Казанский федеральный университет

ВЫСШАЯ ШКОЛА информационных технологий и информационных систем

Введение в технологию CUDA

Compute Unified Device Architecture

- NVIDIA CUDA – фреймворк для написания кода исполняемого на GPU
- Набор расширений над C/C++ и Fortran
- Существуют обертки для Python, Java, etc.
- Для чего нужна – в некоторых задачах CUDA позволяет на несколько порядков повысить быстродействие кода
- Области применения: научно-инженерные расчеты, финансы и аналитика, машинное обучение и обработка данных

Описание курса

- Цель курса – научится применять API NVIDIA CUDA в своем коде
- Репозиторий курса на Github
- Итог – решение конкретной задачи с использованием CUDA
- Итоговая оценка за семестр

$$R_{general} = \frac{R_{cuda} + R_{opencl}}{2}$$

Технические вопросы

- Что нужно для программирования с CUDA
  - PC с GPU NVIDIA
  - Установленный компилятор C/C++
  - Свежий драйвер
  - Установленный и настроенный NVIDIA CUDA Toolkit
- В курсе используются язык C++, ОС Linux и строчный компилятор (gcc/nvcc)
- NVIDIA CUDA также поддерживает Windows и MacOS

Технические вопросы

- Список поддерживаемых видеокарт
- Каждая версия CUDA Toolkit имеет свой список актуальных драйверов
- Если GPU устарел, можно установить старую версию CUDA Toolkit

Технические вопросы

- Алгоритм
  - Определить модель GPU
  - Определить последнюю версию CUDA для данного GPU
  - Поставить последний драйвер поддерживающий данную версию CUDA
  - Найти соответствующую документацию CUDA и следовать инструкциям

Кластер КФУ

- Можно выполнять задания на кластере КФУ
- IP и порт у преподавателя
- Логин и пароль у преподавателя
- Работает только из сети КФУ
- ОС – Linux RedHat
- Версия CUDA 7.5
- Версия gcc 4.8.5

Кластер КФУ

- Соединение по SSH на управляющий узел, оттуда на узел с GPU
- Узлы с GPU:
  - bmk-x2-a1-ch1-10 – 2x NVIDIA Tesla K80
  - bmk-x2-a1-ch1-9 – 2x NVIDIA Tesla K80
  - bmk-x4-a1-ch1-8 – 4x NVIDIA Tesla K80
  - bmk-x4-a1-ch1-7 – 4x NVIDIA Tesla K80

