Основы технологии CUDA

Лекция 2:Архитектура графических адаптеров Nvidia

Автор курса:

- Казённов Андрей

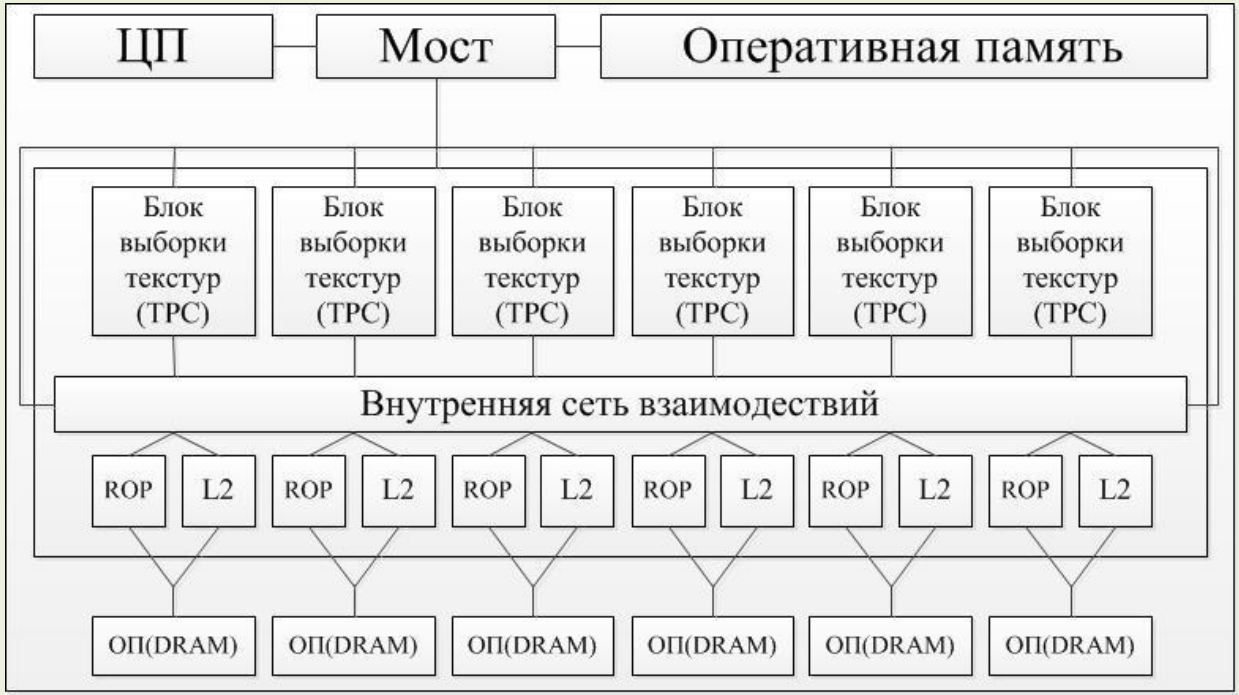
Введем основные термины курса

- **Хост (Host)** – центральный процессор, который управляет выполнением программы.
- **Устройство (Device)** – видеокарта, являющаяся сопроцессором к центральному процессору (хосту).

Введем основные термины курса

- **Ядро (Kernel)** – Параллельная часть алгоритма, выполняется на гриде.
- **Грид (Grid)** – объединение блоков, которые выполняются на одном устройстве.
- **Блок (Block)** – объединение потоков, которое выполняется целиком на одном SM. Имеет свой уникальный идентификатор внутри грида.
- **Тред (Thread)** – единица выполнения программы. Имеет свой уникальный идентификатор внутри блока.
- **Варп (Warp)** – 32 последовательно идущих треда, выполняется физически одновременно.

