Theme: Towards Reconfigurable High-Performance Computing

Lecture **4**

Platforms II: Special Purpose Accelerators

Andrzej Nowak

CERN openlab (Geneva, Switzerland)

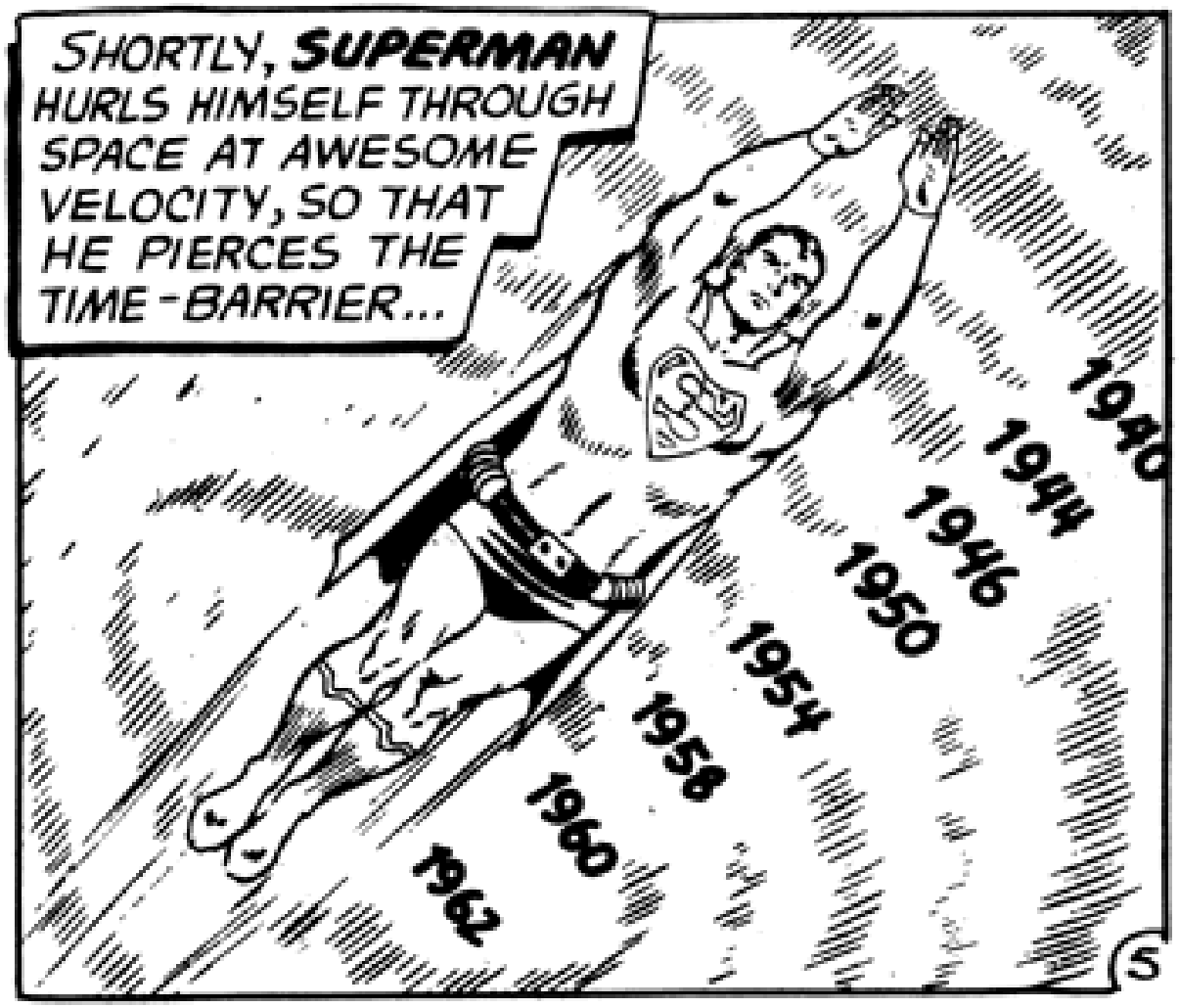
Inverted CERN School of Computing, 3-5 March 2008

