Казанский федеральный университет

ВЫСШАЯ ШКОЛА информационных технологий и информационных систем

Архитектура GPU

Эдуард Храмченков

История

- Первые графические ускорители – Vodoo 3DFx
- Задачи:
  - Растеризация
  - Наложение текстур
  - Альфа-блендинг
- Обработка вершин проводилась на CPU
- Riva TNT – обработка вершин без участия CPU

История

- GeForce 256 – блоки register combiners для выполнения простых вычислительных операций
- Сложный эффект – несколько последовательных вычислительных блоков
- GeForce 2 – ассемблер для создания вершинных программ
- Программы выполнялись параллельно для каждой вершины

История

- Начало 2000-х – выскоуровневые шейдерные языки (HLSL, Cg, GLSL)
- Схема работы
  - Входные данные загружаются в текстуры при помощи OpenGL/DirectX
  - Обработка этих данных через операции рендеринга на ускорителе
  - Выгрузка результата в системную память
- 2 блока программы – для CPU и для GPU

История

- Графические API имеют ряд ограничений и неудобны для вычислений общего назначения
- В 2007 году компания Nvidia выпустила API для вычислений общего назначения на графических ускорителях (GPGPU) – Nvidia CUDA
- 2016 год – приложения и научные статьи с использованием Nvidia CUDA, тысячи их

Особенности GPU

- GPU создавались для решения задач рендеринга изображений
- Рендеринг – применение одной и той же функции (освещенность, поворот, и т.д.) ко всем вершинам и элементам сцены
- Задачи рендеринга обладают значительным ресурсом параллелизации
- В таксономии Флинна архитектура GPU относится к классу SIMD

Особенности GPU

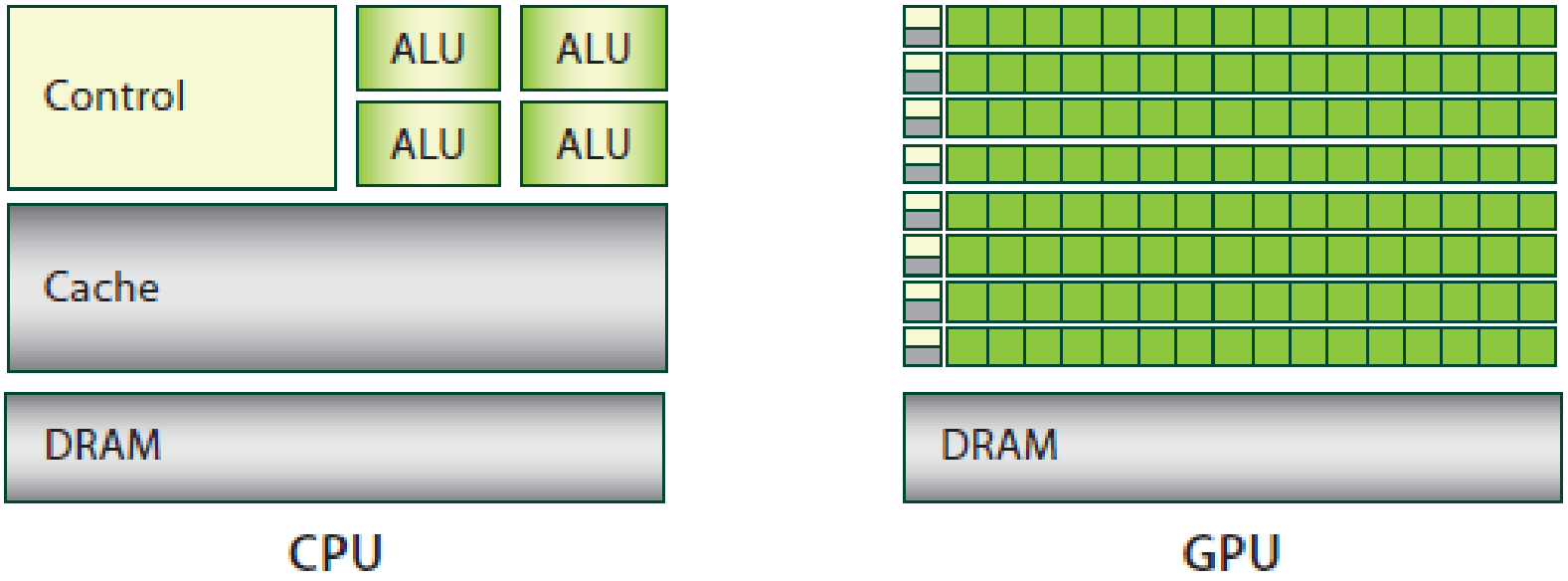
- GPU наиболее эффективны для решения задач параллельных по данным
- Количество арифметических операций >> количество операций с памятью
- В задачах компьютерной графики достаточно работы с числами в одинарной точности (SP – Single Precision, тип float)
- Для вычислительных задач необходима поддержка двойной точности (DP – Double Precision, тип double)

SIMD

Почему GPGPU

- Появление технологий GPGPU совпало с замедлением темпов роста производительности CPU
- В то же время GPU достигли такого уровня, что стали способны решать задачи, выходящие за рамки компьютерной графики
- CPU и GPU основаны на разных архитектурах, что диктует разный подход к их использованию

