирование после включения) и некоторых ОС при выводе консоли. Обращение к Video BIOS производится только на этапе инициализации и при конфигурировании.

Вывод текста в графическом режиме, как правило, выполняется средствами 2D-процессора.

Архитектура графической карты



Графический процессор

(Graphics processing unit — графическое процессорное устройство) — занимается расчётами выводимого изображения, освобождая от этой обязанности центральный процессор, производит расчёты для обработки команд трёхмерной графики. Является основой графической платы, именно от него зависят быстродействие и возможности всего устройства.

Сегодня различают три независимых компонента графического процессора:

- 3D-процессор – наиболее сложная часть видеокарты, отвечает за 3D-рендеринг изображений. Сегодня он представляет собой связку блоков фиксированной обработки и универсальных ALU.
- 2D-процессор – наиболее простая и практически не развивающаяся часть, постепенно ее функции берет на себя 3D-процессор.
- Видеопроцессор – обработка видеоданных различного формата, формирование оверлея. Для декомпрессии видеоданных и дополнительной фильтрации может использовать ALU (собственные или 3D-процессора).

Видеопамять

Выполняет роль кадрового буфера, в котором хранится изображение, генерируемое и постоянно изменяемое графическим процессором и выводимое на экран монитора (или нескольких мониторов).

В видеопамяти хранятся также промежуточные невидимые на экране элементы изображения и другие данные.

Видеопамять бывает нескольких типов, различающихся по скорости доступа и рабочей частоте.

Современные видеокарты комплектуются памятью типа DDR, GDDR2 (двойная скорость передачи графических данных, Graphics Double Data Rate), GDDR3, GDDR4 и GDDR5 (5 ГГц, при использовании 256-битного интерфейса, GDDR5 позволяет передавать данные со скоростью 120 ГБ/с). Помимо видеопамяти, современные графические процессоры обычно используют в своей работе часть общей системной памяти компьютера, прямой доступ к которой организуется драйвером видеоадаптера через шину AGP или PCIe. В случае использования архитектуры Uniform Memory Access в качестве видеопамяти используется часть системной памяти компьютера.

Графический контроллер

В его функции входит обработка команд от хоста и формирование буфера кадра в растровом формате в видеопамяти, даёт команды RAMDAC на формирование сигналов развёртки для монитора и осуществляет обработку запросов центрального процессора.

Кроме этого, обычно присутствуют контроллер внешней шины данных (например, PCI или AGP), контроллер внутренней шины данных и контроллер видеопамяти.

Ширина внутренней шины и шины видеопамяти обычно больше, чем внешней (64, 128 или 256 разрядов против 16 или 32), во многие видеоконтроллеры встраивается ещё и RAMDAC. Современные графические адаптеры (ATI, nVidia) обычно имеют не менее двух видеоконтроллеров, работающих независимо друг от друга и управляющих одновременно одним или несколькими дисплеями каждый.

RAMDAC, TMDS

(Random Access Memory Digital to Analog Converter) — это устройство преобразования изображения в **цифровом** представлении из видеопамати в **аналоговые** сигналы для видеовыхода. (VGA выход, ТВ-выход)

ЦАП служит для преобразования изображения, формируемого видеоконтроллером, в уровни интенсивности цвета, подаваемые на аналоговый монитор. Возможный диапазон цветности изображения определяется только параметрами RAMDAC.

Чаще всего RAMDAC имеет четыре основных блока: три ЦАП, по одному на каждый цветовой канал (RGB), и SRAM для хранения данных о гамма-коррекции.

Большинство ЦАП имеют разрядность 8 бит на канал — по 256 уровней яркости на каждый основной цвет, что в сумме дает 16,7 млн цветов (а за счёт гамма-коррекции есть возможность отображать исходные 16,7 млн цветов в гораздо большее цветовое пространство). Некоторые RAMDAC имеют разрядность по каждому каналу 10 бит (1024 уровня яркости), что позволяет сразу отображать более 1 млрд цветов, но эта возможность практически не используется.

Для поддержки второго монитора часто устанавливают второй ЦАП. (Мониторы и видеопроекторы, подключаемые к цифровому DVI выходу видеокарты, для преобразования потока цифровых данных используют собственные ЦАП и от характеристик ЦАП видеокарты не зависят.

(Transition Minimized Differential Signaling) — это протокол передачи данных, используемый в интерфейсе DVI. Термин применяется также для обозначения самого электрического канала передачи данных.

3D-процессор

Представляет собой сложное устройство, сочетающее **специализированные** блоки (фильтрации, выборки, преобразования координат и т.п.) и **универсальные** вычислительные блоки (ALU), управляемые сложным диспетчером.

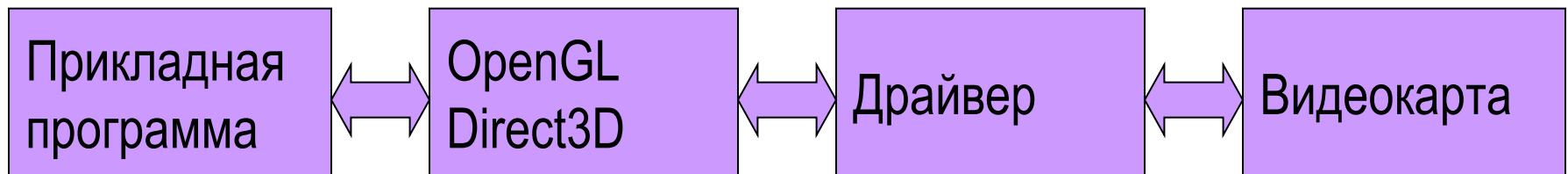
Количество ALU может превышать несколько тысяч. Однако единого подхода к архитектуре пока нет, ALU могут быть как векторными суперскалярными, так и обычными скалярными. Взаимосвязь и количественные соотношения между ALU, диспетчером, блоками фильтрации, преобразования координат, Z-буферизации, отставки, записи в память также могут быть различными.

Специальные блоки 3D-процессора можно поделить на блоки

- геометрической обработки (трансформация, освещение/затенение, преобразование координат, настройка треугольников),
- текстурирования,
- фильтрации,
- пост-обработки (отсечение, Z-буферизации, преобразования цветов, отставки).

3D-ускорители

- “Ускоряются” этапы T&L и растеризации
- Взаимодействие с программой при помощи специальных API



Создание графического объекта

- Моделирование — создание трёхмерной математической модели сцены и объектов в ней. Выполняет CPU.
- Рендеринг (визуализация) — построение проекции в соответствии с выбранной физической моделью. **Выполняет GPU.**
- Вывод полученного изображения на устройство вывода - дисплей или принтер.

Рендеринг состоит в преобразовании 3D объекта в 2D кадр, при этом часть информации теряется, прежде всего, о глубине объекта.