Общая схема устройства графического адаптера



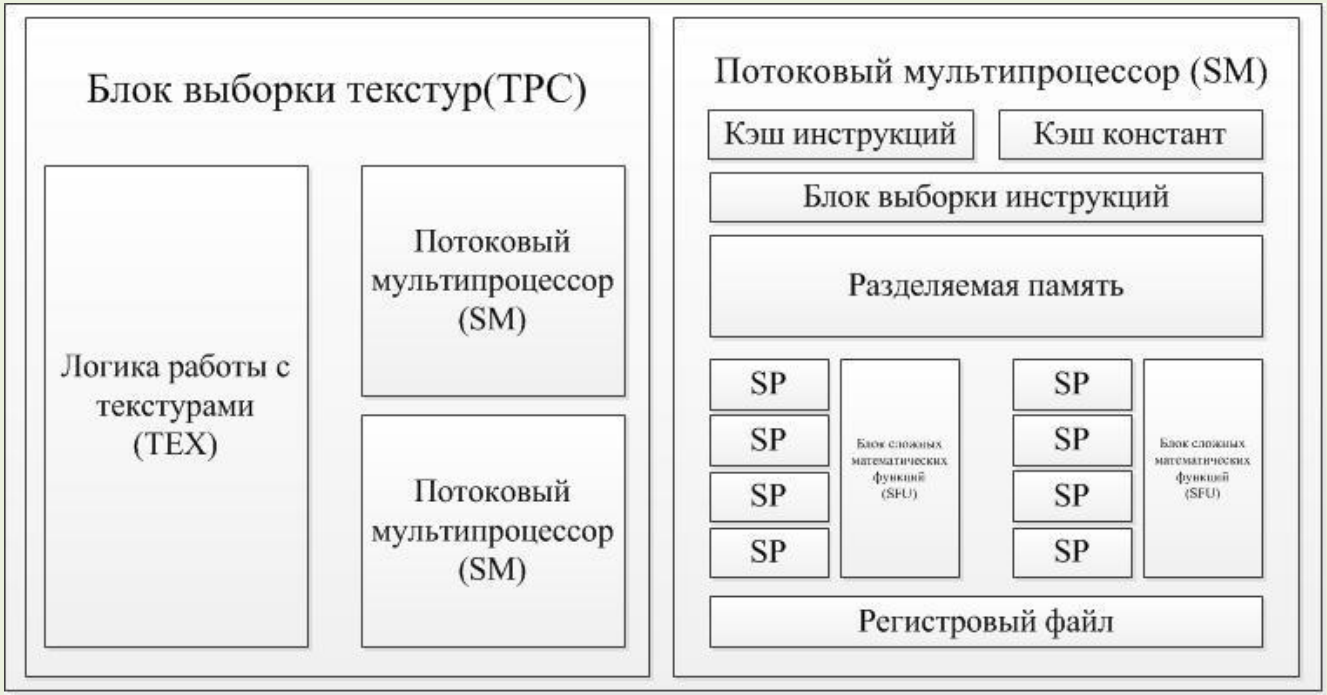
Основы технологии CUDA

Устройство чипа графического процессора

- Чип G80
- Чип G200
- Чип Fermi

Основы технологии CUDA

Чип G80



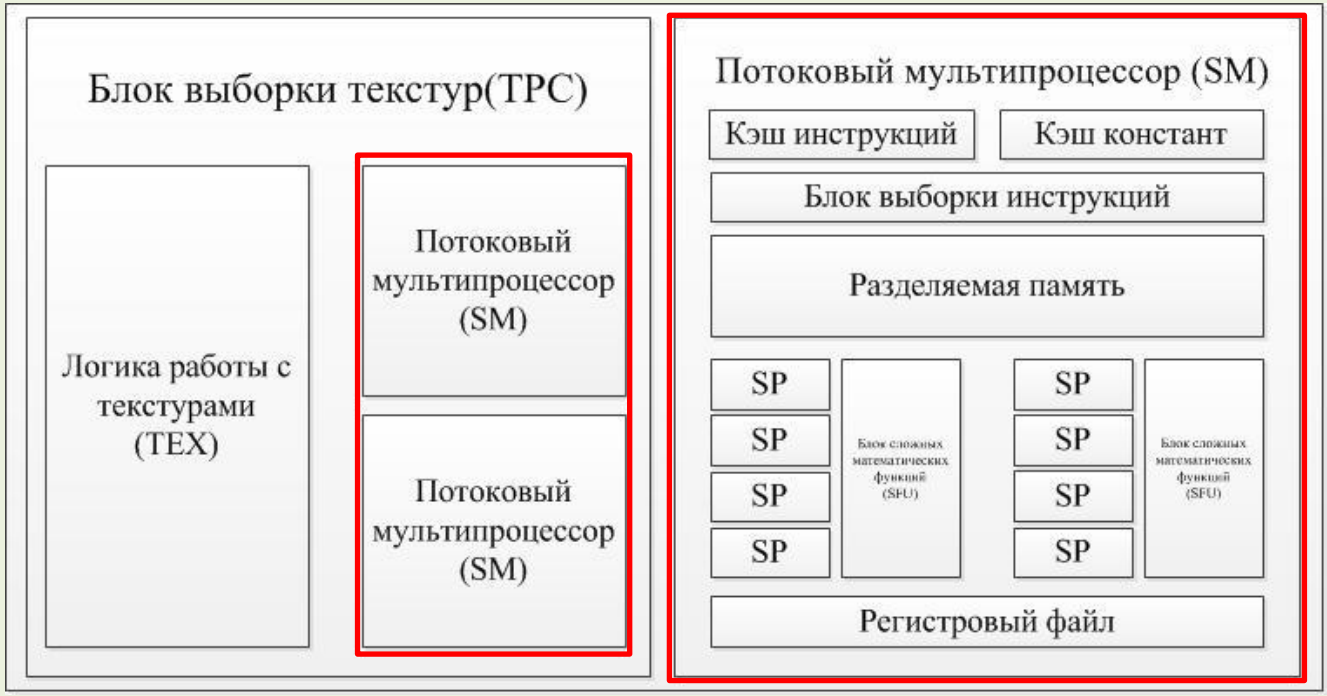
Детали аппаратной архитектуры G80: TPC
Блок выборки текстур

- Самостоятельная архитектурная единица
- Содержит в себе логику работы с текстурами
- Содержит несколько Потоковых мультипроцессоров

Детали аппаратной архитектуры G80: SM
Потоковый мультипроцессор

Самостоятельная вычислительная единица

➢**Место выполнения блока**



Детали аппаратной архитектуры G80: SP
Потоковый процессор

- Непосредственно вычислительный модуль
- Умеет совершать арифметические операции с целочисленными операндами и с операндами с плавающей точкой **(одинарная точность)**

Детали аппаратной архитектуры G80: SFU
Блок обработки сложных функций

- Проводит вычисления сложных математический функций (exp, sqr, log)
- Использует вычислительные мощности SP

Детали аппаратной архитектуры G80: Reg
Регистровый файл

- Место хранения локальных переменных
- Регистры выделяются отдельно на каждый тред

Детали аппаратной архитектуры G80: SM
Разделяемая память

- Особый тип памяти, позволяющий совместное использование данных всеми тредами отдельного блока