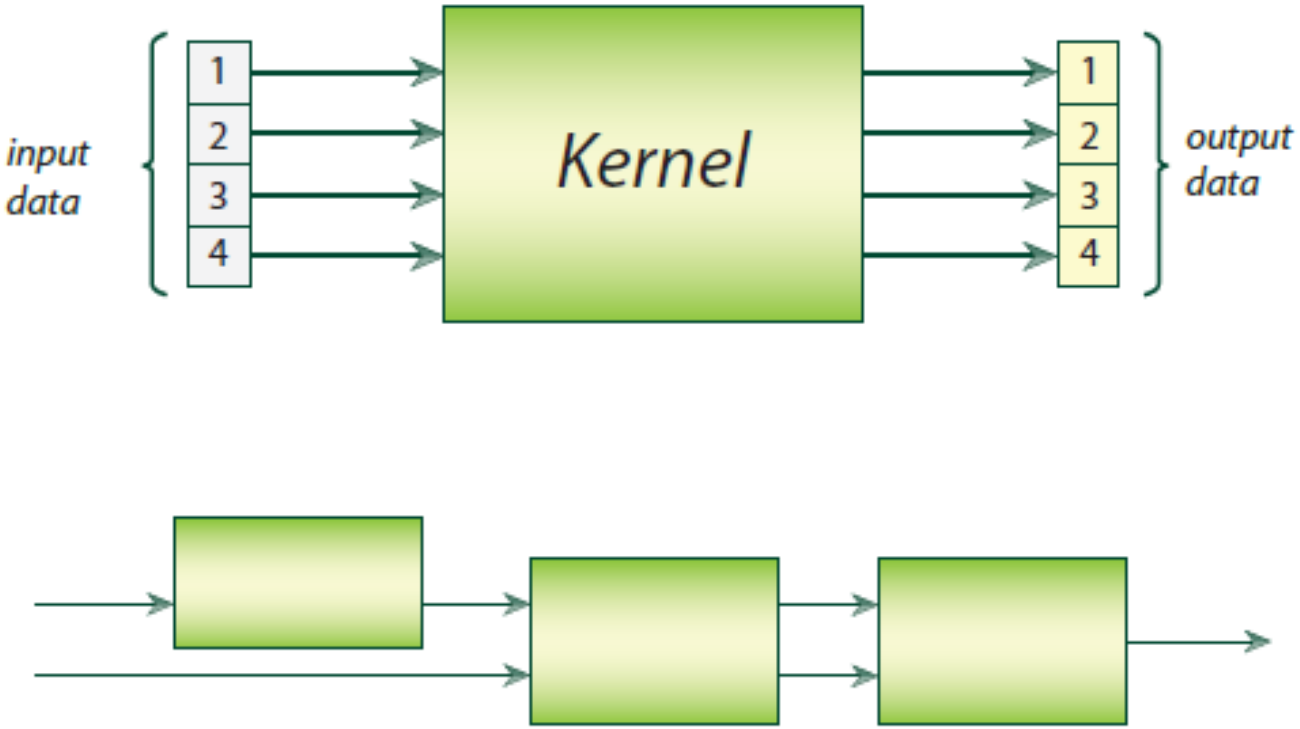
Особенности GPU

- Архитектура GPU относится к классу SIMD – Single Instruction Multiple Data
- GPU создавались для рендеринга изображений
- GPU эффективны для решения задач параллельных по данным
- Количество арифметических операций >> количество операций с памятью

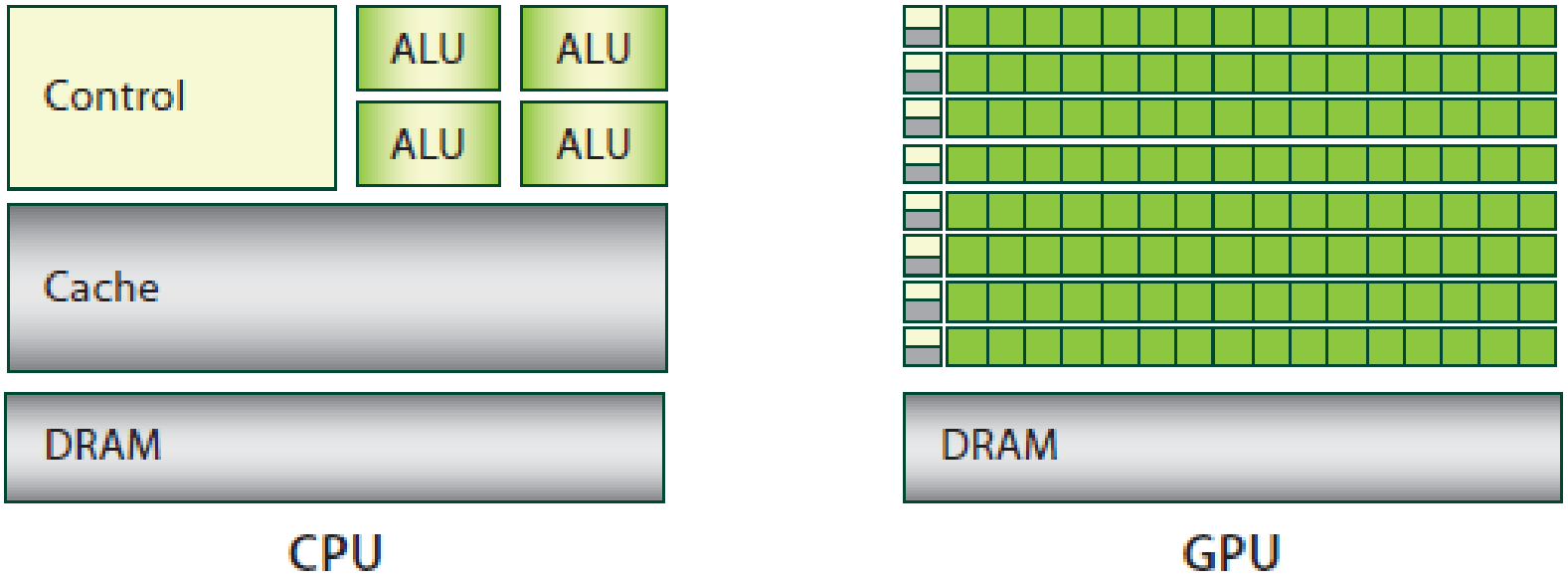
Особенности GPU

- Несколько типов с плавающей точкой:
  - Single precision (float32) – достаточная точность для 3d графики
  - Double precision (float64) – точность для научно-технических расчетов
  - Half precision (float16) – появились в последних архитектурах для тензорных ядер машинного обучения