Чтобы сделать объект реалистичным, объекты проходят несколько стадий обработки. Самые важные стадии это:

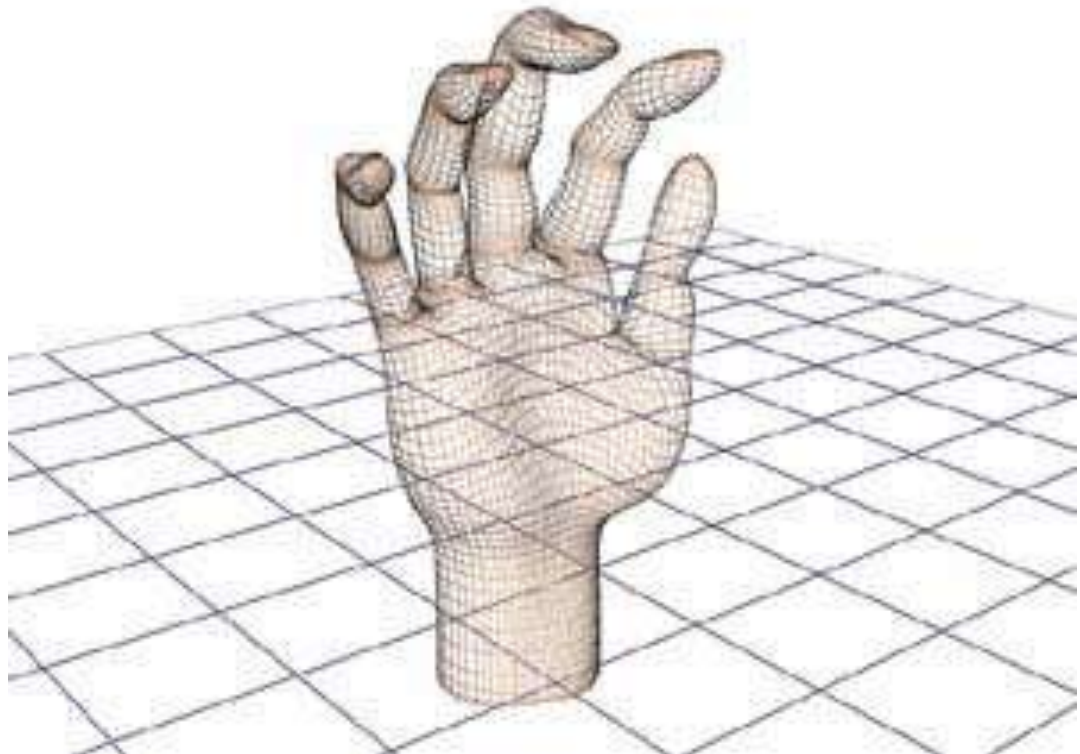
- создание формы (shape),
- обтягивание текстурами,
- освещение,
- создание перспективы,
- глубины резкости (depth of field)
- сглаживания (anti-aliasing).

Выполняются эти шаги CPU и GPU

Создание формы

- Для того чтобы составить достоверную картинку с кривыми линиями как в окружающем мире, приходится компоновать форму из множества мелких формочек (**полигонов**). Вместе они будут образовывать структуру, называемую **каркасом**.

каркас руки,
составленный
из 862
полигонов

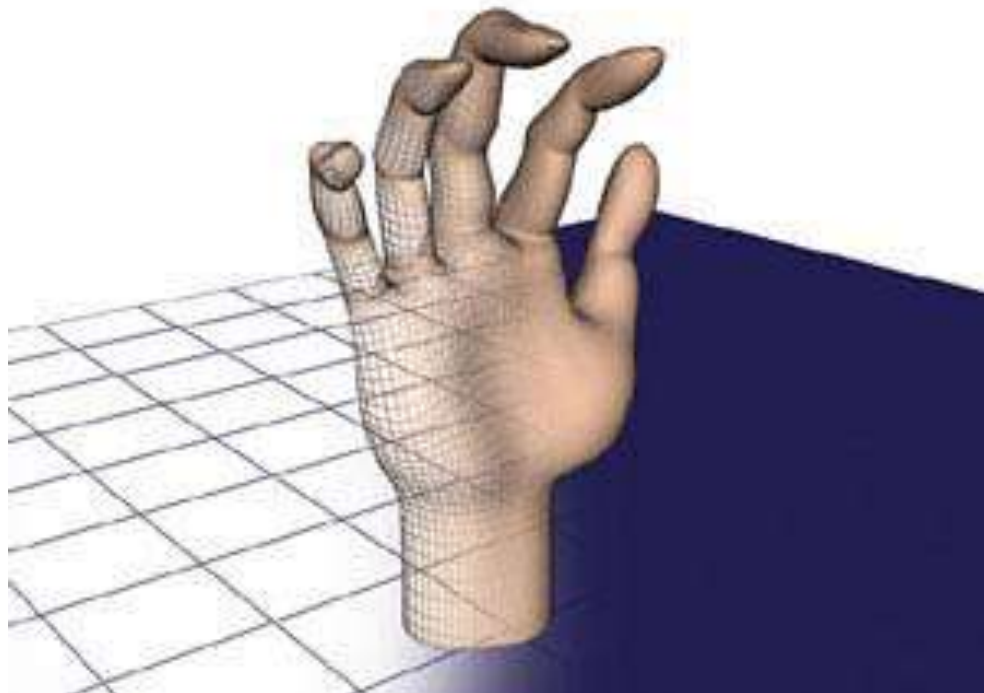


Создание поверхности каркаса

- Цвет: какого поверхность цвета? Однородно ли она окрашена?
- Текстура: ровная ли поверхность, есть вмятины, бугры, рихтовка?
- Отражающая способность: отражает ли свет? Четкость отражения ?

Придание "реальности" объекту состоит в подборе комбинации этих трех составляющих в различных частях изображения.

Добавление
поверхности к
каркасу
улучшает
изображение



Освещение

Освещение играет ключевую роль в двух эффектах, придающих ощущение веса и цельности объектам:

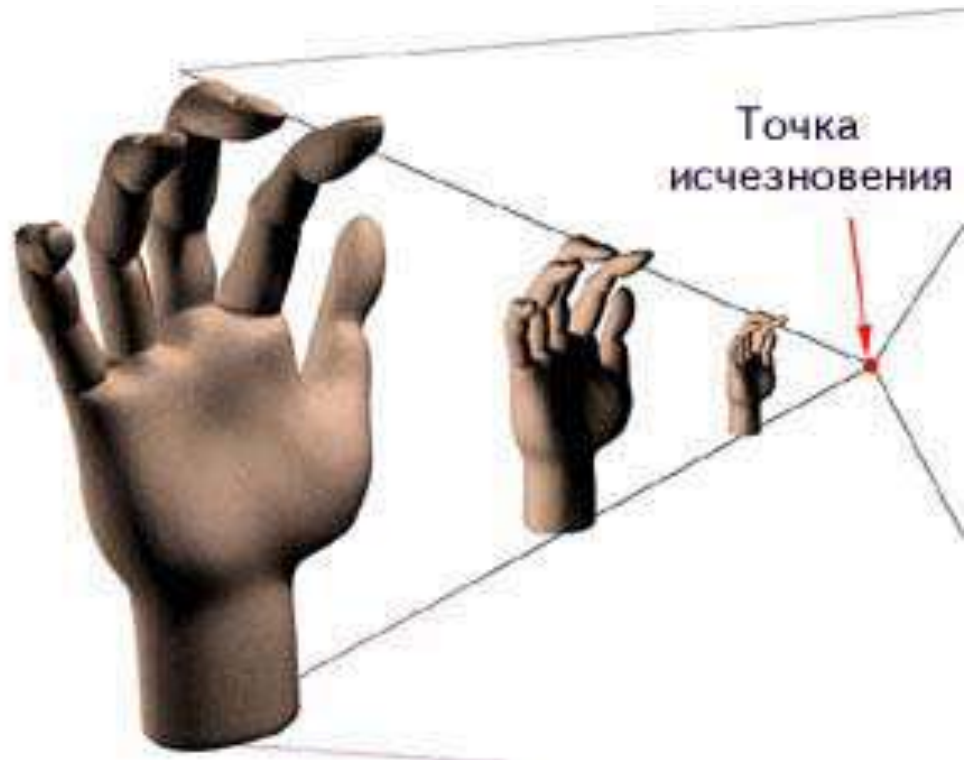
- затенения (shading) - изменение интенсивности освещения объекта от одной его стороны к другой
- тени (shadow).