Детали аппаратной архитектуры G80: Логика

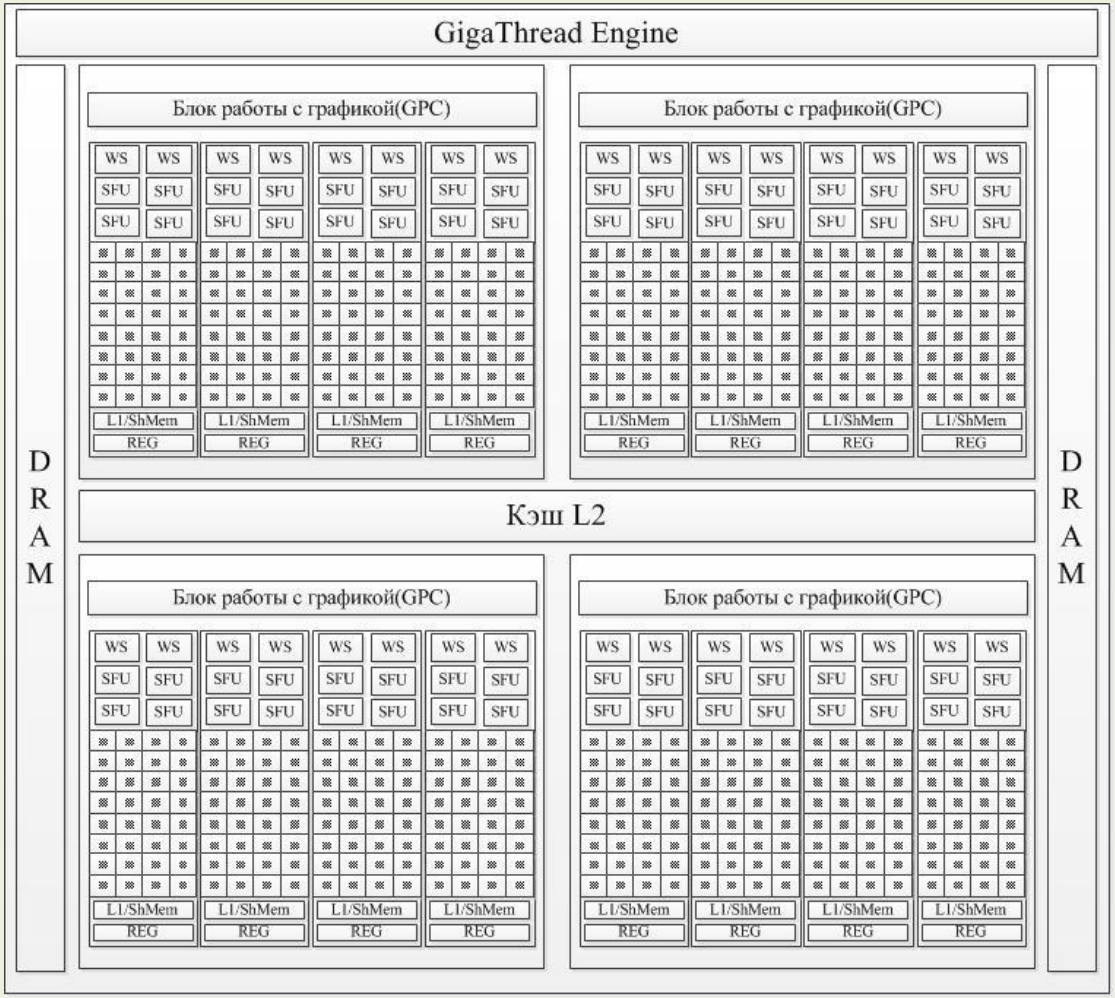
- Система управления SM

Детали аппаратной архитектуры G200: Double