CERN School of Computing

Introduction

- **Recap:**
  - General purpose processors excel at various jobs, but are no match for accelerators when dealing with specialized tasks
- **Objectives:**
  - Define the role and purpose of modern accelerators
  - Provide information about General Purpose GPU computing
- **Contents:**
  - Hardware accelerators
  - GPUs and general purpose computing on GPUs
  - Related hardware and software technologies

Hardware acceleration philosophy



Popular accelerators in general

- **Floating point units**
  - Old CPUs were really slow
  - Embedded CPUs often don't have a hardware FPU
  - 1980's PCs – the FPU was an optional add on, separate sockets for the 8087 coprocessor
- **Video and image processing**
  - MPEG decoders
  - DV decoders
  - HD decoders
- **Digital signal processing (including audio)**
  - Sound Blaster Live and friends

Mainstream accelerators today

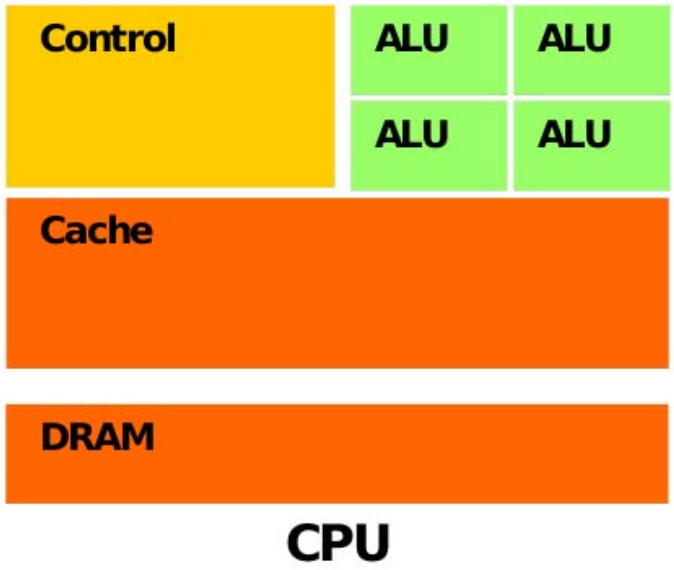
- **Integrated FPUs**
- **Realtime graphics**
  - Gaming cards
- **Gaming physics**
  - AGEIA PhysX gaming card
- **Digital audio processing**
  - Creative Sound Blaster X-FI
- **Networking**
  - KillerNIC
- **Encryption**
  - Add on and on-board dedicated crypto modules
- **Platform development**
  - AMD Torrenza (coprocessor integration initiative)
  - Intel/IBM Geneseo (PCIe extensions)

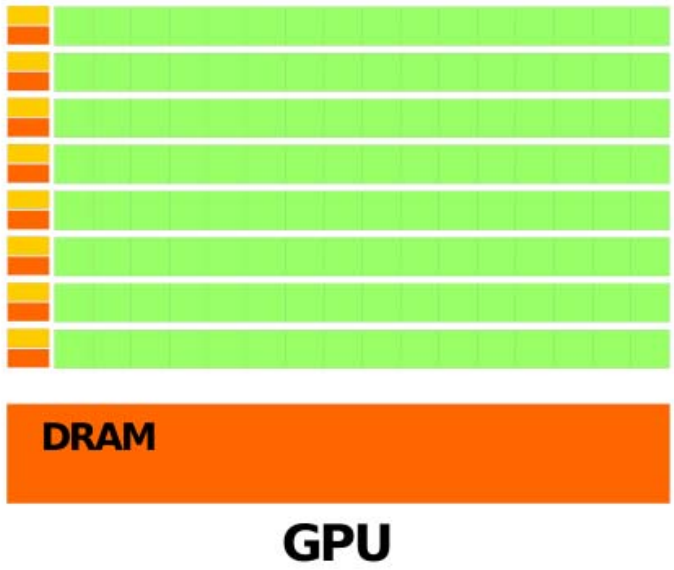
GPUs

Bobby wants to play a game

The rise of the GPUs

- **Graphics Processing Units – A mainstream, market-driven vector computing accelerator family**
  - Simple operations
  - Large width and throughput
  - Medium frequencies