Перспектива

Если все объекты на экране будут сходиться в одну точку, то это и будет называться перспективой.

Для таких сцен необходимо учитывать информацию, какие объекты закрывают другие и насколько сильно. Наиболее часто для этого используется Z-буфер

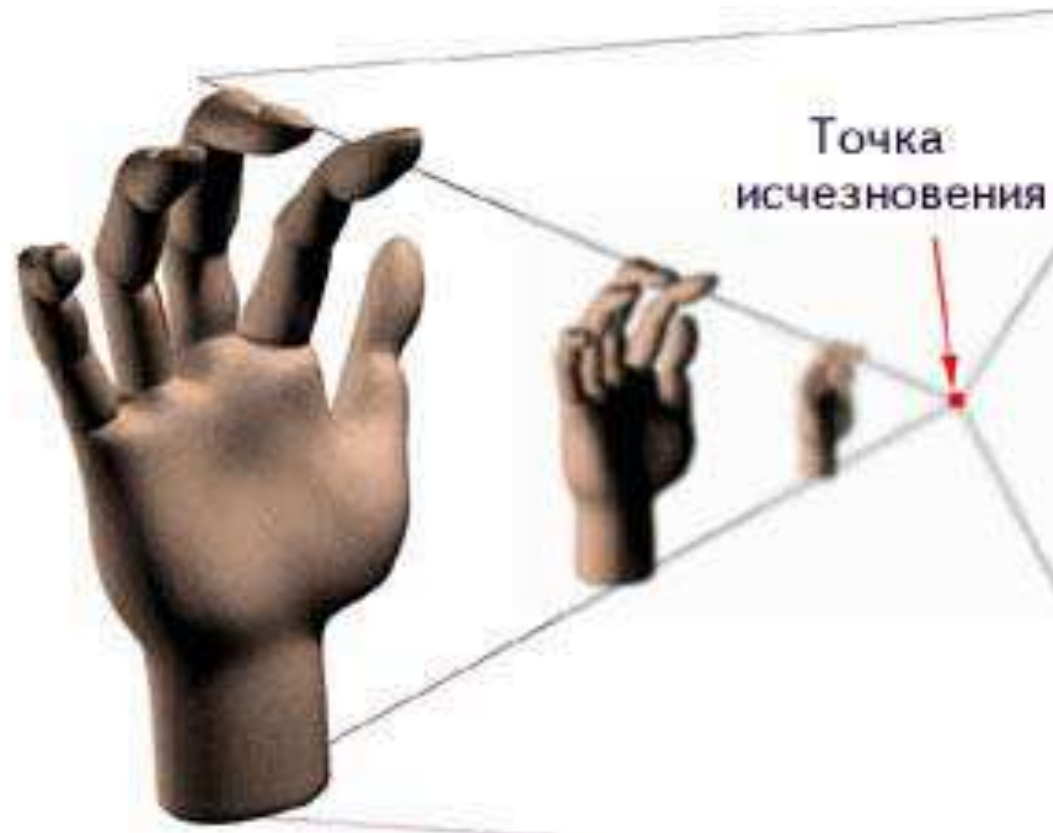


Z-буфер

Z-буфер присваивает каждому полигону номер в зависимости от того, насколько близко к переднему краю сцены располагается объект, содержащий этот полигон. Обычно меньшие номера присваиваются ближайшим к экрану полигонам. Объект с самым маленьким Z-значением будет полностью прорисовываться, другие же объекты с большими значениями будут прорисованы лишь частично.

Глубина резкости

По мере удаления объекта от наблюдателя будет потеря резкости.
Это тоже надо реализовать.



Типы шейдеров

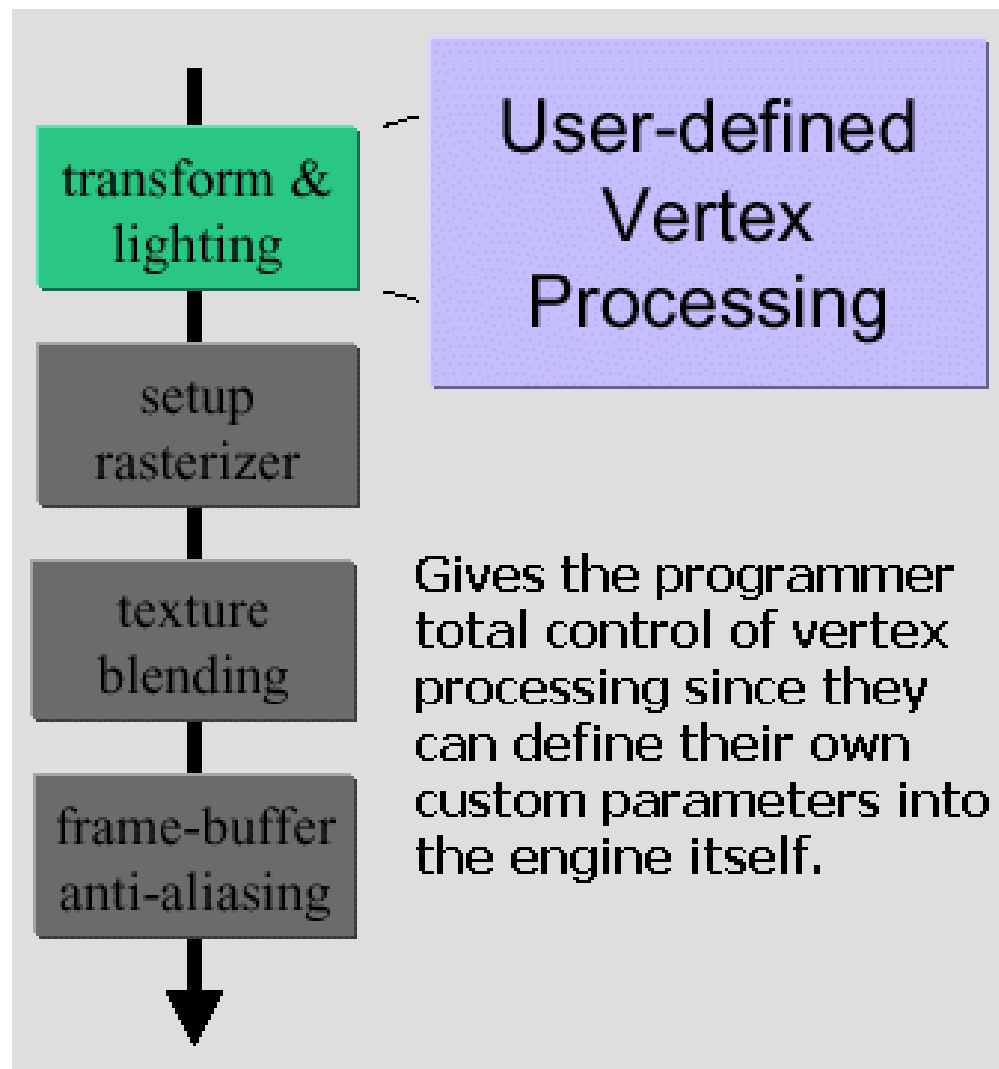
- **Вершинный шейдер** оперирует расположением узлов пространственной сетки, которая формирует каркас 3D-модели. Как мы знаем, точка в 3D-графике задается, как правило, набором из 4-х значений (x, y, z, w) . Компонент w является масштабом.
- Путем программирования вершинных шейдеров можно изменять расположение объекта в пространстве и рассчитывать эффекты его освещения.
- Пиксельные шейдеры позволяют изменить текстуру виртуальной кожи объекта, придавая ей соответствующую фактуру и цвет.