Блок расчета с двойной точностью

- Использует вычислительные мощности SP

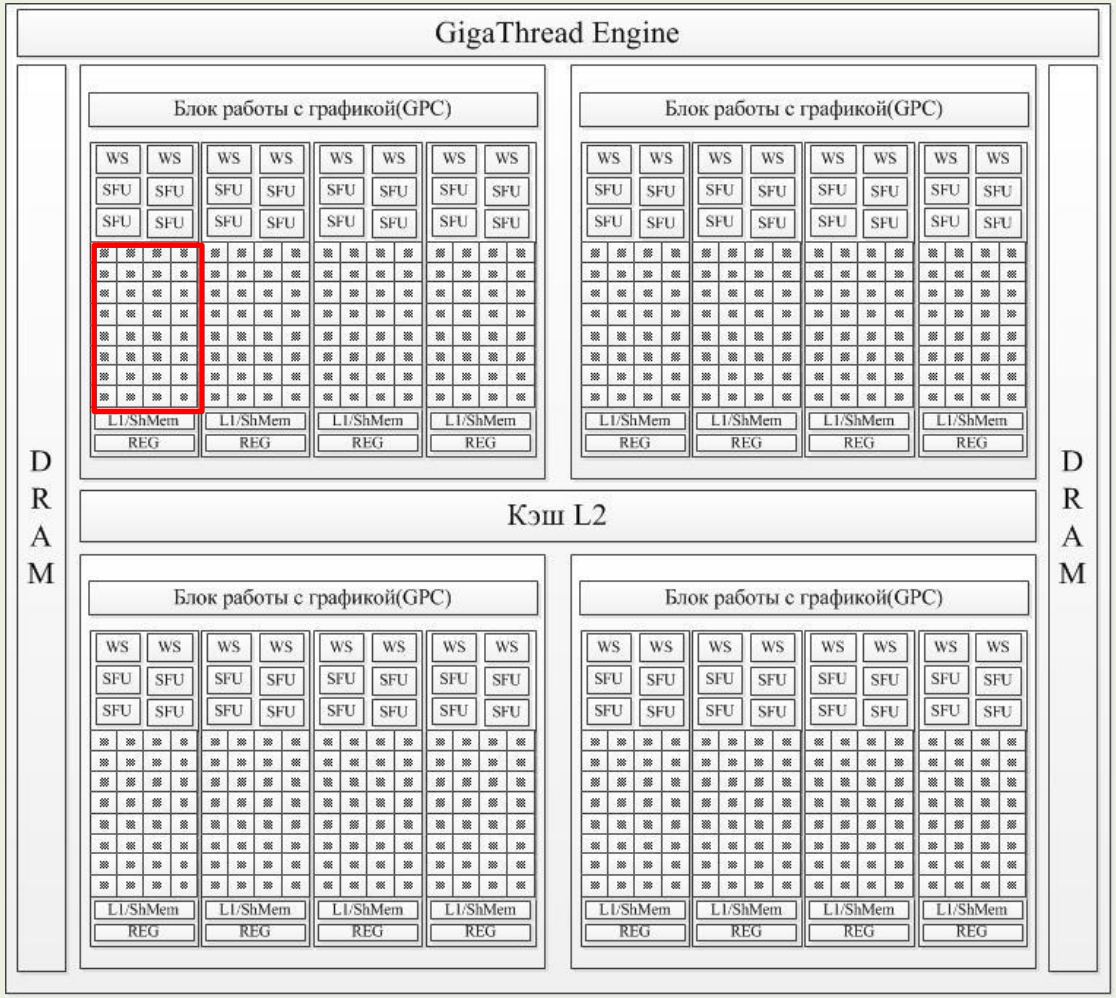
Основы технологии CUDA

