GPGPU, 4th Meeting

Mordechai Butrashvily, CEO

moti@gass-ltd.co.il

GASS Company for Advanced Supercomputing Solutions

Agenda

- 3rd meeting
- 4th meeting
- Future meetings
- Activities

3rd meeting

- Dr. Avi Mendelson presented Intel “Larrabee” architecture
- Covered hardware details and design information

4th meeting

- GPU computing with AMD (ATI)
- StreamComputing programming
- CAL.NET
- FireStream platform
- GPGPU for IT
- Questions

Future meetings

- Software stacks and frameworks by NVIDIA and ATI:
  - **CUDA -** √
  - **StreamComputing -** √
- Upcoming OpenCL standard
- 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

AMD Hardware

GPU Computing for programmers

AMD Hardware

	HD3870	HD4870	HD4870 X2	FirePro V8700	FireStream 9250
Core#	320	800	1600	800	800
Tflops	0.5	1.2	2.4	1.2	1.2
Core Freq.	775 Mhz	750 Mhz	750 Mhz	750 Mhz	750 Mhz
Memory	0.5 GB	1 GB	2 GB	1 GB	2 GB
Bandwidth	72 GB/s	115 GB/s	115 GB/s	108 GB/s	108 GB/s
Power	110 W	184 W	200 W	180 W	180 W
Price	150$	300$	550$	2000$	1000$

Stream Processor

- For example, HD3870, 320 cores:
  - 4 SIMD engines
  - 16 thread processors each
  - 5 stream cores per thread