Геометрические шейдеры активируются при быстром приближении объекта к зрителю, добавляя изображению необходимые подробности для реализма.

Вершинный шейдер

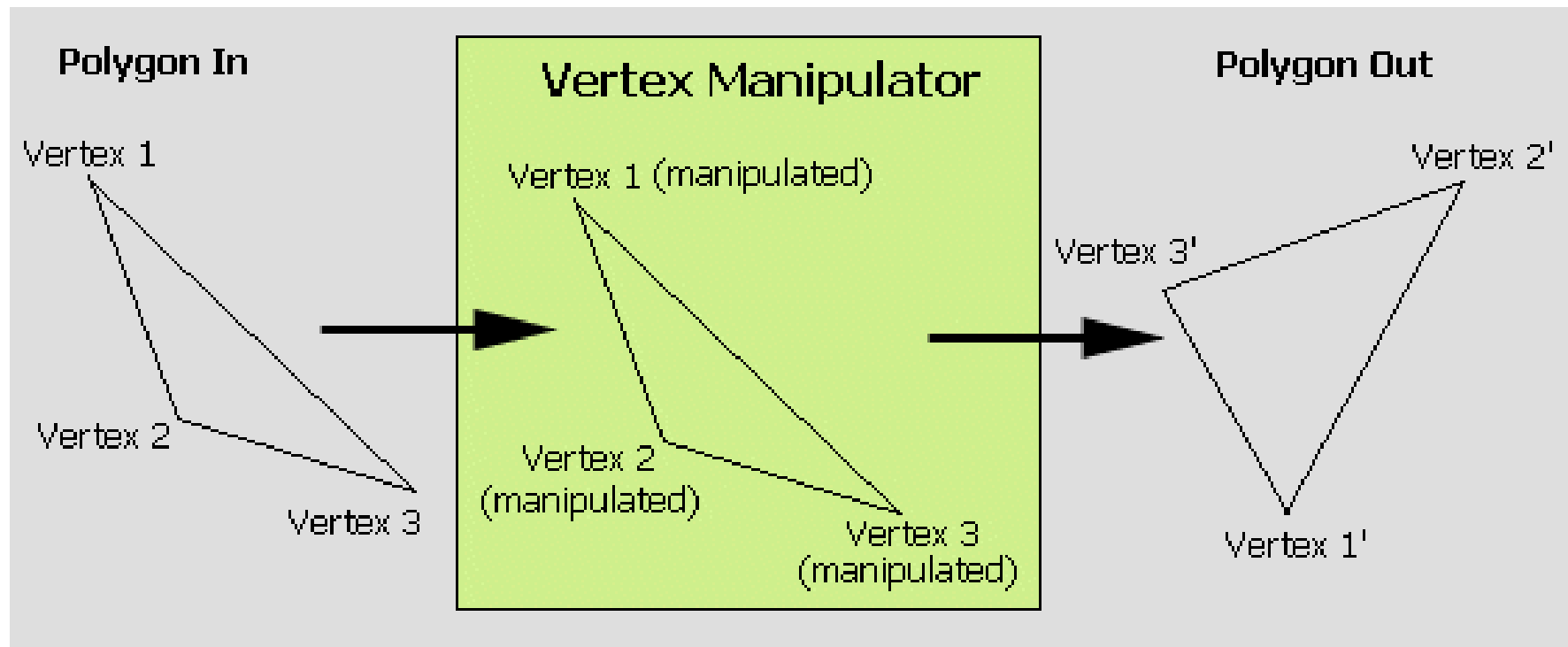
Путем программирования вершинных шейдеров можно

- изменять расположение объекта в пространстве и рассчитывать эффекты его освещения
- динамически вставить кусок кода на ассемблере прямо в конвейер,
- изменить различные настройки и затем продолжить процесс.

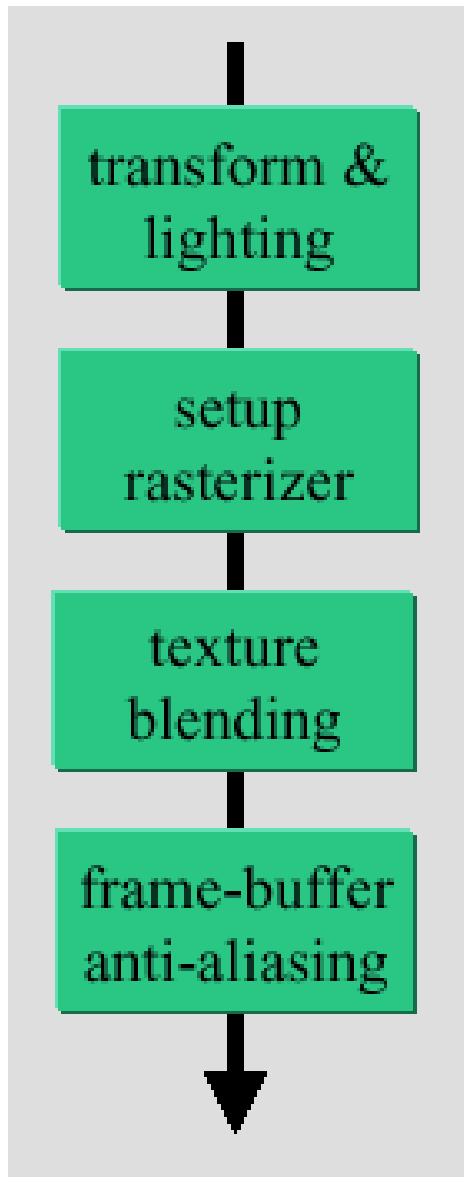


Преимущества вершинных шейдеров

- Полное управление аппаратным T&L;
- Сложные вершинные операции аппаратно ускоряются;
- Попиксельное наложение карт среды может опираться на вершинные данные (pre-vertex set up);
- Морфинг объектов (character morphing) и теневая проекция (shadow volume projection);
- Настраиваемое вершинное освещение (vertex lighting);
- Настраиваемое обтягивание скелета (skinning) и смешение текстур (blending);
- Настраиваемая генерация координат текстур;
- Настраиваемые матричные операции с текстурами (texture matrix operations);
- Настраиваемое освещение в стиле мультфильма (cartoon-style lightning);
- Программируемое вычисление вершин (vertex computations);
- Освобождаются ресурсы центрального процессора.



Графический конвейер



Все приведенные последовательно реализуемые действия над изображением составляют графический конвейер, который реализуется программно-аппаратными средствами.

