

Основы ускорения вычислений на GPU в Matlab

Докладчик: к.т.н., доцент кафедры радиотехнических и телекоммуникационных устройств, ИРТСУ ЮФУ г. Таганрог.

Ведущий инженер АО «Региональный межотраслевой центр информации и технологий», г. Ростов-на-Дону.

Михаил Потипак

E-mail: potipak@sfedu.ru

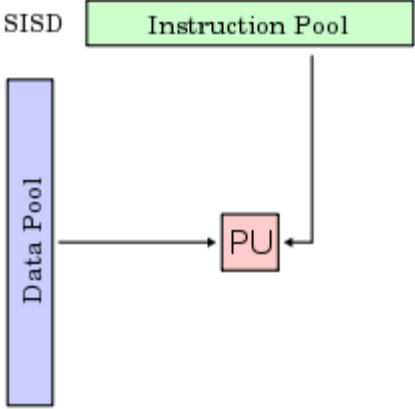
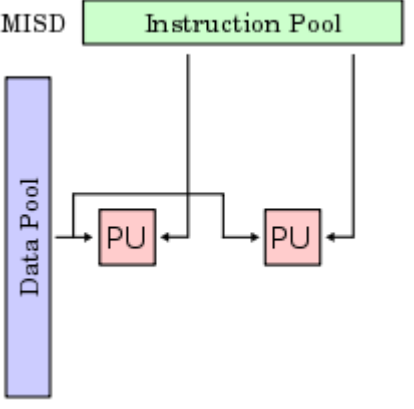
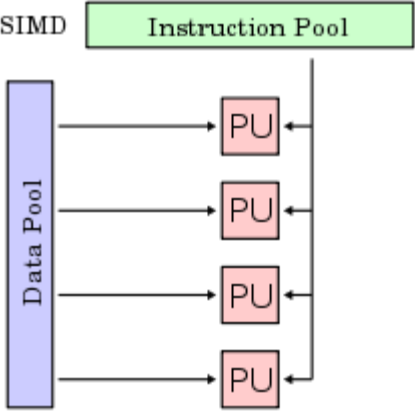
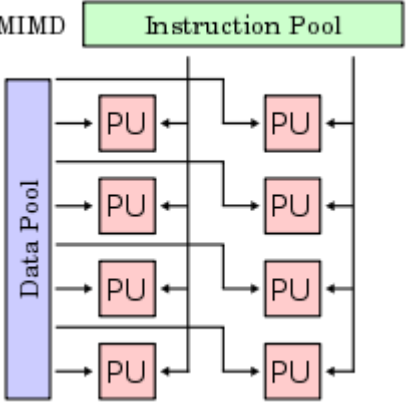
GPGPU

General-Purpose Graphics Processing Units (2003 г.) – («GPU общего назначения») – техника использования графического процессора видеокарты для общих (неграфических) вычислений, которые обычно проводит центральный процессор.

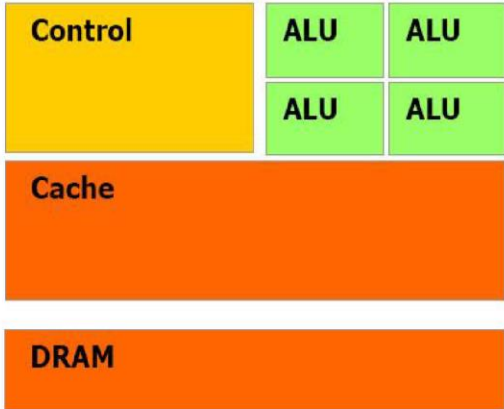
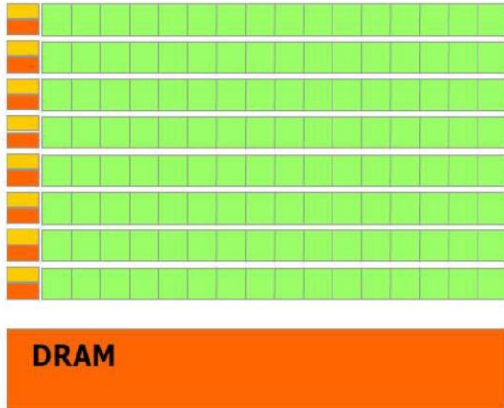
Применение GPGPU:

- Вычислительная математика
- Вычислительная биология
- Вычислительная экономика
- Моделирование в физике
- Обработка сигналов...

