ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

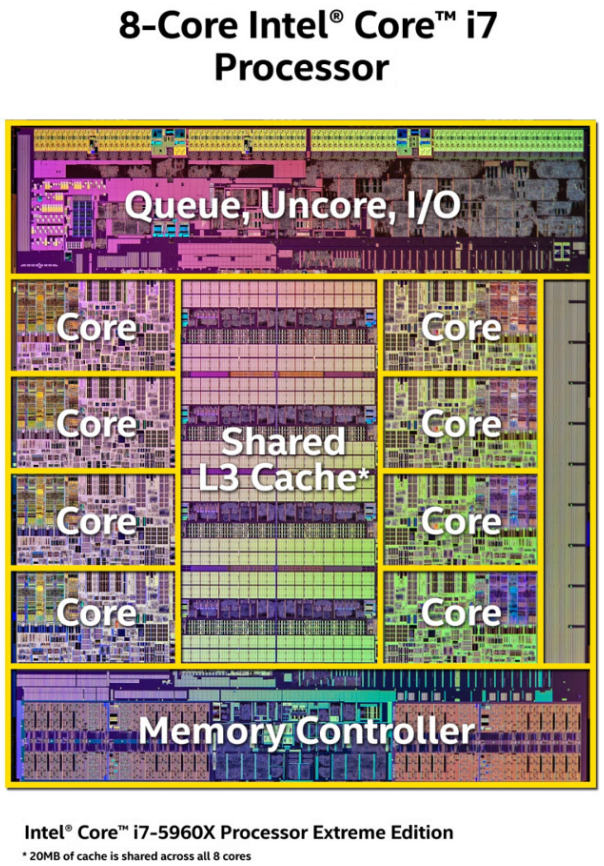
МИЭМ

Параллельные вычисления на графических адаптерах nVidia в задаче обучения нейронных сетей

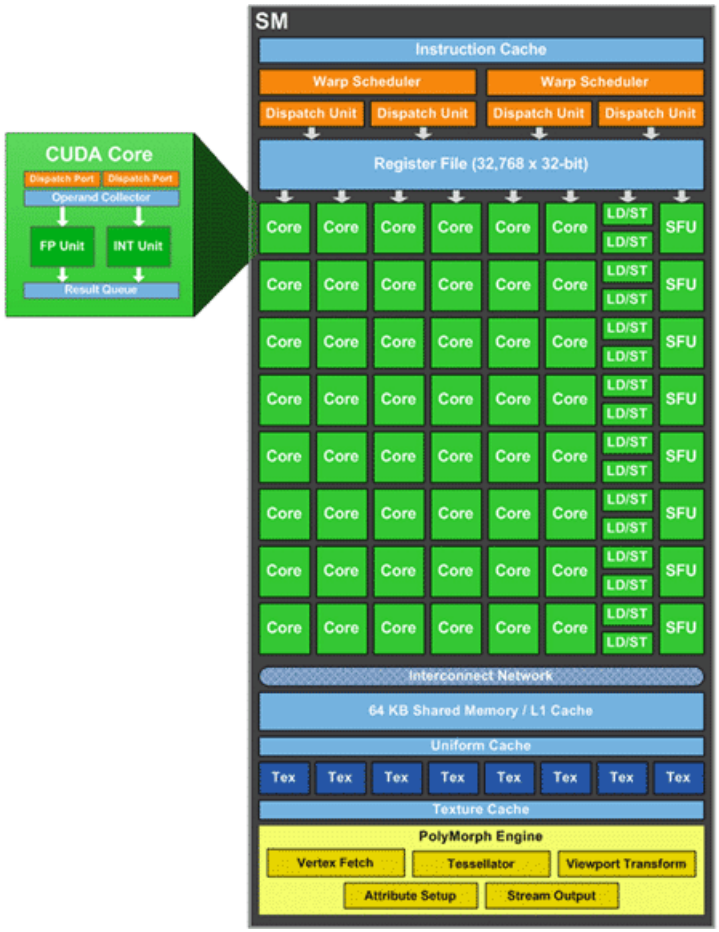
Автор работы:
Федин Н. А., аспирант 1 г.о.
Научный руководитель:
к.т.н., профессор
Истратов Анатолий Юрьевич

Сравнение архитектур CPU и GPU

- Небольшое количество ядер (от 2 до 8)
- Максимальные частоты отдельного ядра (2,3 – 4 ГГц)
- Ядра независимы по инструкциям и данным
- Небольшое количество планируемых потоков
- Переключение контекста является затратной процедурой



Потоковый мультипроцессор (Stream Multiprocessor)



- Десятки более простых ядер (ALU + FPU) на мультипроцессор (обычно кратно 32)
- В пределах мультипроцессора группы ядер реализуют SIMD выполнение команд
- Более низкие частоты работы отдельных ядер (1 – 1,5 ГГц)
- Сотни потоков, планируемых одним мультипроцессором. Переключение контекста не снижает производительности

Общий вид GPU

- 16 SM
- 512 - 640 ядер
- Кеш L2 (около 1-2 Mb)
- Контроллеры памяти

Программная модель CUDA

- CUDA (англ. Compute Unified Device Architecture) – архитектура вычислений общего назначения на GPU компании nVidia.