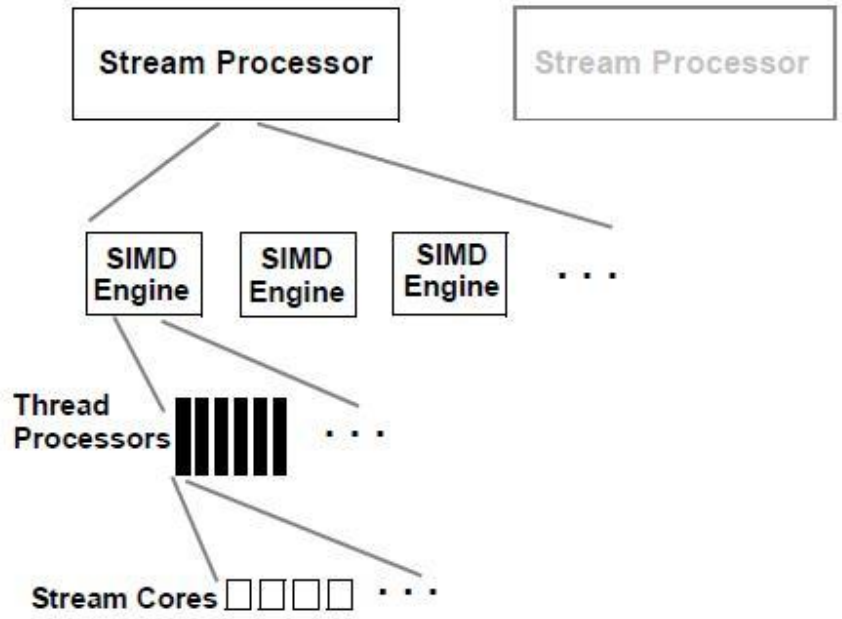


Figure 1.8 Generalized Stream Processor Structure

Stream Core

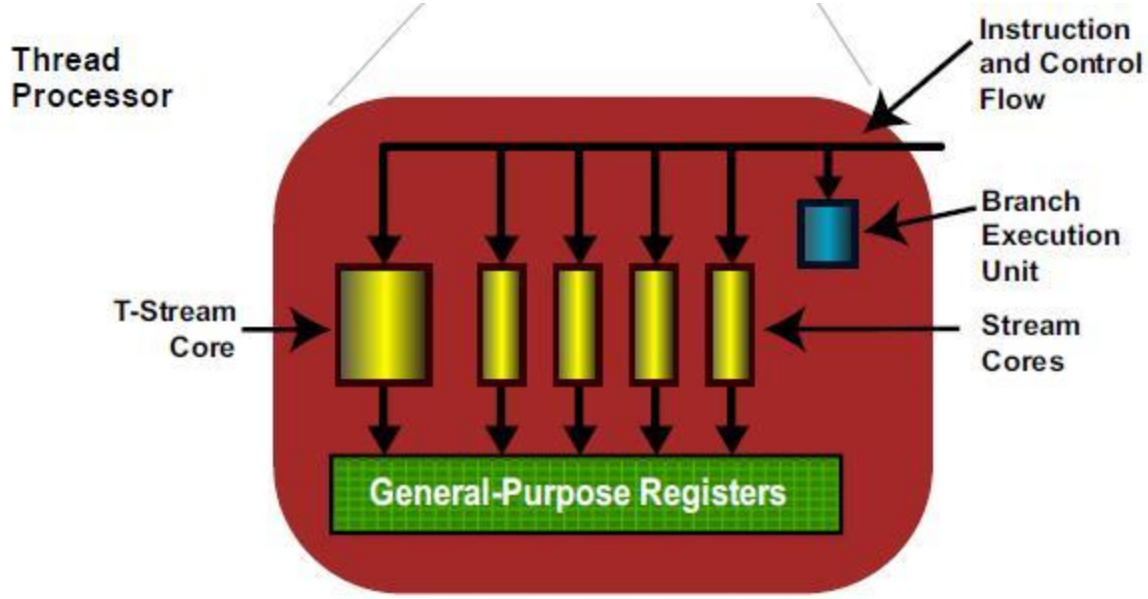


Figure 1.9 Simplified Block Diagram of the Stream Processor[1]

GPU Performance

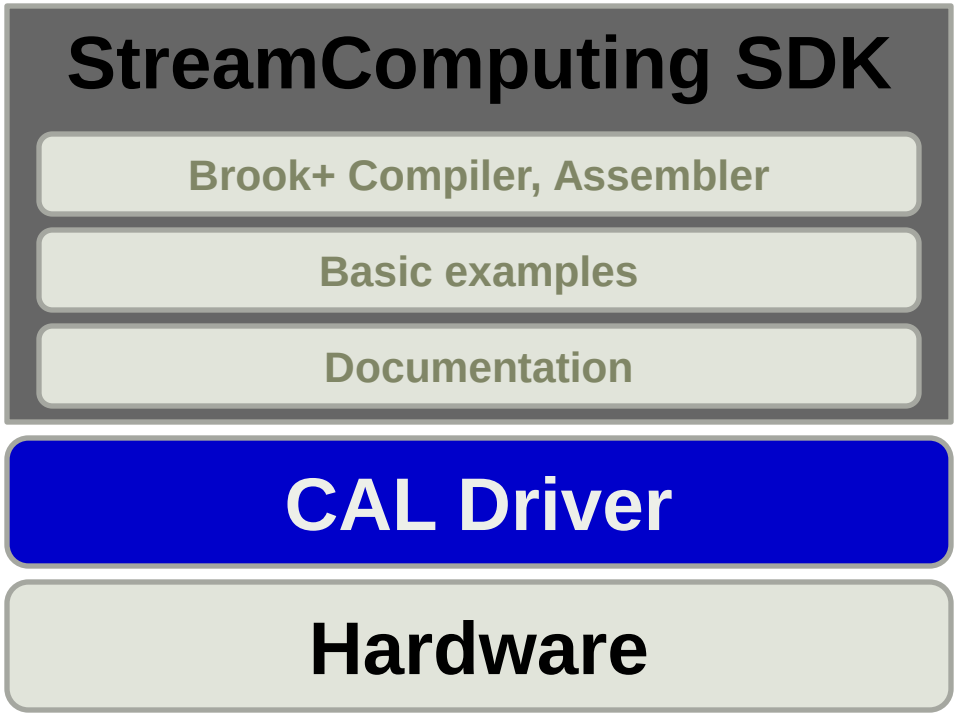
- ATI formula
- 320 cores
- Each runs at 775 Mhz
- 1 MAD per cycle
- FLOPS = Cores * Freq. * FLOPS_CYCLE
- FLOPS = 320 * 775e6 * 2
- FLOPS = 496 GFLOPS