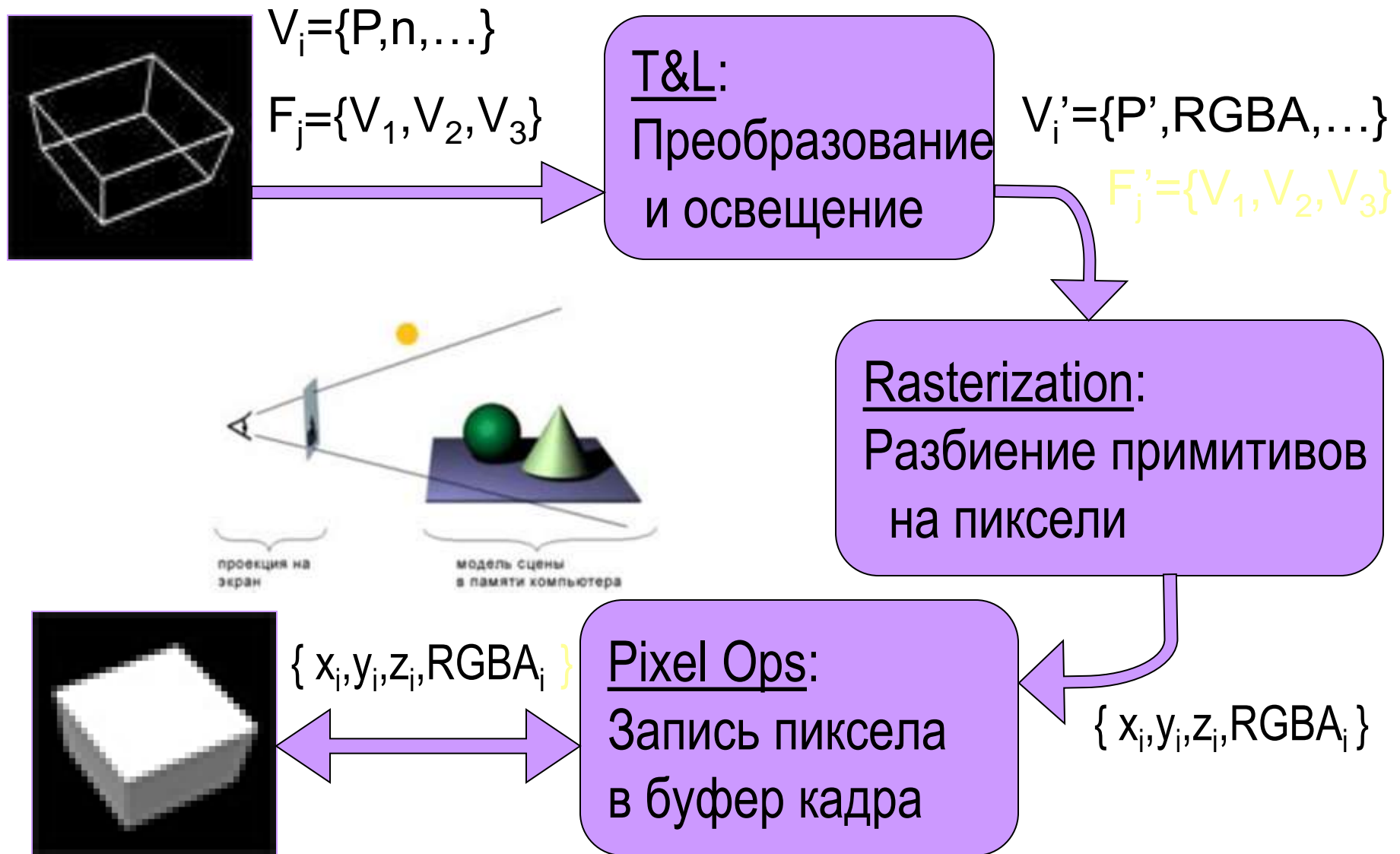
Графический конвейер переводит объекты, описанные в трехмерном пространстве XYZ, с учетом положения наблюдателя, во множество пикселей на экране монитора.

Обработка объектов в GPU производится с помощью специальных программ - шейдеров, выполняемых внутри GPU.

Этапы графического конвейера

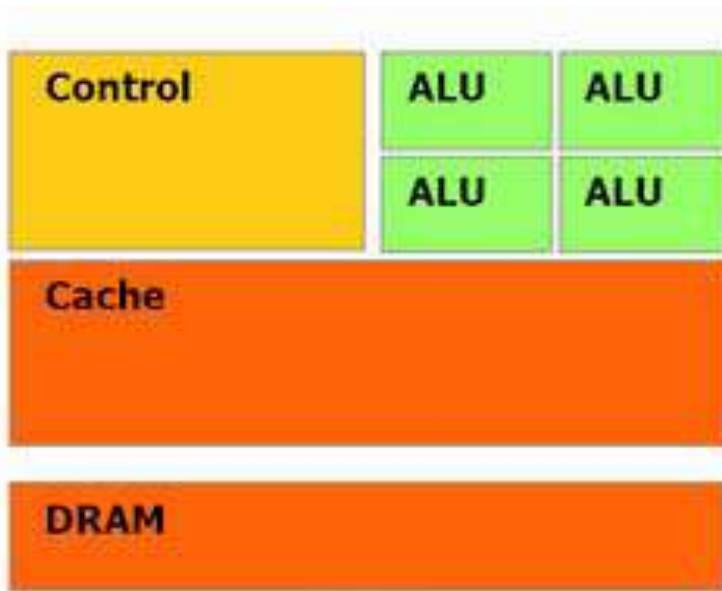
- **Этап 1.** Видеопроцессор получает от CPU информацию об объектах, которые необходимо обработать, и сцене.
- **Этап 2.** Вершинный процессор ядра (ядер может быть много) строит конкретный объект в пространстве сцены с фиксированными координатами, называемый вершиной (vertice). Режим MIMD.
- **Этап 3.** Сборка - вершины собираются в примитивы – треугольники (полигоны), линии или точки.
- **Этап 4.** Пиксельный процессор определяет конечные пиксели, которые будут выведены в кадровый буфер, и проводит над ними различные операции (см. стадии рендинга).
- **Этап 5.** Из Z-буфера вычитываются данные о расположении конкретных пикселей, чтобы отбросить те, которые будут скрыты другими объектами и не видны пользователю. Фрагменты снова собираются в полигоны, состоящие из отдельных пикселей, и весь массив уже отработанной картинки передаётся в кадровый буфер для последующего вывода на экран. Эта текстура обрабатывается в пиксельных процессорах (режим SIMD).

