GPGPU, 2nd Meeting

Mordechai Butrashvily, CEO

moti@gass-ltd.co.il

GASS Company for Advanced Supercomputing Solutions

Agenda

- 1st meeting
- 2nd meeting
- Future meetings
- Activities

1st meeting

- Got familiar with GPU technology and trends
- Covered historical developments
- Current and latest technologies from NVIDIA and ATI

2nd meeting

- GPU computing using NVIDIA software stack and products
- CUDA programming
- Short example
- Tesla platform
- GPGPU for IT
- Questions

Future meetings

- Software stacks and frameworks by NVIDIA and ATI:
  - **CUDA - √**
  - StreamComputing
- Developments and general talks about programming and hardware issues
- More advanced topics
- Looking for ideas ☺

Activities

- Basis for a platform to exchange knowledge, ideas and information
- Cooperation and collaborations between parties in the Israeli industry
- Representing parties against commercial and international companies
- Training, courses and meetings with leading companies

NVIDIA software stack

GPU Computing for programmers

Agenda

- Software stack for GPU computing:
  - CUDA Toolkit
    - Compiler, Assembler
    - Libraries
    - Documentation
  - CUDA SDK

CUDA Toolkit

- CUDA –

  Compute **U**nified **D**evice **A**rchitecture
- Provides the runtime required to run CUDA based solutions
- Supporting all GPUs starting from G80 (GeForce 8x00 series)

Cont.

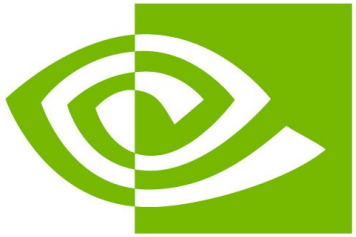
- CUDA 1.1 released in Q4 2007:
  - Single precision arithmetic
  - Supports Windows XP/Linux/MacOS 32 bit
- CUDA 2.0, final in Aug 2008:
  - Single/Double precision arithmetic
  - Support for XP/Vista/Linux in both 32 and 64 bits

Cont.

- Includes:
  - Compiler for CUDA code (binary that runs on the GPU)
  - Assembler for PTX language
  - Documentation
  - Runtime, FFT and BLAS libraries

CUDA SDK

- Provides additional information for developers
- Examples covering many aspects of CUDA programming
- Supported by every platform as the toolkit



NVIDIA

CUDA Programming

Syntax, capabilities etc.

Agenda

- What is CUDA
- Why is it good?
- What can be done with it?
- Summary of capabilities by CUDA
- Additional tools

What is CUDA

- CUDA can be considered as another shader language for GPUs
- Providing low level access to the hardware
- Without knowing graphics API (DX, GL)
- CUDA is a framework that provides:
  - Development tools
  - Runtime
  - Defines a language

Why is it good?

- Provides low level access to the GPU hardware
- Much faster than traditional Graphics API
- Language that is specific for computing, without graphics terms
- C/C++ based syntax (upcoming support for Fortran)
- Porting existing code isn't that difficult

What can be done with it?

- Every computation that fits a GPU
- Using CUDA we can:
  - Allocate and transfer memory between a device and host
  - Run specific “kernel”s (math computations)
  - Configure the amount of cores to utilize
  - Access DirectX and OpenGL resources (texture data) during process

Summary of capabilities by CUDA

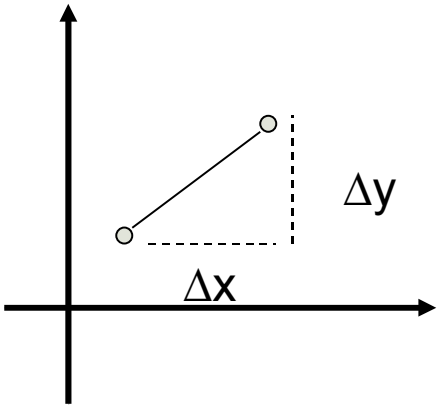
- Vector types:
  - char{1-4}, short{1-4}, int{1-4}, long{1-4}, float{1-4}
  - Unsigned version of integers
- No vector operators!
- Intrinsic functions (sin, cos, exp etc.)
- Procedural programming

Cont.