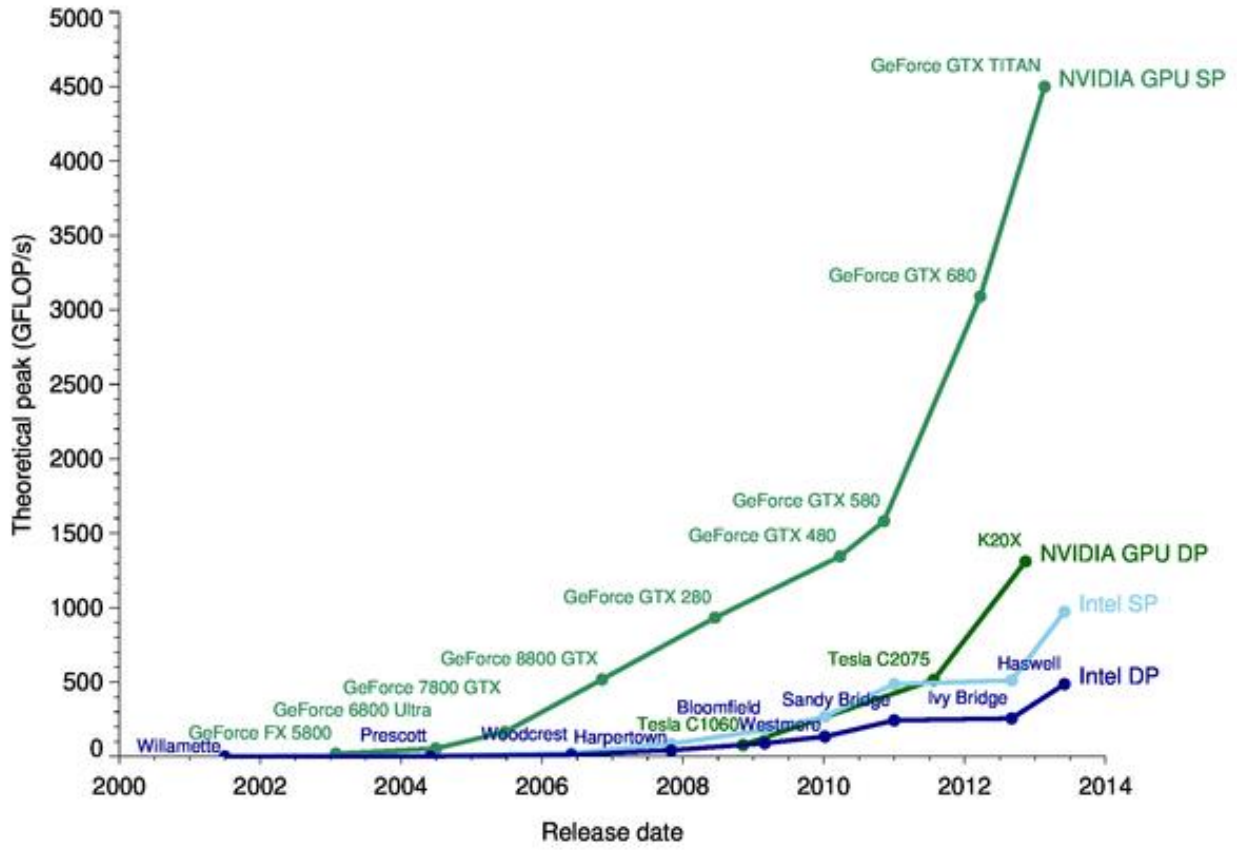
GPU vs. CPU: архитектура



GPU vs. CPU: архитектура

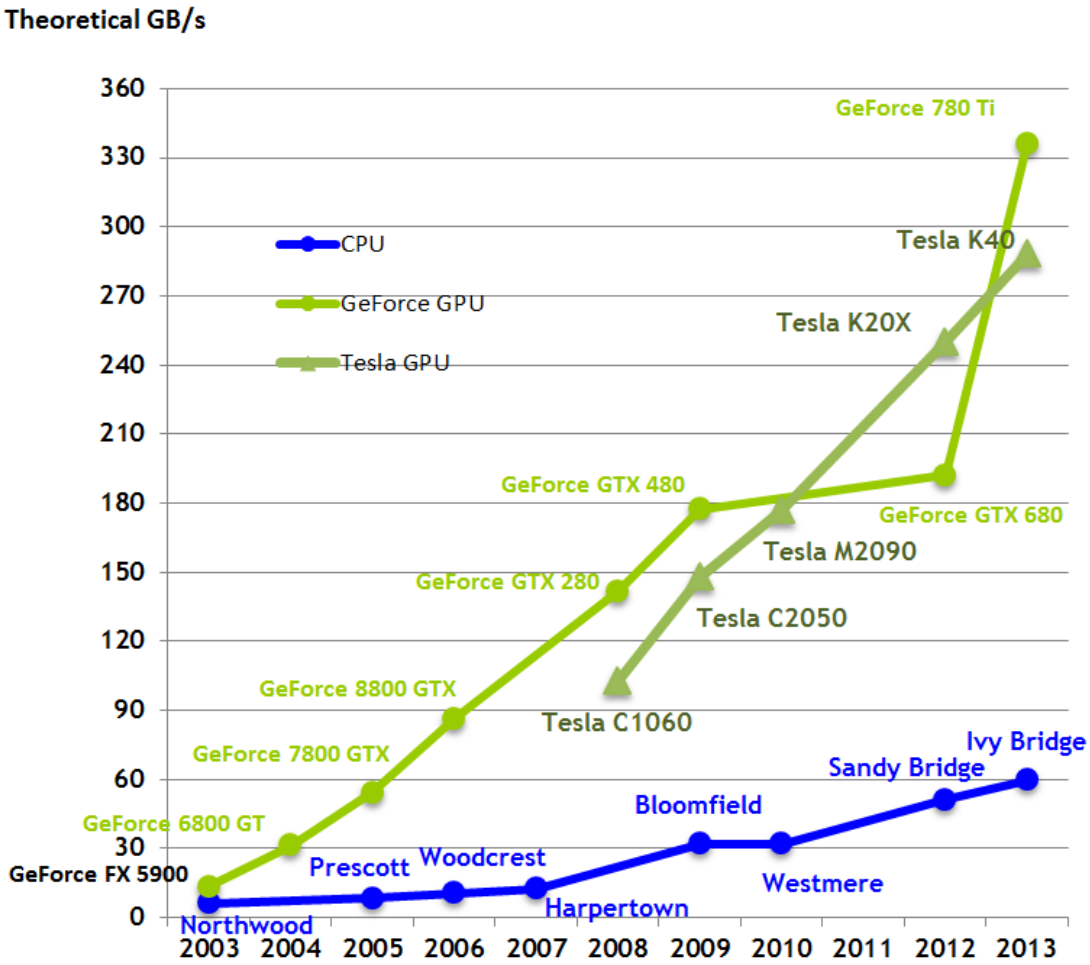
Характеристика	CPU	GPU
Количество ядер	<20	>100
Тактовая частота ядра	Высокая	Низкая
SSE, ветвление, и т.д.	Есть	Нет
Кеш	Большой	Незначительный