Чип Fermi



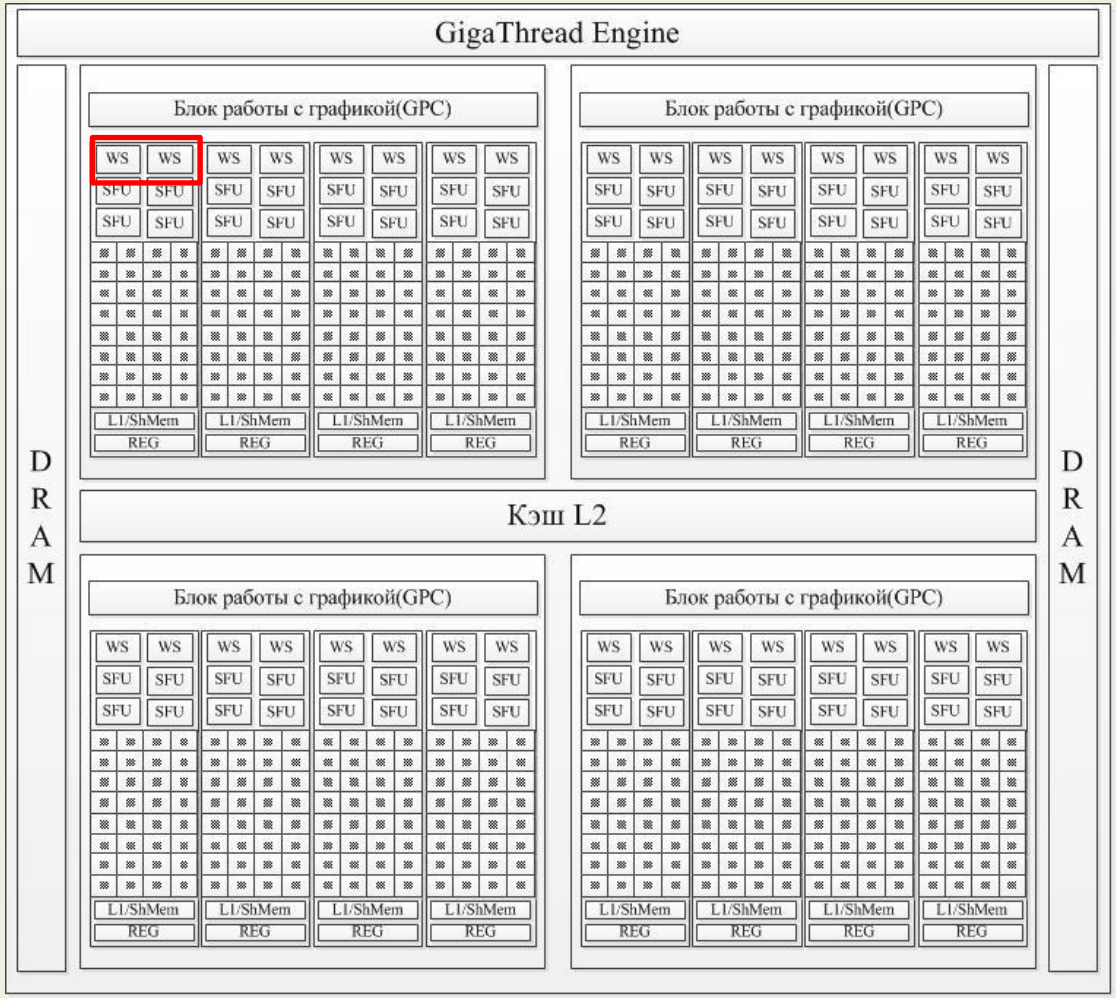
Детали аппаратной архитектуры G200: SP

- 32 SP на SM
- SP считает с двойной точностью