Графический конвейер

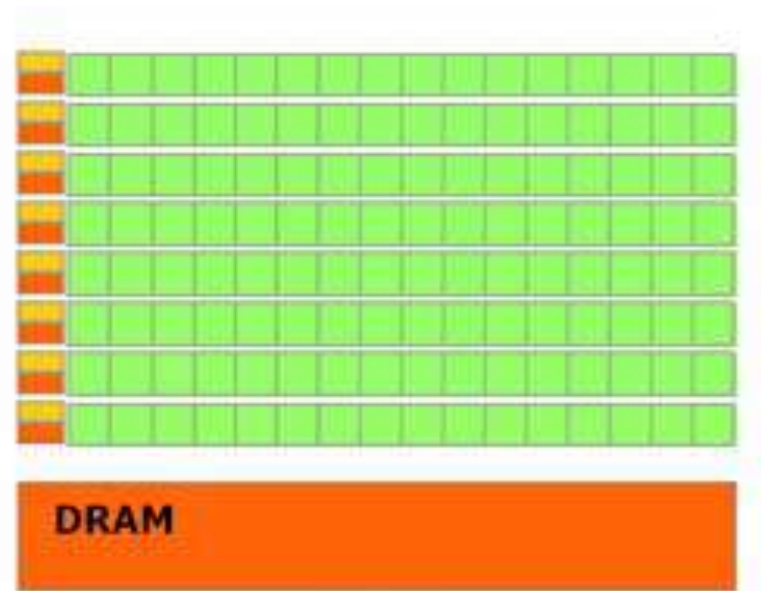


Потоковый процессор

Основной вычислительный элемент графического процессора – потоковый процессор (Streaming Processor – SP). Количество SP на кристалле графического процессора может составлять сотни и тысячи.



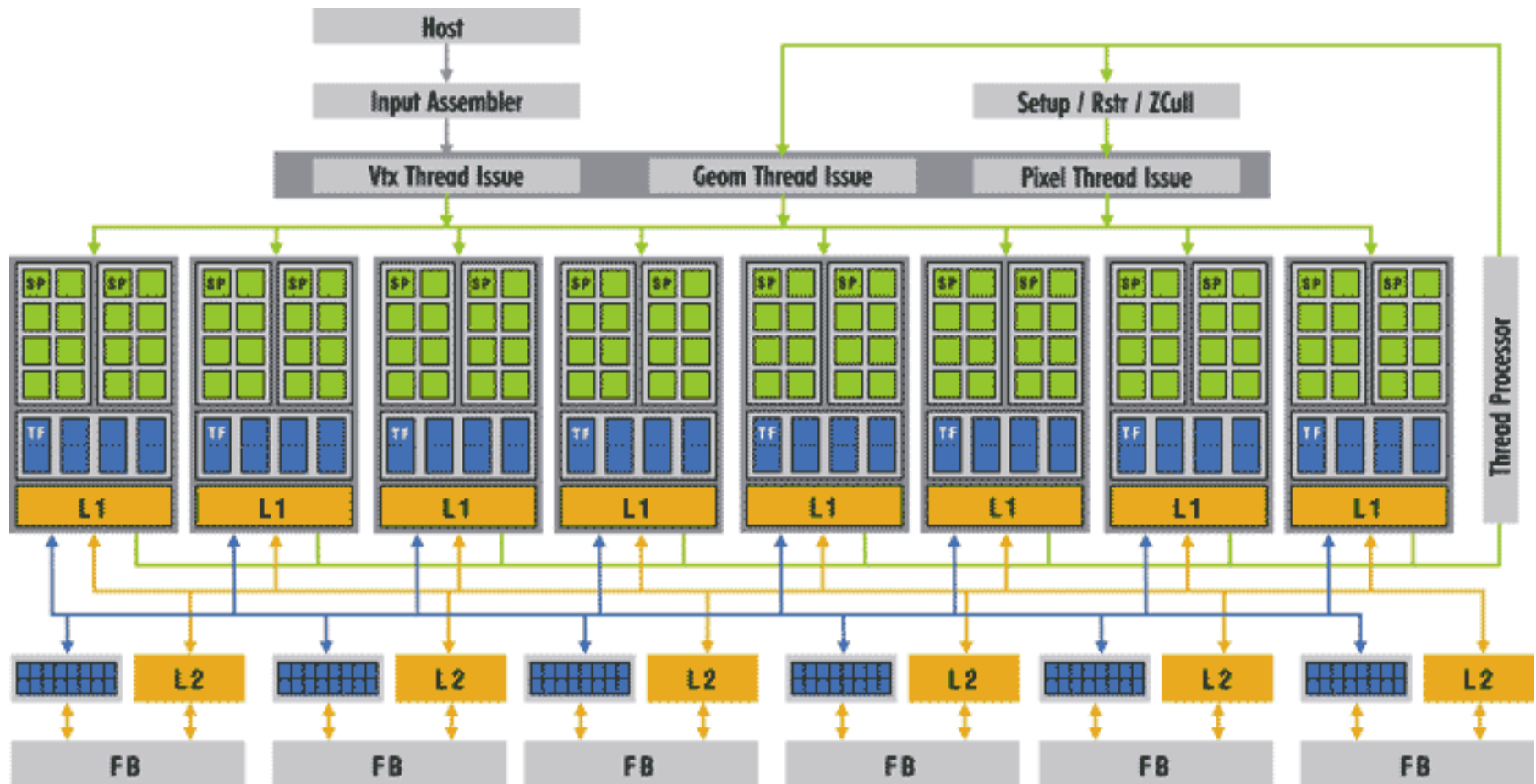
CPU



GPU

NVIDIA GeForce 8800

128 АЛУ, которые конструктивно объединены в 8 мультипроцессоров (ядер), каждый из которых оснащен четырьмя текстурными модулями и общим L1-кэшем.



Устройство

Мультипроцессор N

⋮

Мультипроцессор 2

Мультипроцессор 1

Разделяемая статическая память

Регистры

Регистры

Регистры

Устройство
управления

Процессор 1

Процессор 2

⋮


Процессор N

Кэш
констант

Кэш
текстур

Видеопамять

128 ПП объединены в SIMD-группы по 16 мультипроцессоров (МП), но при этом разные МП работают независимо друг от друга, хотя и исполняют один и тот же шейдер. Каждый ПП является суперскалярным устройством и может выполнять до двух команд за такт. При обращении в видеопамять ему доступна вся память, как на чтение, так и на запись. Однако на практике ввиду слабости средств синхронизации между различными МП желательно процесс обработки строить так, чтобы адреса записи не пересекались