Классификация по Флинну

	Одиночный поток команд (single instruction)	Множество потоков команд (multiple instruction)
Одиночный поток данных (single data)	<p style="text-align: center;"><u>SISD</u> (ОКОД)</p> 	<p style="text-align: center;"><u>MISD</u> (МКОД)</p> 
Множество потоков данных (multiple data)	<p style="text-align: center;"><u>SIMD</u> (ОКМД)</p> 	<p style="text-align: center;">MIMD (МКМД)</p> 

Требования к оборудованию

CPU	GPU
<ul style="list-style-type: none">• Память оптимизирована под минимальную латентность (система “кэшей”).• Много транзисторов “управления” (предсказание ветвлений, планировщики и пр.).• Архитектура оптимизирована для программ со сложным управлением (эффективная обработка ветвлений).  <p>The diagram illustrates the CPU architecture. It features a yellow 'Control' block at the top left. To its right are four green 'ALU' blocks arranged in a 2x2 grid. Below the 'Control' block is a large orange 'Cache' block. At the bottom is another large orange 'DRAM' block. The entire diagram is labeled 'CPU' at the bottom center.</p>	<ul style="list-style-type: none">• Память оптимизирована под максимальную пропускную способность.• Большая часть транзисторов для вычислений.• Архитектура оптимизирована для программ с большим объемом вычислений (параллелизм по данным типа SIMD).• Латентность скрывается вычислениями во время запросов к памяти.  <p>The diagram illustrates the GPU architecture. It features a large grid of 10 rows and 20 columns of small green squares, representing parallel processing units. To the left of each row is a small orange square. Below the grid is a large orange 'DRAM' block. The entire diagram is labeled 'GPU' at the bottom center.</p>

GPU

демонстрируют хорошие результаты в параллельной обработке данных:

- с одной и той же последовательностью действий, применяемых к большому объёму данных (многопоточные вычисления), что подразумевает меньшие требования к управлению исполнением,
- с высокой плотностью арифметики - высоким отношением числа арифметических операций к числу обращений к памяти, что означает возможность покрытия латентности памяти вычислениями.

CUDA

Compute Unified Device Architecture (2007 г.) - новая программно-аппаратная архитектура NVIDIA для параллельных вычислений на GPU, предоставляющая средства (toolkit) для организации вычислений общего назначения на GPU

Присутствует в GPU NVidia:

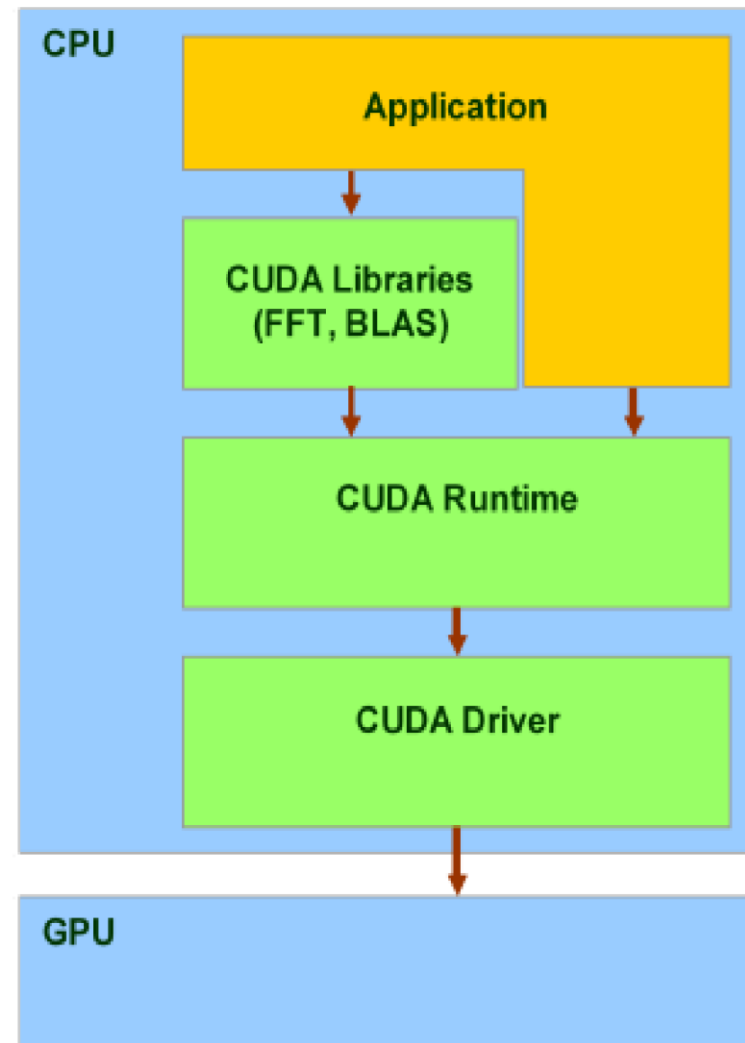
- GeForce 8800 и выше,
- Quadro FX 5600/4600 и выше,
- Tesla серии 10,
- Tesla серии 20 (Fermi).