GPU vs. CPU: производительность



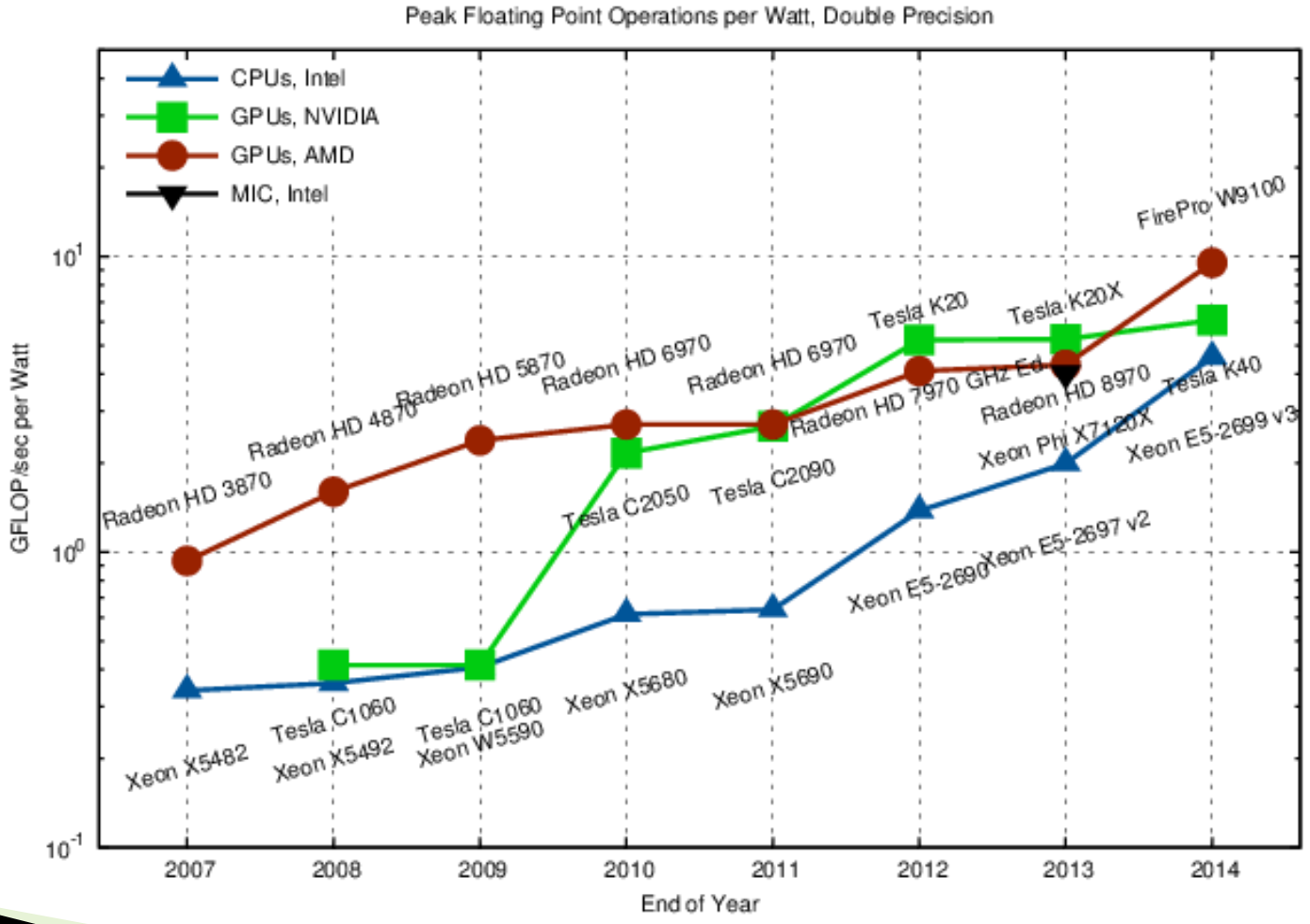
Казанский федеральный университет

GPU vs. CPU: скорость памяти



Казанский федеральный университет

GPU vs. CPU: энергоэффективность



Казанский федеральный университет

Железо

- Компании Nvidia и AMD выпускают несколько линеек графических ускорителей:
  - Игровые карты Nvidia GeForce и AMD Radeon – массовый сегмент для обычных пользователей
  - Профессиональные карты Nvidia Quadro и AMD Firepro – обработка графики и видео
  - Профессиональные карты Nvidia Tesla и AMD Firepro – суперкомпьютеры, высокопроизводительные вычисления