  - Компилятор nvcc
  - Диалект языка программирования C
  - Набор библиотек, поддерживающих параллельные вычисления
  - Run Time API

Блоки и потоки в CUDA

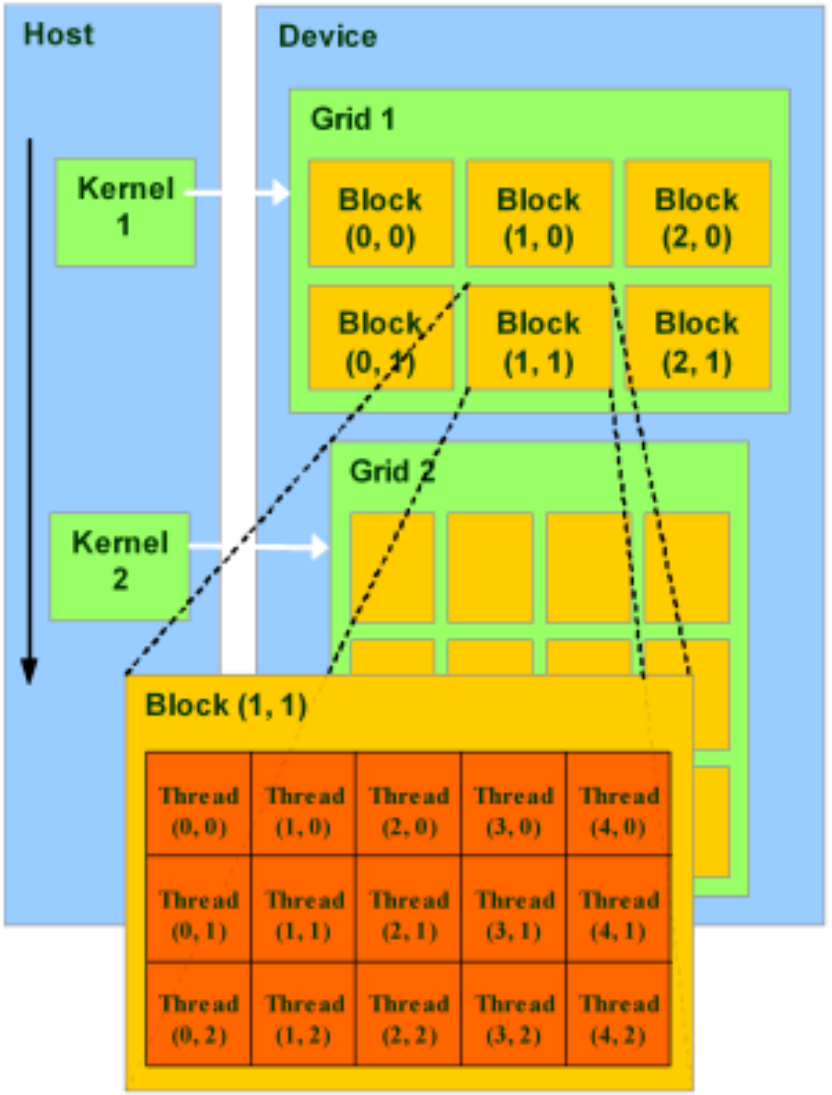
Поток – элементарный поток инструкций, имеющий собственные регистры, ядро на мультипроцессоре и контекст

Блок – совокупность потоков, в пределах которой возможен обмен информацией с помощью разделяемой памяти.