http://www.nvidia.ru/object/cuda_home_new_ru.html

CUDA Toolkit

- компилятор nvcc;
- библиотеки CuFFT и CuBLAS;
- профилировщик;
- отладчик gdb для GPU;
- API высокого уровня (CUDA Runtime) и API низкого уровня (CUDA Driver);
- руководство по программированию;
- CUDA Developer SDK (исходный код, утилиты и документация).



http://developer.nvidia.com/object/cuda_3_2_downloads.html



Parallel Computing Toolbox

lets you solve computationally and data-intensive problems using multicore processors, GPUs, and computer clusters. High-level constructs - parallel for-loops, special array types, and parallelized numerical algorithms let you parallelize MATLAB applications without CUDA or MPI programming. You can use the toolbox with Simulink to run multiple simulations of a model in parallel.

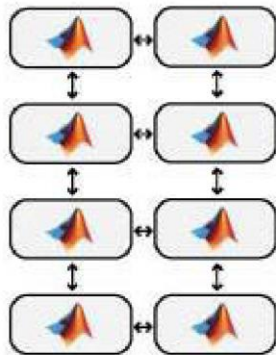
The toolbox provides eight workers (MATLAB computational engines) to execute applications locally on a multicore desktop. Without changing the code, you can run the same application on a computer cluster or grid (using MATLAB Distributed Computing Server).

<http://www.mathworks.com/products/parallel-computing/>

Multicore Desktop with GPUs

Parallel Computing Toolbox

Local Workers



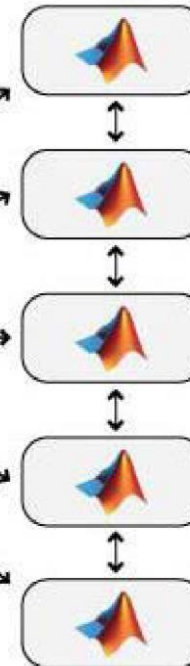
Simulink, Blocksets,
and Other Toolboxes

MATLAB

Computer Cluster

MATLAB Distributed Computing Server

Workers



Scheduler



Поддержка GPU в Parallel Computing Toolbox

- NVIDIA GPUs с вычислительной способностью 1.3 или выше

- включая Tesla 10-серий и 20-серий

- (напр., NVIDIA Tesla C2075 GPU: 448 процессоров, 6 Гб памяти)

- http://www.nvidia.com/object/cuda_gpus.html

Работа с GPU с версии R2010b



- Почему требуется вычислительная способность 1.3

- Поддерживает doubles (базовый тип данных в MATLAB)