Железо

- GeForce/Radeon
  - Низкая цена – до 1200$ (Nvidia Titan X)
  - Урезанная двойная точность – производительность ⩽25% от HPC решений
  - Объем памяти – до 12Gb (Nvidia Titan X)
- Tesla/FirePro
  - Высокая стоимость – от 4000$
  - Полноценная производительность в двойной точности
  - Объем памяти – до 16Gb

Железо

- Видеокарты от AMD и Nvidia используют разные API для параллельных вычислений
- Карты AMD поддерживают OpenCL – открытый стандарт, позволяющий писать программы как под многоядерные CPU, так и под GPU
- Карты Nvidia оптимизированы под фреймворк CUDA – проприетарный продукт Nvidia; карты Nvidia так же поддерживают API OpenCL

Железо

- Для работы с Nvidia CUDA необходима видеокарта от Nvidia
- Не все видеокарты Nvidia способны работать с двойной точностью
- В актуальной версии Nvidia CUDA 8.0 поддерживаются видеокарты начиная с архитектуры Fermi
- Любая совместимая с CUDA видеокарта имеет числовую характеристику Compute Capability

Compute Capability

- Числовая характеристика видеокарты Nvidia вида X.Y
- X указывает на номер основной ревизии, Y на минорную ревизию
- По этому числу можно определить функции CUDA, поддерживаемые этой картой
- Устройства с одним номером основной ревизии относятся к одной архитектуре
- Чем больше CC тем лучше GPU

Compute Capability

Compute Capability	Архитектура
1.y	Tesla
2.y	Fermi
3.y	Kepler
5.y	Maxwell
6.y	Pascal

Compute Capability

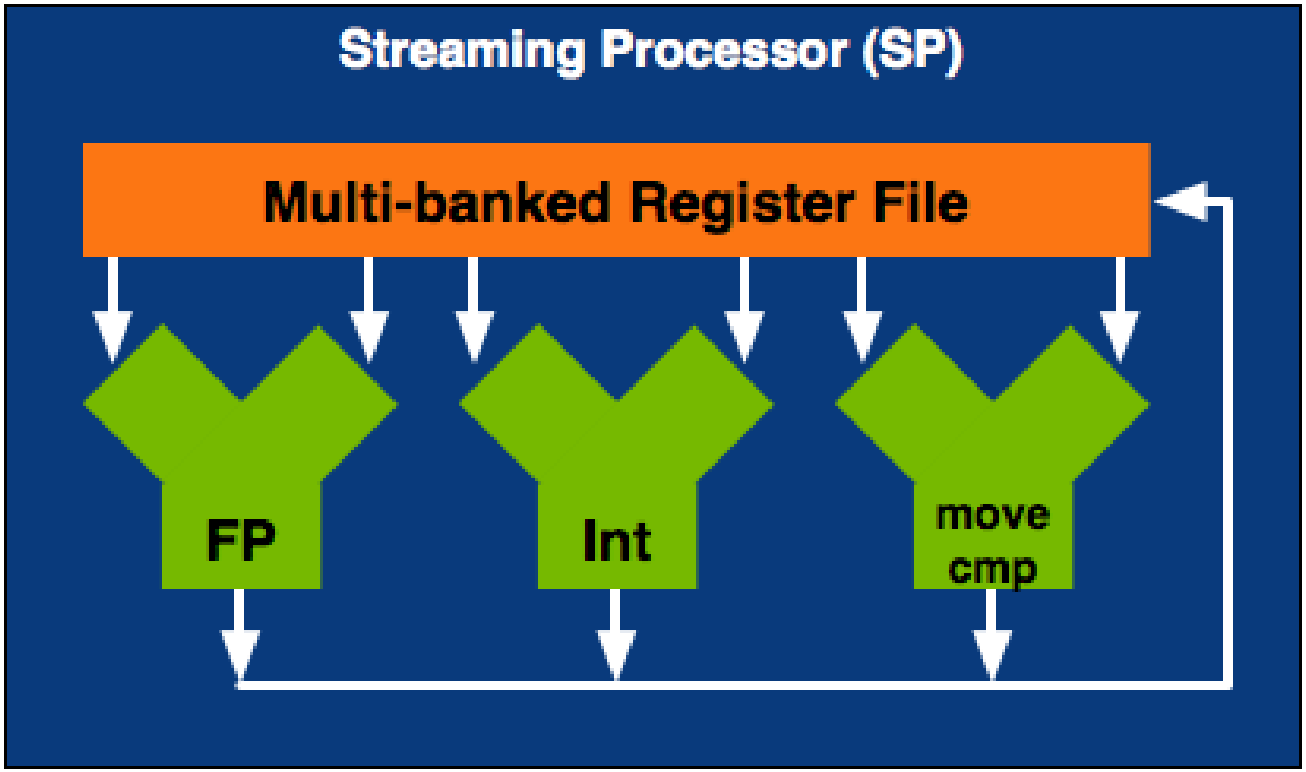
Compute Capability	Архитектура
1.y	Tesla
2.y	Fermi
3.y	Kepler
5.y	Maxwell
6.y	Pascal