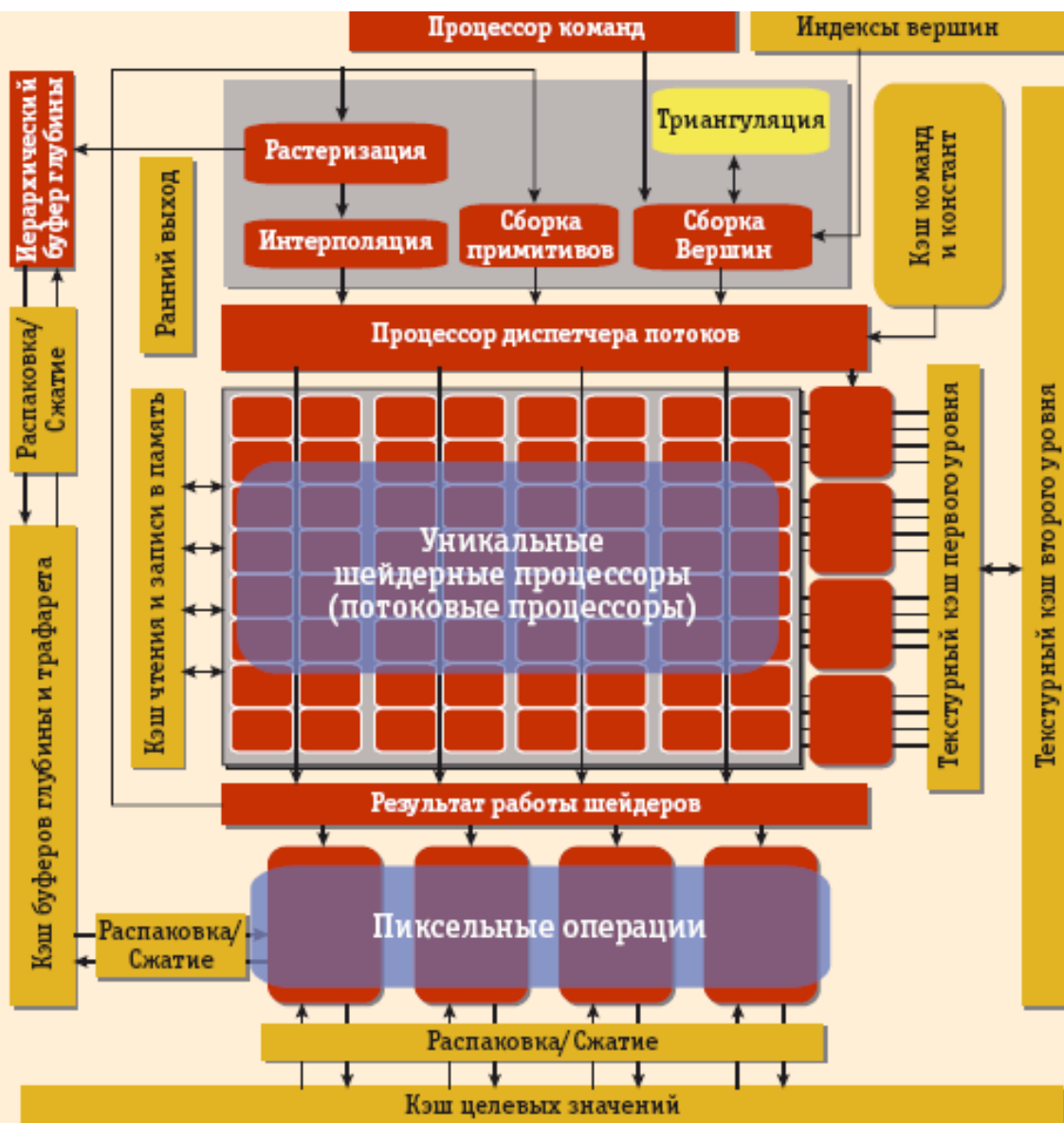


Каждое ядро представляет собой два шейдерных процессора (состоящих из восьми потоковых процессоров каждый), при этом все восемь блоков имеют доступ к любому из шести L2-кэшей и к любому из шести массивов регистров общего назначения.

На каждые четыре потоковых процессора приходится один текстурный блок, включающий один блок адресации текстур (Texture Address Unit, TA) и два блока фильтрации текстур (Texture Filtering Unit, TF).

- Архитектура GeForce 8 достаточно популярна, причина этого—интерфейс программирования CUDA (C Unified Driver Architecture).
- В nVidia первыми обеспечили возможность написания полных программ на диалекте языка Си. Это означает, что на этом языке можно одновременно писать код, исполняемый на графическом процессоре, и код, исполняемый на центральном процессоре, и все это в рамках одного проекта. Язык Си был расширен дополнительными квалификаторами памяти и функций для представления статической памяти и шейдеров соответственно. И хотя остался ряд ограничений, например отсутствие рекурсии на GPU, и программирование перемещения данных по-прежнему лежит на разработчике, язык впервые позволил программировать на GPU в терминах, понятных обычным программистам.

GPU компании AMD (Radeon 2K, 3K)



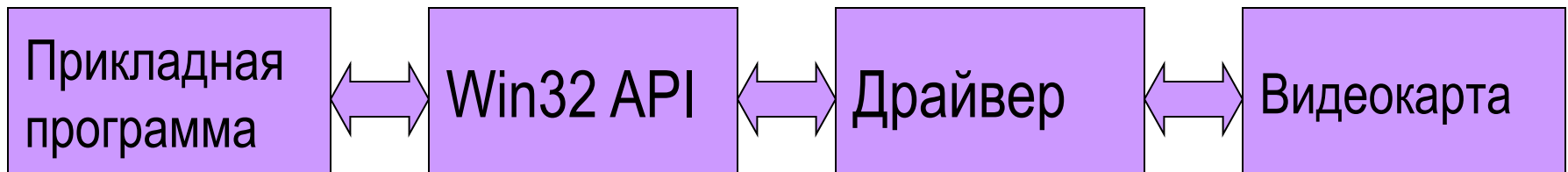
Современные GPU компании AMD, представленные сериями HD 2K и HD 3K, больше похожи на традиционные GPU, чем продукты nVidia.

2D-процессор

Используется для аппаратного ускорения GUI

Основные функции:

- Прорисовка примитивов – линий, кривых, полигонов
- Растеризация – вывод шрифтов, заливка, растяжение/сжатие, масштабирование
- Поддержка окон и спрайтов
- Поддержка курсора мыши