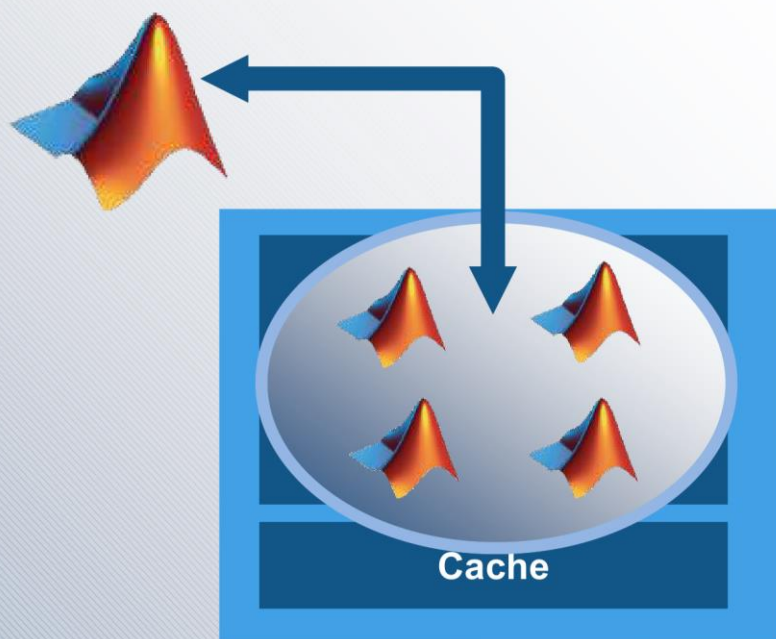
- Операции соответствуют стандарту IEEE

- Поддержка кроссплатформенности

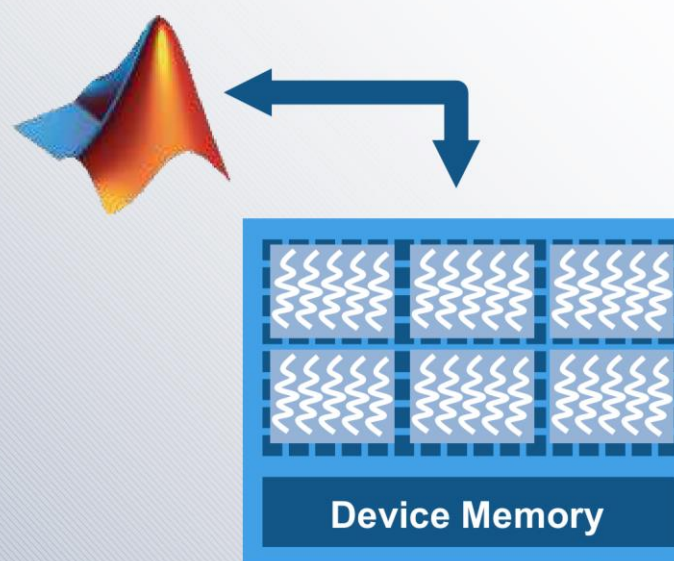
Быстро развивается –
используйте
последнюю версию

Возможности увеличения производительности

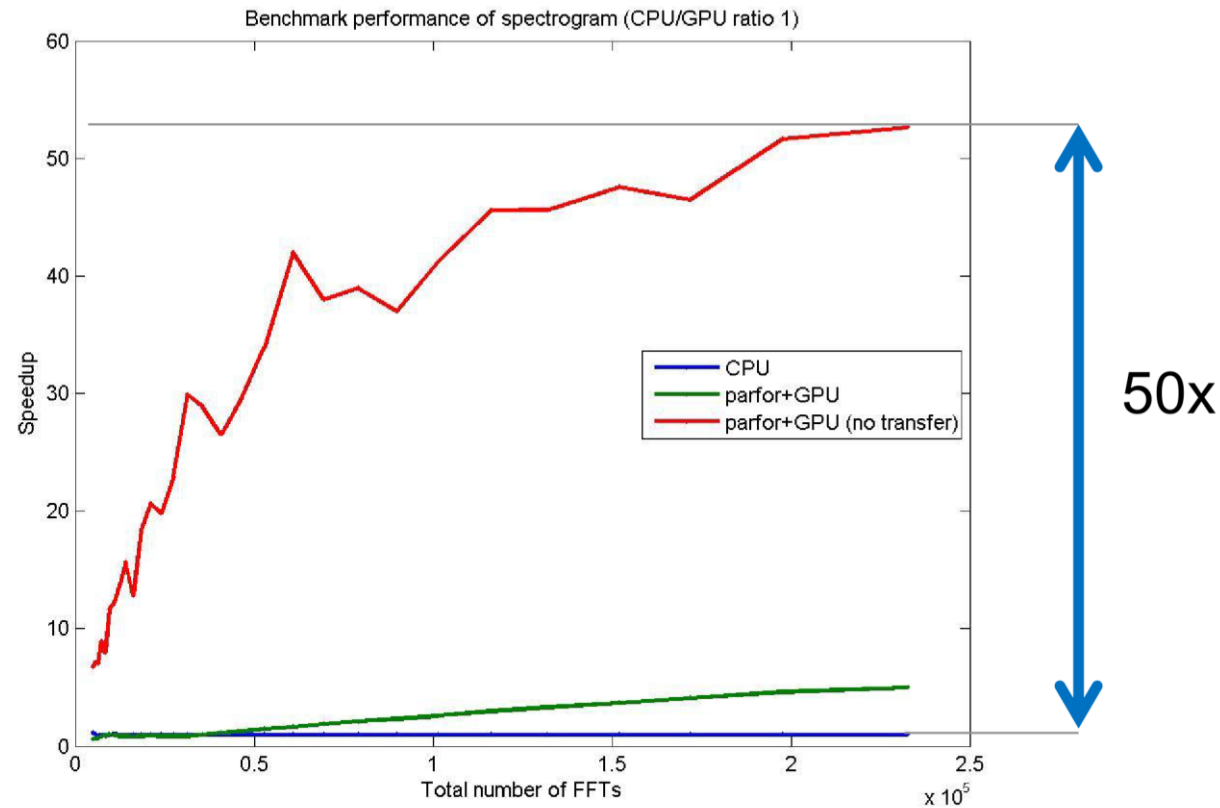
Использование больше ядер (CPUs)