Deprecated

Архитектура GPU

- Самый низкий уровень архитектуры составляют потоковые процессоры (SP – Streaming Processor)
- Каждый SP – микропроцессор с очередным типом исполнения команд, обладающий полноценным конвейером, парой ALU и FPU
- У SP нет кэш-памяти, он эффективен только в выполнении большого количества математических расчетов

Архитектура GPU: SP



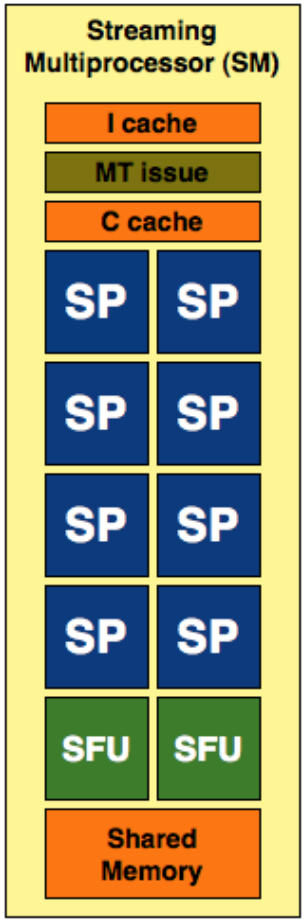
Архитектура GPU

- SP объединены в группы – потоковые мультипроцессоры (SM – Streaming Multiprocessor)
- SM – массив из нескольких SP и модулей специальных функций SFU – Special Function Units
- В каждом SM есть SP для работы с float(FP32) и double(FP64)
- Соотношение ядер FP32 и FP64 зависит от архитектуры

Архитектура GPU

- Каждый SFU содержит в своем составе FPU для выполнения трансцендентных операций (sin, cos и т.д.) и интерполяции
- В SM входит диспетчер исполнения команд MT, который занимается распределением нагрузки по SP и SFU
- В SM содержится кэш (⩾16 Кб), общий для всех SP

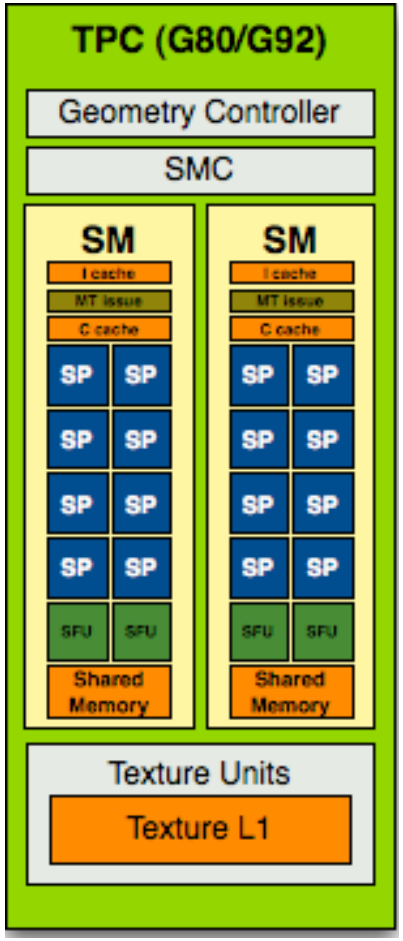
Архитектура GPU: SM



Архитектура GPU

- Следующим уровнем объединения является кластер SM, называемый Texture/Processor Cluster (TPC)
- Каждый TPC содержит контролирующий модуль TM и высокоуровневую управляющую логика
- Память – текстурный блок, в котором располагаются модули текстурной адресации и фильтрации, а так же текстурный кэш L1