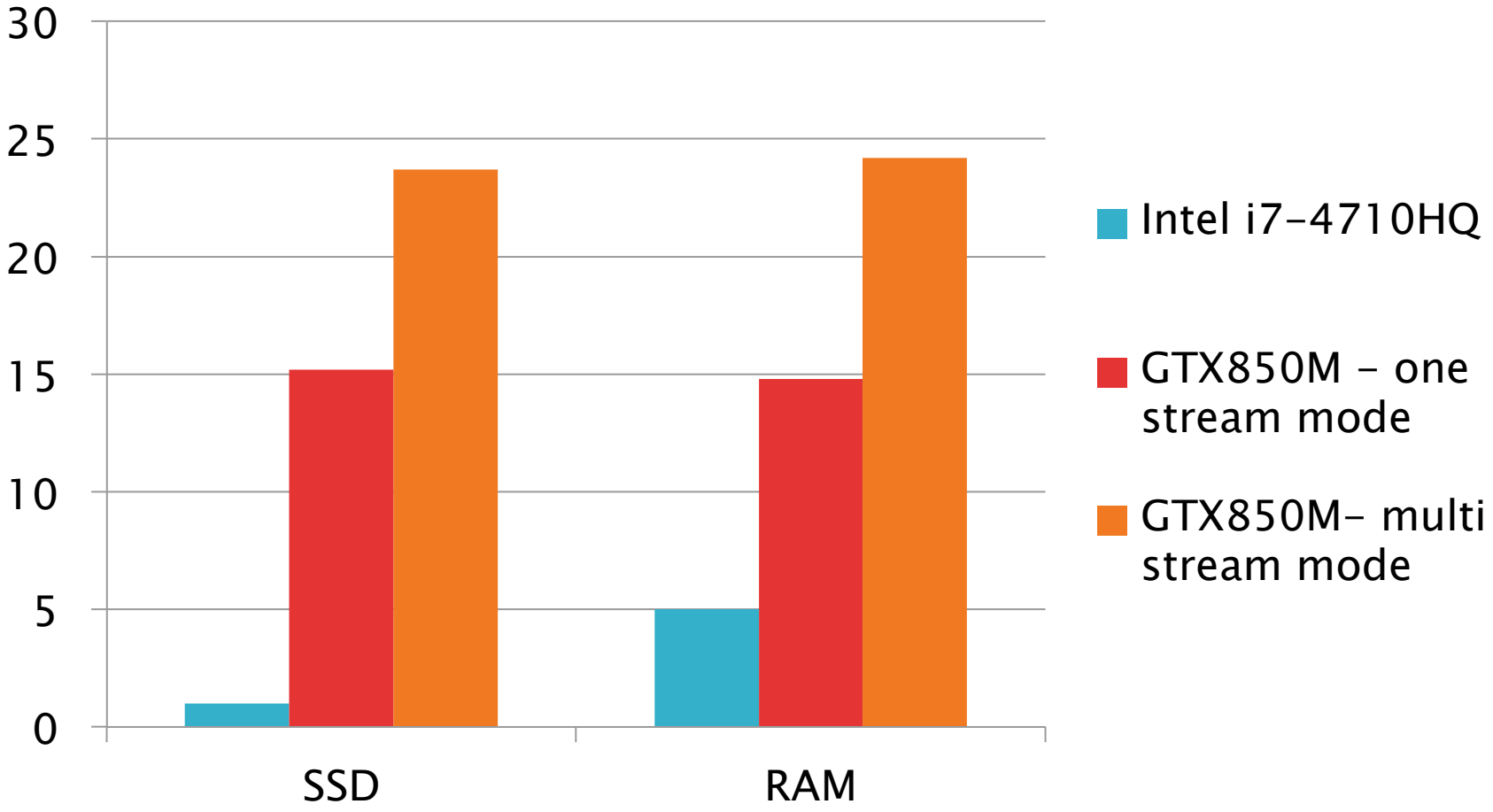
Варп (warp) – группа потоков в блоке, реализующая SIMD исполнение команд.

Сетки блоков и потоков

Блоки и потоки образуют не более чем трёхмерную сетку, в которой получают свои координаты. Общее количество потоков в блоке около 1000, общее количество блоков в сетке – порядка миллиона.



Эффект применения GPU



Использованные инструменты и оборудование

- Microsoft Visual Studio Ultimate 2013
- nVidia cuda toolkit v7.5.18
- nVidia driver 353.90
- MS Windows 8.1
- GeForce GTX 850M (4 Gb ddr3 RAM, GM107)
- CPU i7-4710HQ
- SSD Plextor M6Pro 512Gb, SATA III.

Заключение

1. Изучена технология параллельного программирования и архитектура GPU
2. Разработано ПО для обучения нейронных сетей
3. Достигнуто десятикратное сокращение времени обучения для эпохи

Полезные ссылки

- http://www.nvidia.ru/object/cuda-parallel-computing-ru.html

- Параллельные вычисления на GPU. Архитектура и программная модель CUDA: Учеб. Пособие / А.В. Боресков и д.р. – М.: Издательство Московского университета, 2012. – 336 с., илл.

- Документация к CUDA Toolkit v 7.5

Спасибо за внимание!