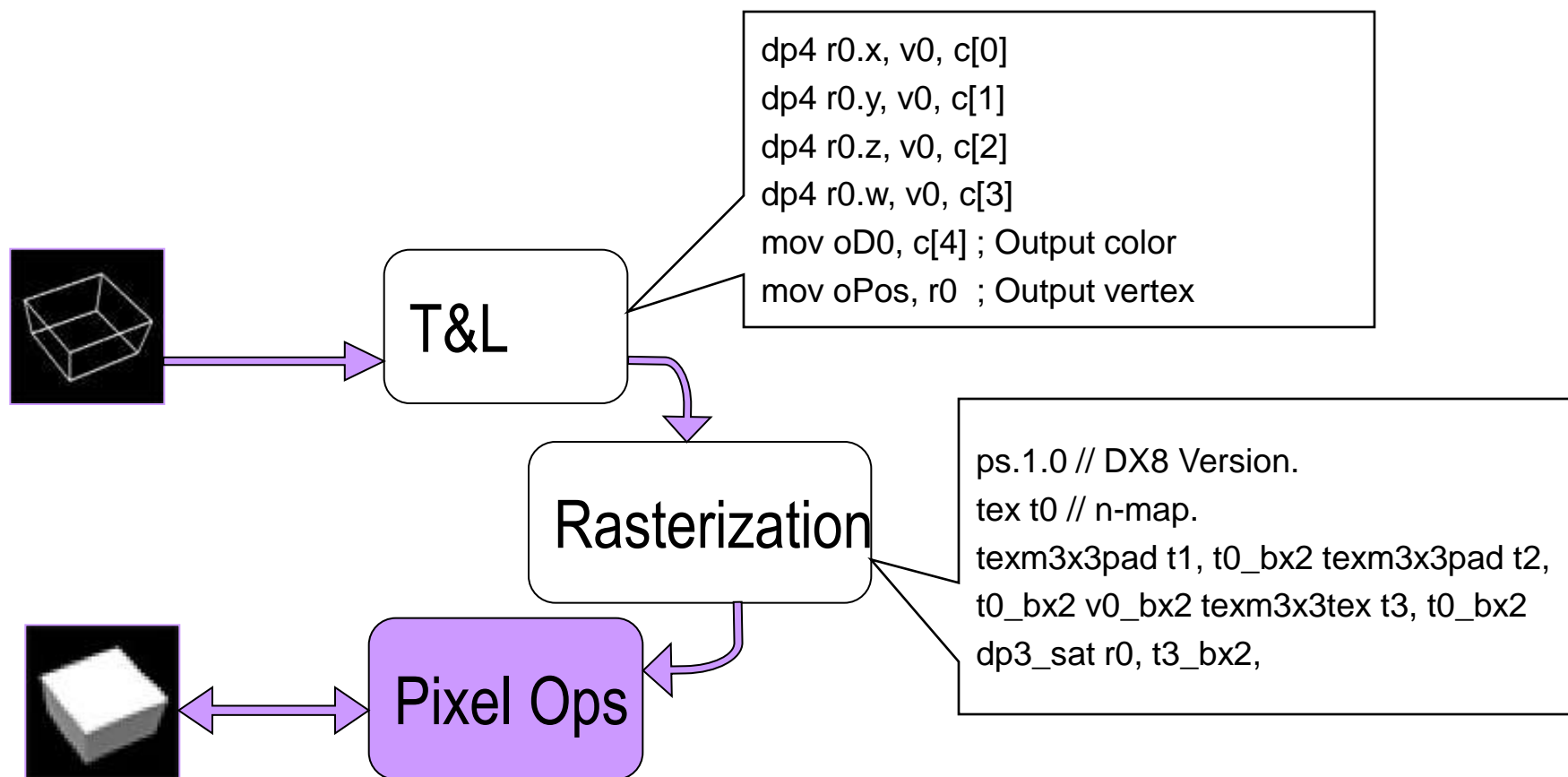
Видеопроцессор

Основные функции:

- Преобразование цветового пространства
- Коррекция гаммы, цветности, резкости и пр.
- Де-интерлейсинг(**deinterlacing**) - преобразование видеосигнала из чересстрочного режима в прогрессивный. Нужен для устранения эффекта гребенки, вызванного тем, при чересстрочной развертке выводятся два последовательных полукадра, а при прогрессивной один, склеенный из двух разных частей (события во втором полукадре происходят чуть позже, чем в первом). Эффект хорошо заметен на движущихся в горизонтальном направлении объектах.
- Компенсация движения
- Удаление «блочности» (выражается в разбиении кадров сжатого видео на квадраты) и др. функции улучшения качества
- Масштабирование
- Специфические функции для декодирования сжатого видео

Построенная картинка отображается с помощью механизма оверлеев.

Современные ускорители: шейдеры



Локальная видеопамять

Как правило, предпочтение отдается динамической памяти с наивысшей пропускной способностью. Задержки доступа в данном случае менее важны. Увеличение «ширины канала» достигается в том числе и за счет увеличения физической ширины шины памяти, несмотря на существенное усложнение дизайна печатной платы.

Как правило, видеопамять, полностью или частично, отображена на область системных адресов. При этом диапазон адресов видеопамети обычно не кэшируется, о чем делается пометка в MTRR.

Технология DiME и производные отводят ряд адресов под апертуру GART. Обращение к этим адресам приводит к обращению к страницам системной памяти. С исчезновением AGP механизм GART стал ненужным. Однако к идее пришлось вернуться, и сегодня большинство видеокарт способно использовать память в разделяемом режиме, с возможностью обращаться к заданным адресам системной памяти для чтения текстур.



В случае AGP графический акселератор практически напрямую обменивается информацией с системной памятью. Благодаря высокой скорости передачи между графическим акселератором и основной памятью AGP сделала возможным использование оперативной памяти в дополнение к локальной видеопамяти в таких случаях, как, например, работа с текстурой (т.е. для объемного раскрашивания рисунка). Корпорация Intel назвала такую технологию DIME (Direct Memory Execute). Следует отличать DIME от предложенной ранее технологии UMA (Unified Memory Architecture), при которой основная память уже использовалась как дополнение к видеопамяти. Эти две технологии имеют два существенных отличия.

Контроллер CRT

Его задача – генерация сигналов доступа к видеопамяти и сигналов синхронизации интерфейса подключения дисплея.

Возможно, и другие функции также отводятся этому контроллеру, в частности, функции графического контроллера (запись/чтение пикселей, модификация цвета, коррекцию гаммы и т.п.).

За разрешение и глубину цвета отвечает именно CRTC.

К CRTC подключаются преобразователи интерфейса, часто – по два:

- RAMDAC для аналогового VGA
- TDMS-трансивер для DVI-I (HDMI, DisplayPort)
- Кодер ТВ-сигнала для телевизионного выхода

Встроенный графический процессор (IGP, Integrated Graphics Processor)

GPU), встроенный (интегрированный) в материнскую плату компьютера и (или) в CPU

Встроенная видеокарта (интегрированная или «onboard») является частью чипсета, как правило, располагается внутри микросхемы его "северного моста".

Виды IGP

3 основных вида

- **С разделяемой памятью** (часть северного моста). В качестве видеопамяти используют ОЗУ. Преимущества данного решения — низкая цена и малое энергопотребление. Недостатки — невысокая производительность в 3D-графике и отрицательное влияние на пропускную способность памяти (Intel, ATI, SiS и NVidia)
- **Дискретная графика** (*Dedicated graphics*). На системной плате или (реже) на отдельном модуле распаяны видеочип и один или несколько модулей видеопамяти. Обеспечивает наивысшую производительность в 3D-графике. Недостатки: более высокая цена (для высокопроизводительных процессоров — очень высокая) и большее энергопотребление (AMD, ATI и Nvidia)
- **Гибридная дискретная графика** (*Hybrid graphics*) - комбинация вышеназванных способов, ставшая возможной с появлением шины PCIe. Наличествует небольшой объём физически распаянной на плате видеопамяти, который может виртуально расширяться за счёт использования основной ОП. Компромиссное решение, с разной степенью успеха пытающееся нивелировать недостатки двух вышеназванных видов. но не устраняет их полностью

APU - Accelerated Processing Unit

AMD Fusion

Аббревиатура APU расшифровывается как Accelerated Processing Unit (ускоренное процессорное устройство). Если обратиться к подробным разъяснениям, то оказывается, что с аппаратной точки зрения это – гибридное устройство, объединяющее на одном полупроводниковом кристалле традиционные вычислительные ядра общего назначения с графическим ядром. Иными словами, тот же CPU с интегрированной графикой. Однако разница всё-таки есть, и кроется она на программном уровне. Графическое ядро, входящее в APU, должно иметь универсальную архитектуру в виде массива потоковых процессоров, способных работать не только над синтезом трёхмерного изображения, но и над решением вычислительных задач.

Внешний графический процессор (eGPU)

- Внешний графический процессор — это графический процессор, расположенный за пределами корпуса компьютера. eGPU иногда используются совместно с портативными компьютерами. Ноутбуки могут иметь большой объем RAM и достаточно мощный CPU, но часто им не хватает мощного GPU, вместо которого используется менее мощный, но более энергоэффективный встроенный графический чип. Встроенные графические чипы обычно недостаточно мощны для воспроизведения новейших игр или для других графически интенсивных задач, таких как редактирование видео.
- Поэтому желательно иметь возможность подключать графический процессор к некоторой внешней шине ноутбука. PCI Express — единственная шина, обычно используемая для этой цели. Порт может представлять собой, к примеру, порт ExpressCard или mPCIe (PCIe × 1, до 5 или 2,5 Гбит / с соответственно) или порт Thunderbolt 1, 2 или 3 (PCIe × 4, до 10, 20 или 40 Гбит / с соответственно). Эти порты доступны только для некоторых ноутбуков.