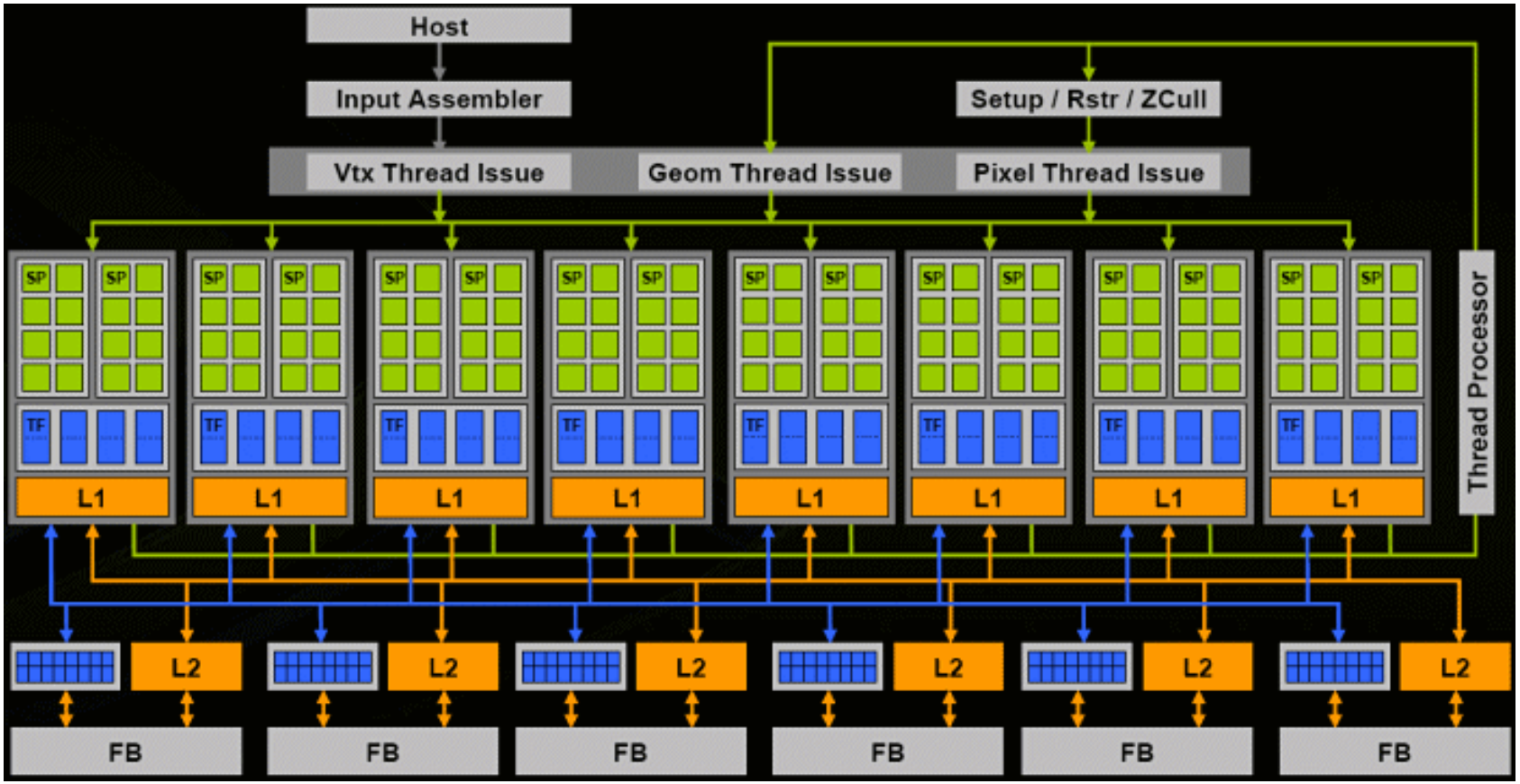
Архитектура GPU: TPC



Архитектура GPU

- Блоки TPC объединены в массив потоковых процессоров (SPA – Streaming Processor Array)
- Верхний уровень
  - Логика чипа распределяющая нагрузки по SPA
  - Контроллер PCI-Express
  - Шина Interconnect Network,
  - L2 кэш текстур
  - Блоки обработки растровой графики (Raster Operation Unit - ROP), которые имеют прямой доступ к фреймбуферу

Архитектура GPU



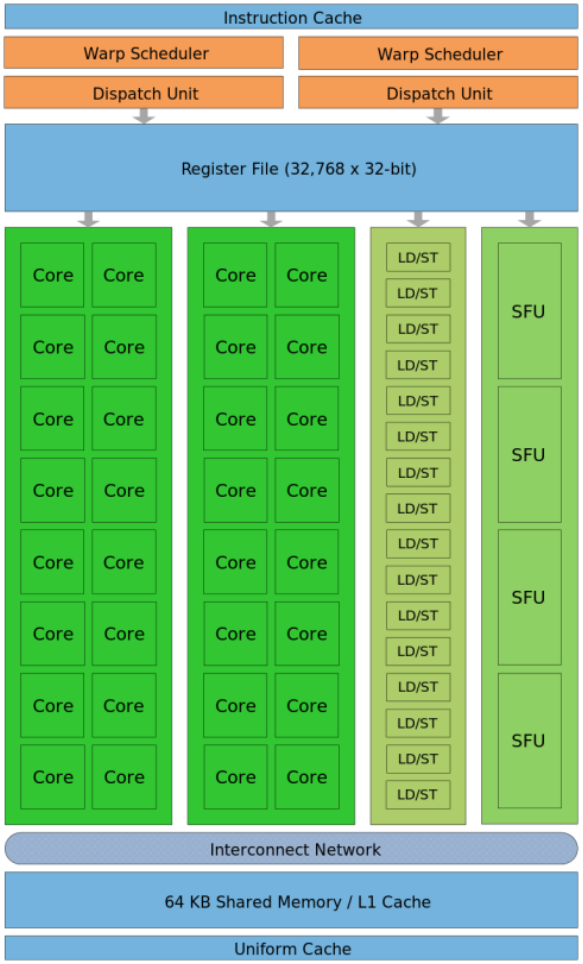
Архитектура Tesla

- Дата релиза: 2006 год
- Игровые карты поколений GeForce 8-9, и GeForce 100-300
- Первая профессиональная карта для HPC Nvidia Tesla
- До 30 SM по 8 SP в каждом
- SM-кэш (shared memory) 16 кб
- Производительность в DP – 77,76 GFLOPs на профессиональной карте (Nvidia Tesla C1060)