Детали аппаратной архитектуры G200: WS

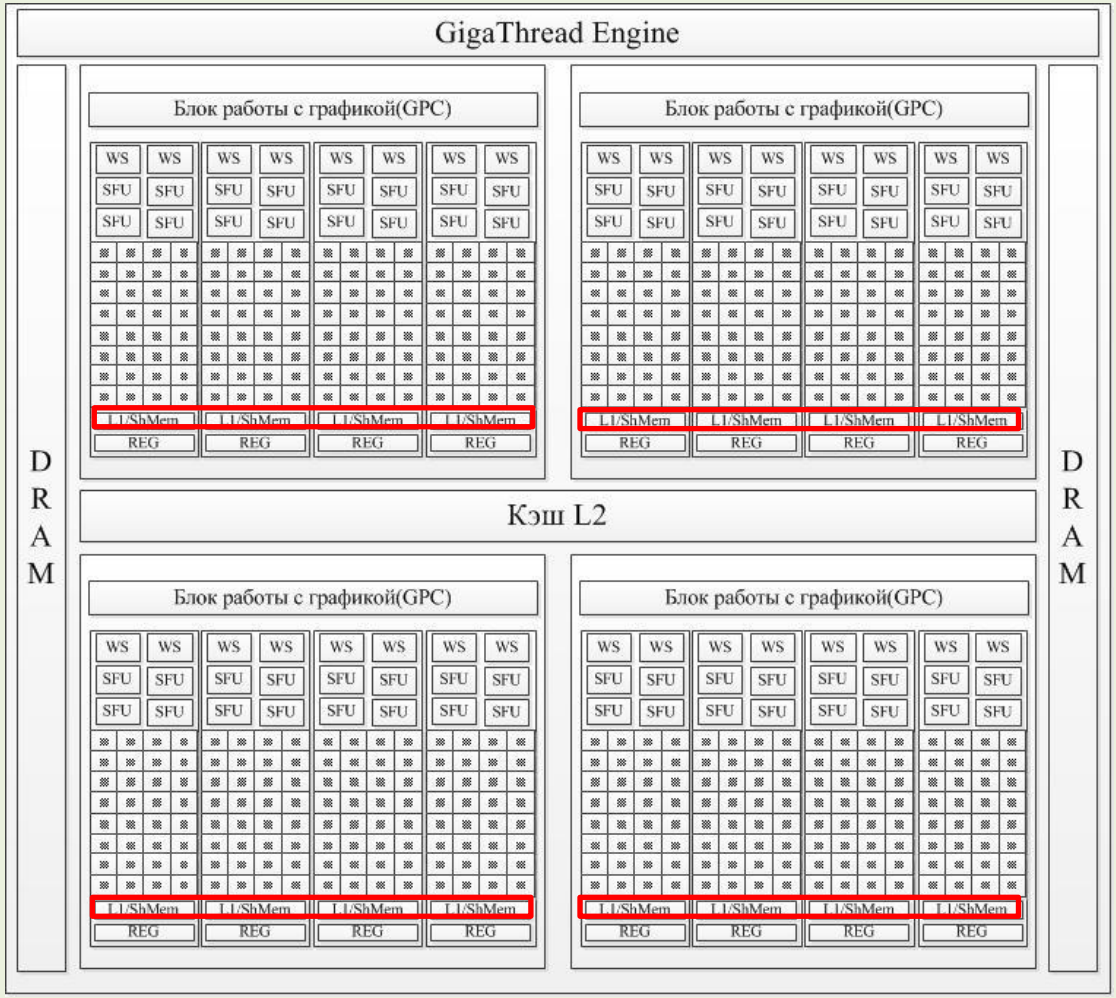
- 2 планировщика варпа на SM
- Одновременное выполнение 2х варпов на SM