AMD software stack

GPU Computing for programmers

Agenda

- Software stack for GPU computing:



CAL Driver

- CAL – “Close to Metal”
- Allows direct communication with GPU hardware - without using graphics API
- Located on top of the “display” driver
- Very similar functionality as CUDA API
- **History** - ATI was first to introduce a low-level interface to GPU

StreamComputing SDK

- Provides the runtime required to run “Stream” based solutions
- Supporting all GPUs starting from RV600 (Radeon HD2x00 series)

Cont.

- Includes:
  - Brook+ compiler (*brcc*, an extension to Brook)
  - C based syntax, can integrate into existing applications
  - Assembler for IL language
  - Documentation
  - Runtime library

Cont.

- Supported platforms:
  - Windows XP 32/64 bit
  - Windows Vista 32/64 bit
  - Linux 32/64 bit

StreamComputing Programming

Syntax, capabilities etc.

Agenda

- What is StreamComputing?
- Why is it good?
- What can be done with it?
- Summary of capabilities

What is StreamComputing?

- Can be considered as another shader language for GPUs
- Providing low level access to the hardware
- Without knowing graphics API (DX, GL)
- 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
- Porting existing code isn't that difficult

What can be done with it?

- Mostly stream operations
- Using StreamComputing 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 resources (texture data) during process

Short example

Matrix multiplication

The code

```
kernel void simple_matmult(float Width, float A[][],
        float B[][], out float result<>)
{
        // vPos - Position of the output matrix i.e. (x,y)
        float2 vPos = indexof(result).xy;
        // index - coordinates of A & B from where the values are fetched
        float4 index = float4(vPos.x, 0.0f, 0.0f, vPos.y);
        // step - represents the step by which index is incremented
        float4 step = float4(0.0f, 1.0f, 1.0f, 0.0f);

        float accumulator = 0.0f;
        float i0 = Width;
        while(i0 > 0)
        {
                // A[i][k] * B[k][j]
                accumulator += A[index.zw]*B[index.xy];
                index += step;
                i0 = i0 - 1.0f;
        }

        // Writing the result back to the buffer
        result = accumulator;
}
```

CAL.NET

.NET library for CAL and StreamComputing

CAL.NET

- A .NET library that provides access to AMD GPU hardware from:
  - Windows XP/Vista
  - Linux
- Manage devices, allocate memory, load and execute code

CAL.NET – The future

- OpenCL standard gains popularity
- Several providers for high level languages:
  - CUDA.NET – CUDA & NVIDIA
  - CAL.NET – StreamComputing & AMD
- Providing a unified interface for GPU hardware for other technologies as well:
  - Java, C++, FORTRAN etc…

FireStream Platform

Hardware platform for GPU computing

Agenda

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

FireStream, another GPU card

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

Some history

- AMD was first to announce one year ago a card with double precision
- BUT, it provided 0.5 TFlops and costs 2000$

Current products

	FireStream 9170
GPU#	1
Cores	320
Memory	1 GB
Performance	0.5 TFlops
Bandwidth	~75 GB/s
Price	2000$

Future products

	FireStream 9250
GPU#	1
Cores	800
Memory	2 GB
Performance	1 TFlops
Bandwidth	108 GB/s
Price	~1000$

GPGPU for IT

GPU Computing in Organizations

Agenda

- GPU computing solutions
- Implementing GPU environment
- IT services

GPU computing solutions

- Like covered previously –
  - FireStream 9250 – single GPU in a workstation
  - Or embed within a 1U server having PCIe

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 FireStream) by global vendors
- Support for systems with replacement parts available immediately

Summary

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

Questions