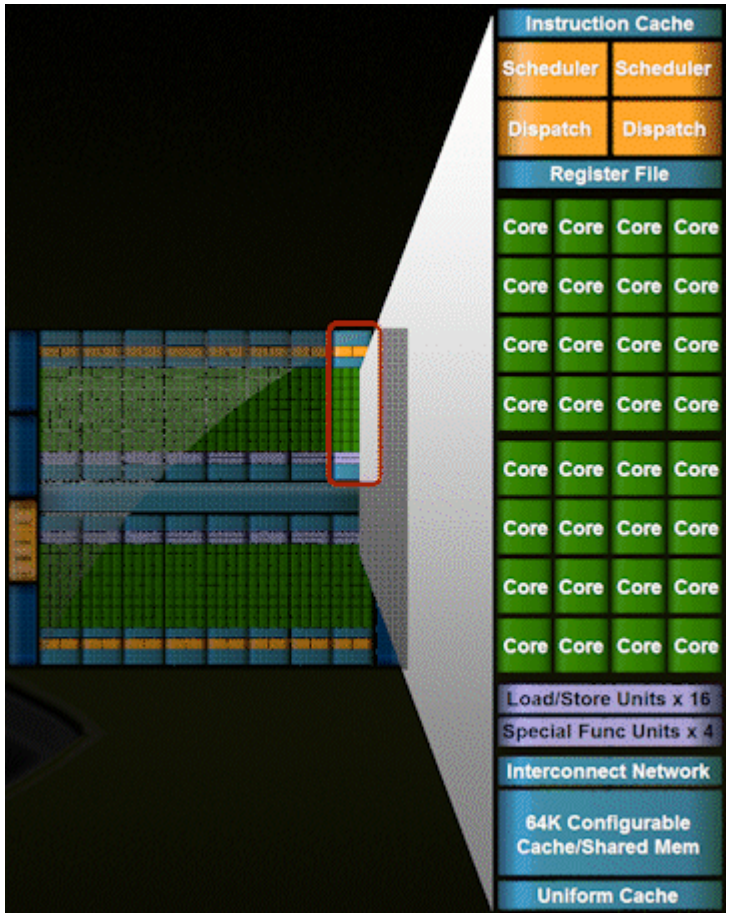
Архитектура Fermi

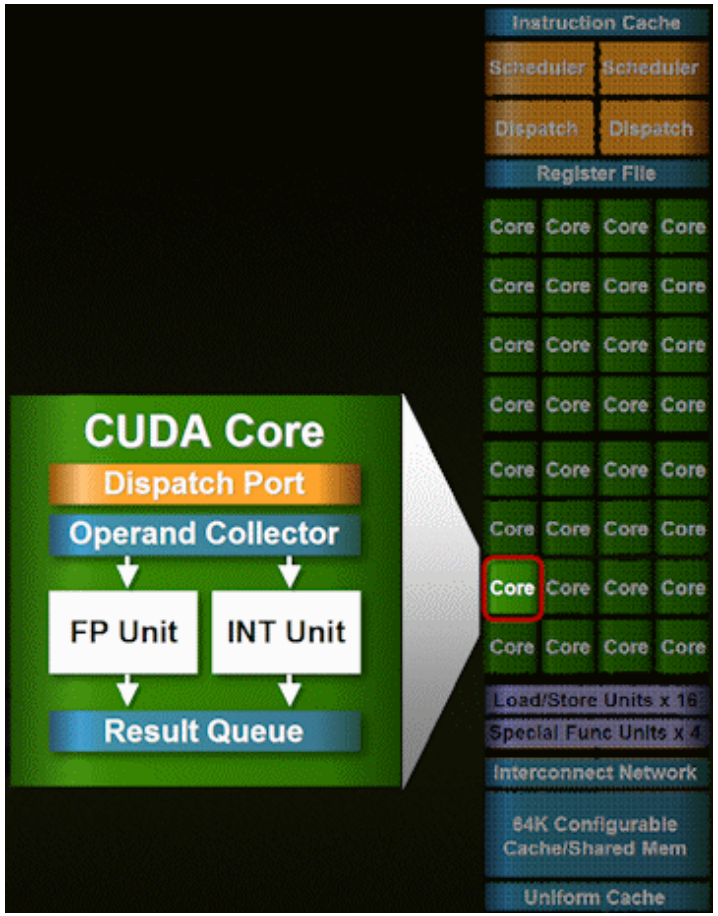
- Дата релиза: 2009 год
- Игровые карты семейств GeForce 400-500
- 16 SM – каждый содержит 32 SP, всего 512 ядер CUDA
- SM объединены в Graphics Processing Clusters (GPC)
- SM-кэш/L1-кэш (shared memory) 64 кб
- Производительность в DP – 515,2 GFLOPs на профессиональной карте (Nvidia Tesla C2070)