Основы технологии CUDA

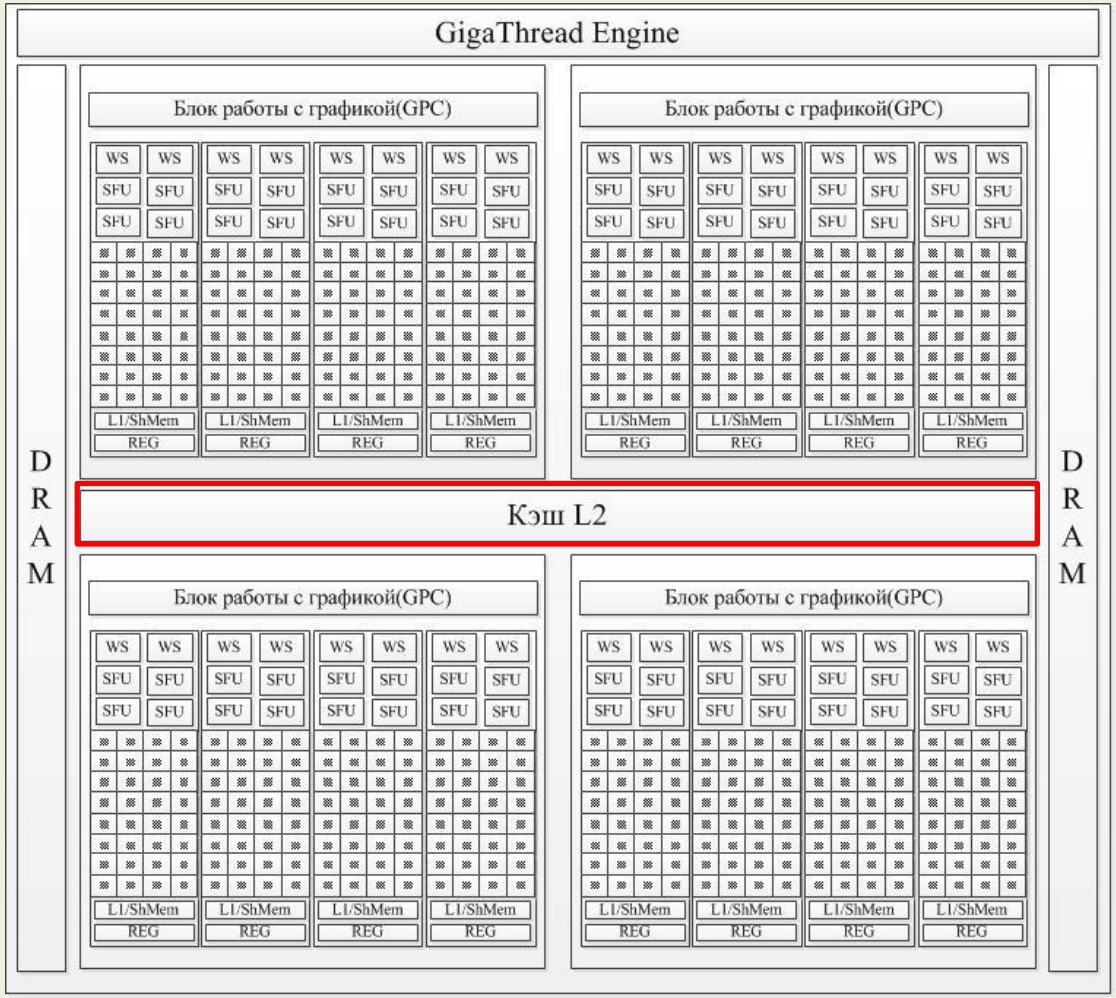
Детали аппаратной архитектуры G200: L1/ShMem

- Кэш 1го уровня на каждом SM
- Банк из 64 Кб разделяемый между ShMem и L1
- 48/16 Кб или 16/48 Кб



Детали аппаратной архитектуры G200: L2

- Кэш 2го уровня на весь кристалл
- 768Кб



Сравнение характеристик различных чипов

Архитектура	G80	GT200	Fermi
Год вывода на рынок	2006	2008	2009
Число транзисторов, млн.	681	1400	3000
Количество CUDA-ядер	128	240	512
Объем разделяемой памяти на SM, Кб	16	16	48 или 16 (конфигурируется)
Объем кэш-памяти первого уровня в расчете на SM, Кб	0	0	16 или 48 (конфигурируется)
Объем кэш-памяти второго уровня, Кб	0	0	768
Функция ECC	нет	нет	есть

Compute Capability

- Разные поколения графических карт обладали различными возможностями
- Возможности конкретной карты можно определить по Compute Capability
- Стандартный способ узнать Compute Capability – использовать программу **nvidia-smi** из **Nvidia Toolkit**

Compute Capability

Compute Capability	Возможности
1.0	Базовые возможности
1.1	Добавлены атомарные операции с Int
1.2	Изменены правила работы с глобальной памятью
1.3	Добавлена работа с двойной точностью
2.0	Добавлены атомарные операции с float

Compute Capability

GPU	Compute Capability
GeForce 8800GTX	1.0
GeForce 9800GTX	1.1
GeForce 210	1.2
GeForce 275GTX	1.3
Tesla C2050	2.0

Ресурсы курса

Сайт: HPC.MIPT.RU

Раздел образование

Автор курса:

Казённов Андрей

- E-mail: kazennov@gmail.com
- ICQ: 622774