- In the host side:
  - Allocating memory and transferring data
  - Support for 2D and 3D block copies
  - Asynchronous memory transfer (bi-di)
  - Asynchronous execution
  - Support for FFT (1D-3D) and BLAS routines (partial support for complex ops)

Additional tools

- **CUDA.NET** – Develop cross-platform CUDA solutions
- **CAPS HMPP** – Accelerate your program using available co-processor (Multi-Core, GPU, Cell), using C/C++/Fortran etc.



Short example

Slope computation

Example

- Using CUDA and Tesla platform
- Computing slopes between coordinates in 2D space:
  - Total of 2^{26} (67,108,864) points pairs
  - 2^{27} total points
  - Result: 2^{26} slopes
  - Total memory:
    - 1 GB input (0.5 GB for each data-set)
    - 256 MB for results

Environment

- CUDA 2.0 beta2
- Tesla C870:
  - 128 cores
  - 1.5 GB RAM
- Linux Fedora Core 9
- CUDA.NET for executing the kernel (Mono 1.9)

Execution Plan

- Number of blocks:
  - X: 512 (2^9)
  - Y: 256 (2^8)
- Each with 512 threads
- Total threads – 2^{26}
- Total memory – 1.25 GB (input & output)

The code


```
extern “C” __global__ compute_slopes(float2* p1,
   float2* p2, float* result)
{
   // Get index into array
   int i = blockIdx.x + blockIdx.y * gridDim.x;
   i += threadIdx.x;

   // Compute (y2-y1) / (x2-x1)
   result[i] = (p2[i].y – p1[i].y) / (p2[i].x – p1[i].x);
}
```

Results

- 132.868 ms on the GPU
- ~1200 ms on the CPU (Intel Quad Q9300) using SSE
- ~x10 factor
- Many optimizations to apply, so the final factor may be higher

Tesla Platform

Hardware platform for GPU computing

Agenda

- Tesla, another GPU card
- Current products
- Future products

Tesla, another GPU card

- Not just
- Another class of GPU cards, between gaming (GeForce) and professional (Quadro)
- No screen output, meant for computations only
- The recommended solution for GPU computing!

Current products

	C870	D870	S870
GPU#	1	2	4
Cores	128	256	512
Memory	1.5 GB	3 GB	6 GB
Performance	0.5 TFlops	1 TFlops	2 TFlops
Bandwidth	76.8 GB/s	153.6 GB/s	307.2 GB/s
Price	1000$	5500$	7500$

Future products

	C1060	S1070 (1U)
GPU#	1	4
Cores	240	960
Memory	4 GB	16 GB
Performance	1 TFlops	4 TFlops
Bandwidth	102 GB/s	408 GB/s
Price	1845$	8690$

GPGPU for IT

GPU Computing in Organizations

Agenda

- GPU computing solutions
- Implementing GPU environment
- IT services

GPU computing solutions

- Like covered previously –
  - C1060 – single GPU in a workstation
  - S1070 – 1U server with 4 GPUs
- It is possible to build a custom computer
- Or use a single GPU

Implementing GPU environment

- Organization usually need to implement a large scale GPU solution
- What about maintenance? And other IT services…
- Training?...

IT services

- This issues are being solved nowadays as organizations start to think about GPU solutions
- At the end, these services will help:
  - Choose the correct hardware
  - Train your IT personnel
  - Know how to manage replacement
  - Monitor GPU as network resources
- The goal is to help executives have a solid ground for using GPUs in their solutions!

Cont.

- Hybrid cluster solutions (Servers with integrated Tesla) by global vendors
- Support for systems with replacement parts available immediately

Summary

- GPU computing using NVIDIA solutions is very effective
- Providing both hardware and software
- Very cost-effective solutions compared to CPU and GRID

Questions