SIMD



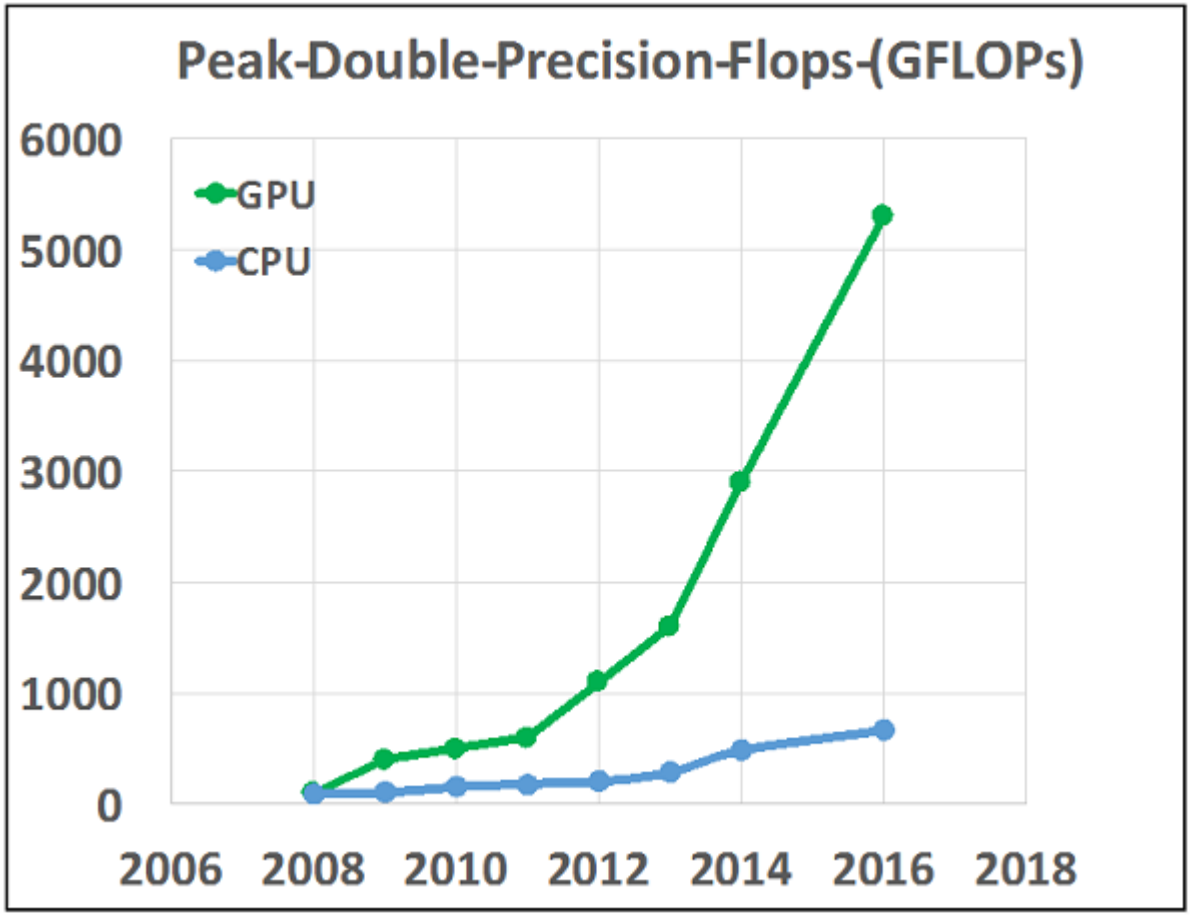
GPU vs. CPU: архитектура



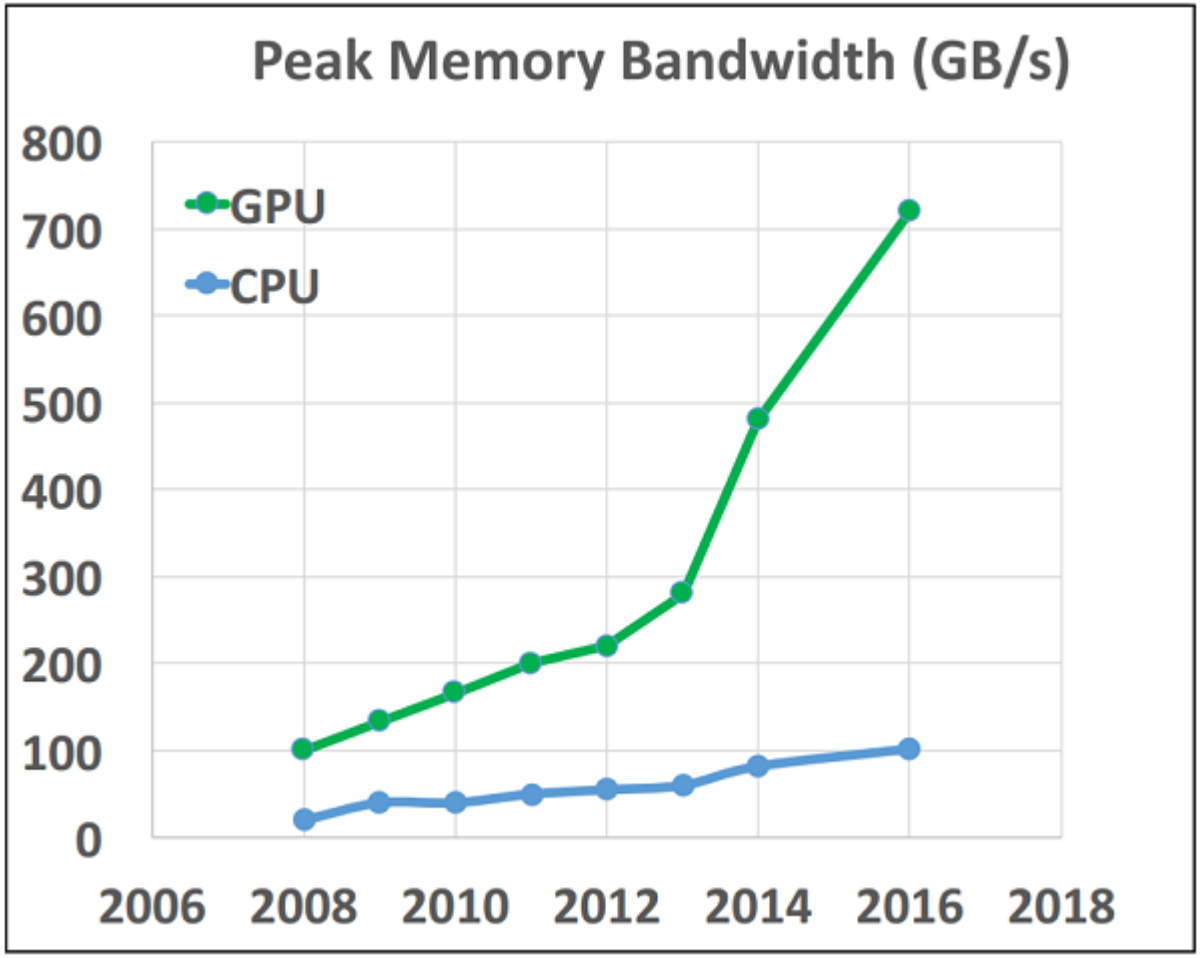
GPU vs. CPU: архитектура

Характеристика	CPU	GPU
Количество ядер	<20	>100
Тактовая частота ядра	Высокая	Низкая
Ветвление и векторизация	Есть	Нет
Кеш	Большой	Незначительный