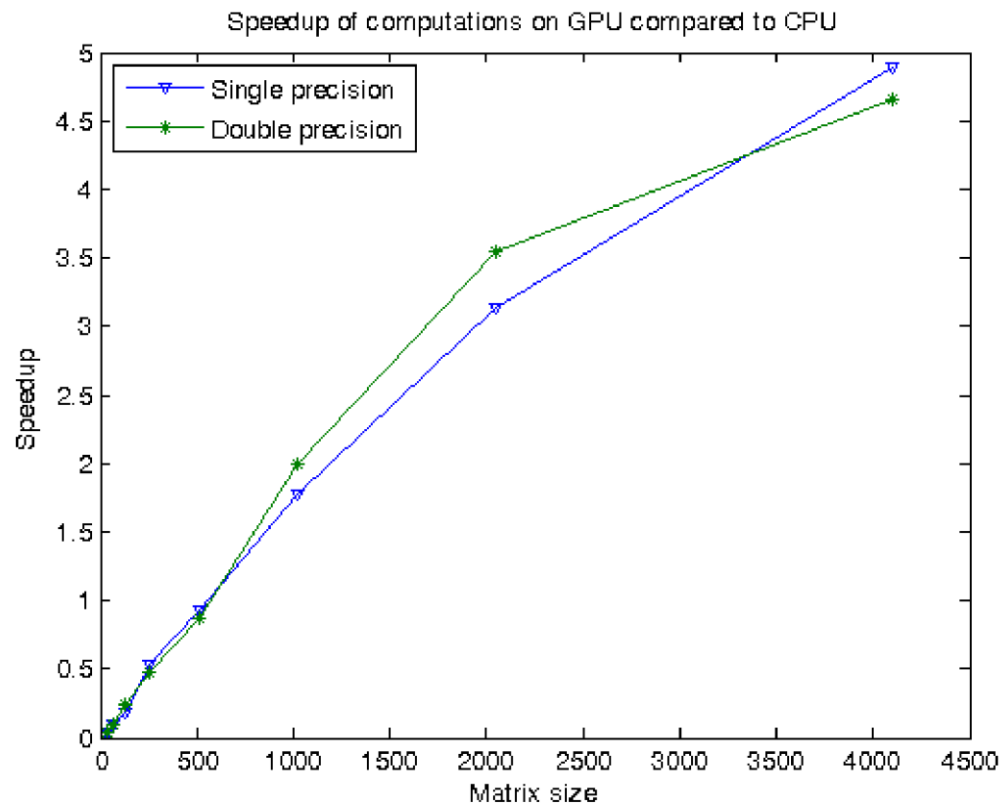
Использование GPUs



Спектрограмма показывает 50ти кратное увеличение скорости вычислений на GPU кластере



Ускорение MATLAB на GPU



<http://www.nvidia.com/object/tesla-matlab-accelerations.html>

Возможности при использовании GPUs

Проще в использовании

Использование интерфейса GPU array со встроенными функциями MATLAB

Запуск пользовательских функций над элементами GPU array

Создание ядер из существующего кода CUDA и PTX файлов

Больше контроля

Спасибо за внимание!