Архитектура Fermi



Архитектура Fermi



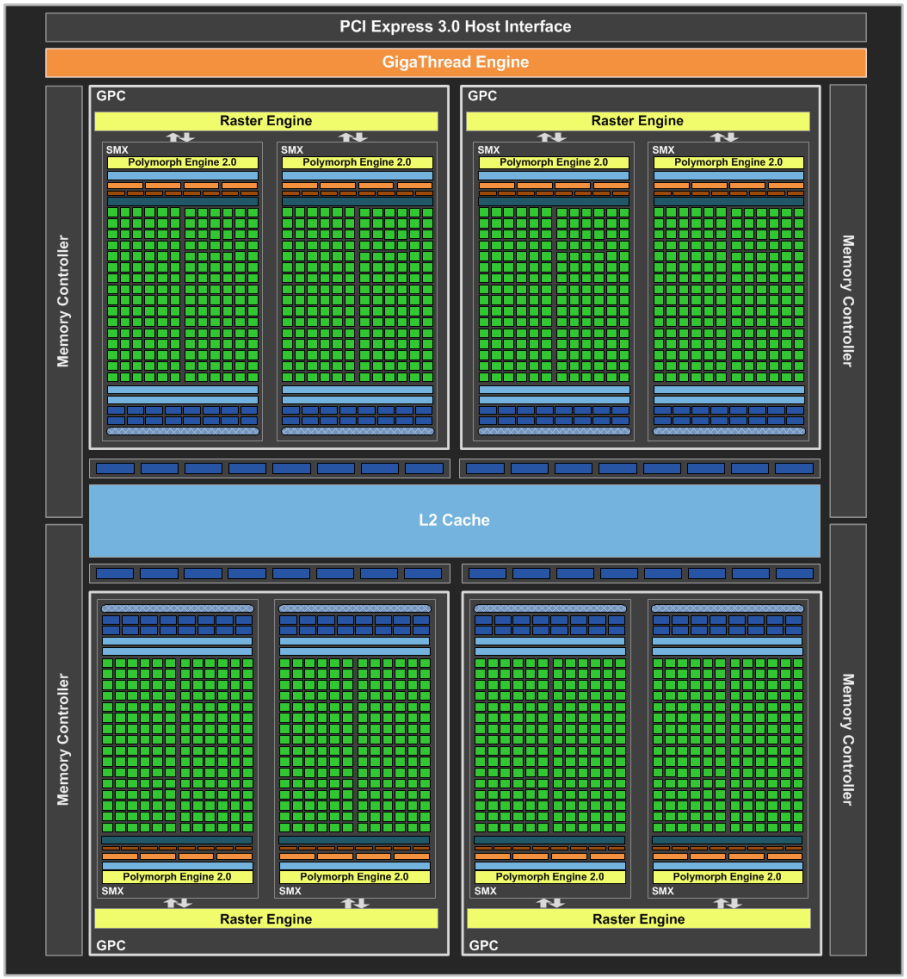


Архитектура Kepler

- Дата релиза: 2012 год
- Игровые карты семейств GeForce 600-700
- 15 NextGen SM (SMX) – каждый содержит 192 вычислительных ядра (SP)
- 2880 ядер CUDA (Nvidia Tesla K40)
- SM-кэш/L1-кэш (shared memory) 64 кб
- Производительность в DP – 1430 GFLOPs (Nvidia Tesla K40)
- FP32/FP64 = 1/5

Архитектура Kepler

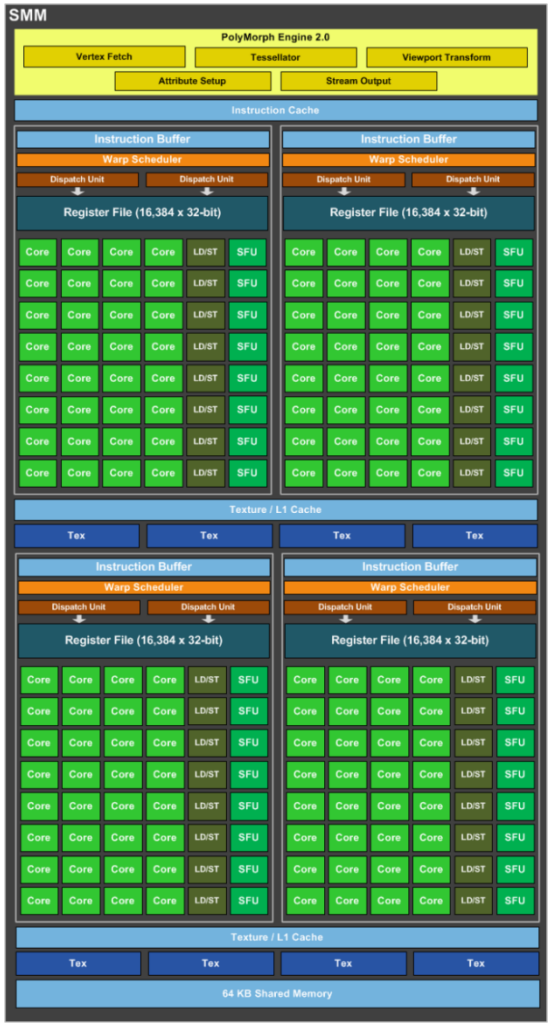
Архитектура Kepler



Архитектура Maxwell

- Дата релиза: 2014 год
- Игровые карты семейств GeForce 800-900, GeForce Titan
- 16 Streaming Multiprocessor (SMM) – каждый содержит 128 вычислительных ядер CUDA
- Shared memory 96 кб
- Texture-кэш/L1-кэш 64 кб
- Аппаратная поддержка DP урезана – FP64/FP32 = 1/32

Архитектура Maxwell



Архитектура Maxwell

Архитектура Pascal

- Дата релиза: 2016 год
- Игровые карты семейств GeForce 10
- 6 GPC по 10 SM в каждом
- Каждый SM содержит 64 SP
- 3840 ядер CUDA
- Вновь появились TPC – объединяют 2 SM и входят в GPC
- Вновь добавлена полноценная поддержка DP (FP64/FP32 = 1/2)

Архитектура Pascal

Архитектура Pascal

Казанский федеральный университет

Казанский федеральный
УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА
информационных технологий
и информационных систем

Вопросы

ekhramch@kpfu.ru