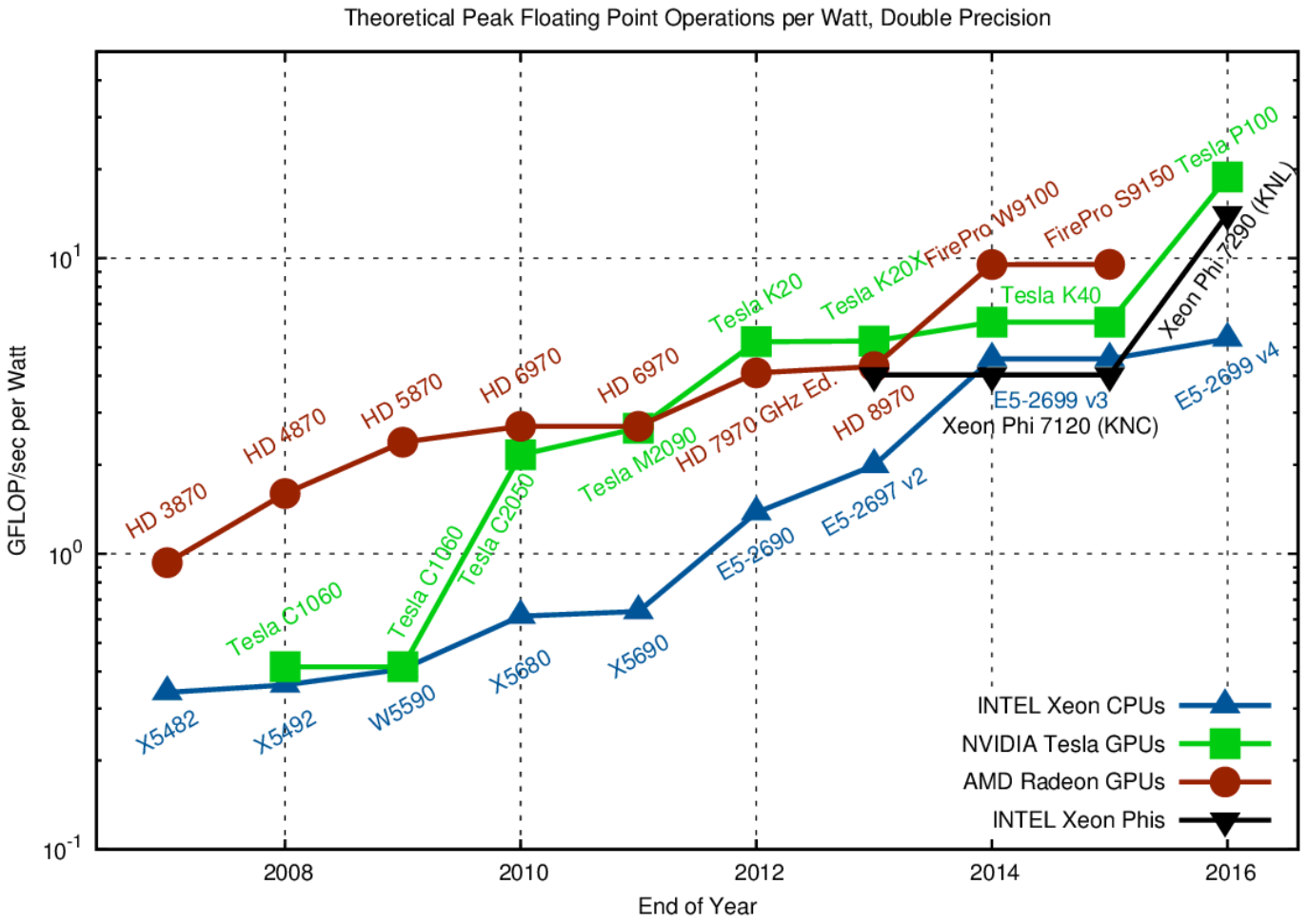
GPU vs. CPU: производительность



GPU vs. CPU: скорость памяти



GPU vs. CPU: энергоэффективность



Казанский федеральный университет

API

- Видеокарты от AMD и NVIDIA используют разные API для параллельных вычислений
- Карты AMD поддерживают только фреймворк OpenCL
- Карты NVIDIA оптимизированы под фреймворк CUDA – проприетарный продукт NVIDIA
- Карты NVIDIA так же поддерживают API OpenCL

Архитектура GPU

- Низший уровень архитектуры – потоковые процессоры (Streaming Processor, SP)
- Каждый SP – микропроцессор с очередным типом исполнения команд
- У SP нет кэш-памяти, эффективен для большого количества математических расчетов

Архитектура GPU

- SP объединены в потоковые мультипроцессоры (Streaming Multiprocessor, SM)
- SM – массив из нескольких SP
- В каждом SM есть SP для работы с float(fp32) и double(fp64)
- В SM содержится кэш общий для всех его SP

Архитектура GPU

- Следующим уровнем объединения является кластер из нескольких SM
- Такие кластеры носят различные названия в разных архитектурах
- Кластеры служат для улучшения баланса загрузки и оптимизации управления ресурсами

Архитектура GPU

- NVIDIA выпускает игровые (GeForce) и профессиональные (Tesla) видеокарты
- На игровых GPU завышены частоты памяти и процессора но урезано количество SP для двойной точности
- Соотношение процессоров fp64 к fp32 колеблется от ½ до 1/32 в зависимости от архитектуры и типа карты
- На профессиональных GPU устанавливается больше памяти

Архитектура GPU

Архитектура	Год релиза	Игровые карты	Профессиональные карты	Ядер CUDA(SP)
Tesla	2006	GeForce 8-9 GeForce 100-300	Tesla C10	240
Fermi	2009	GeForce 400-500	Tesla C20	512
Kepler	2012	GeForce 600-700	Tesla K	2880
Maxwell	2014	GeForce 800-900	-	3072
Pascal	2016	GeForce 10	Tesla P	3840
Volta	2017	-	Tesla V	5384
Turing	2018	GeForce 20	?	4352

Казанский федеральный
УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА
информационных технологий
и информационных систем

Вопросы

ekhramch@kpfu.ru