Graphics: NVIDIA

Modern GPU features

- **Dozens of processing cores**
  - Some cores usually end up disabled due to manufacturers' yield problems
- **A lot of power consumed compared to CPUs - ~150 W**
- **Very fast in vector calculations, up to hundreds of GFLOPS**
- **Market driven features**
  - Main actors: Red, Green, Blue and Alpha
  - DirectX 10 or 10.1 compatibility
  - Different shader model support
- **Active ongoing development**

GPGPU

- **GPGPU – General Purpose GPU computing**
- **GPUs are becoming more universal and versatile**
- **Vast amounts of processing power left unused – what shall we do with it?**
  - Stream processing
- **Main pain – lack of native 64-bit floating point support (double precision)**
- **The domain is moving forward – chip makers are listening to the scientific community**
- **Is GPGPU the answer to your problem?**
  - Large data set
  - High parallelism
  - Small amount of dependencies with the data set
  - 64-bit floating point is not required

Common GPGPU operations

- **Stream filtering**
  - Removing items from a group based on certain criteria
- **Mapping**
  - Run a function on elements inside a group
- **Reducing**
  - Perform calculations on a stream and yield a reduced result
- **Scatter and gather**
- **Sorting**
  - Sorting networks
- **Searching**
  - Parallel searches

Which problems can benefit from GPGPU?

- **Algorithms and applications using the Fast Fourier Transform**
- **Audio processing and DSP**
- **Digital image and video processing**
- **Raytracing**
- **Weather forecasting**
- **Neural networks**
- **Molecular modeling**
- **Database operations**
- **Cryptography and cryptoanalysis**

GPU drawbacks (1)

- **FP representation and precision**
  - Non-IEEE FP representation
  - 128-bit data types but 32-bit precision
  - Low-precision math ops
  - High-precision math ops not always available, usually slow
  - Native 64-bit operations and data types missing
- **Limited amount of simultaneous logic threads**
  - Even though the GPU might have many cores, it has certain limits imposed on threading
- **Limited, high latency communication with the main memory and with the CPU (and sometimes with other cores)**

GPU drawbacks (2)

- **Heat problems**
  - Modern cards can easily achieve 150W
  - Projected Larrabee power is said to be around 150-200W
- **Feeding the beast**
  - A modern CPU is required to feed a modern GPU at full speed
- **Rudimentary development tools**
  - General purpose libraries and utilities are often absent
  - Lacking especially in higher-level languages
- **Vector processor**
  - Limited scientific applications
  - Limited flexibility
- **Data control paths unprotected, fault handling lacks robustness**

FEEDING THE BEASTS

Programming GPUs

Development kits for GPUs - CUDA

- **CUDA stands for “Compute Unified Device Architecture”**
- **General purpose development kit for the G80 chip**
- **C supported**
- **Open64 based compiler**
- **CUDA software includes BLAS and FFT libraries; areas of application:**
  - Parallel bitonic sort
  - Matrix multiplication
  - Matrix transposition
  - Performance profiling using timers
  - Parallel prefix sum of large arrays
  - Image convolution
- **Deviations from the IEEE floating point standard**

Development kits for GPUs – CTM

- **ATI/AMDs counterpart to CUDA**
- **CTM stands for “Close To Metal”**
  - A little bit too close, perhaps…
- **Good access to the native instruction set and memory**
- **Supported by Radeon cards (from R580 on) and FireStream processors (based on the X1900)**
- **AMD claims CTM delivers 8x the performance of “traditional” GPGPU methods – OpenGL or DirectX**
- **Open source**

Development kits for GPUs – Rapid Mind (1)

- **Multi-core and GPGPU development platform**
- **Mostly for graphics processing**
- **An API library for C++**

Graphics: Rapid Mind

Development kits for GPUs – Rapid Mind (2)

- **Features**
  - Code optimization
  - Automatic load balancing
  - Data management and diagnostics
- **Backends:**
  - Intel and AMD CPUs
  - NVIDIA and ATI/AMD GPUs
  - IBM Cell Processor
    - Cell Blade
    - Cell Accelerator Board
    - Sony Playstation 3

Development kits for GPUs - Brook

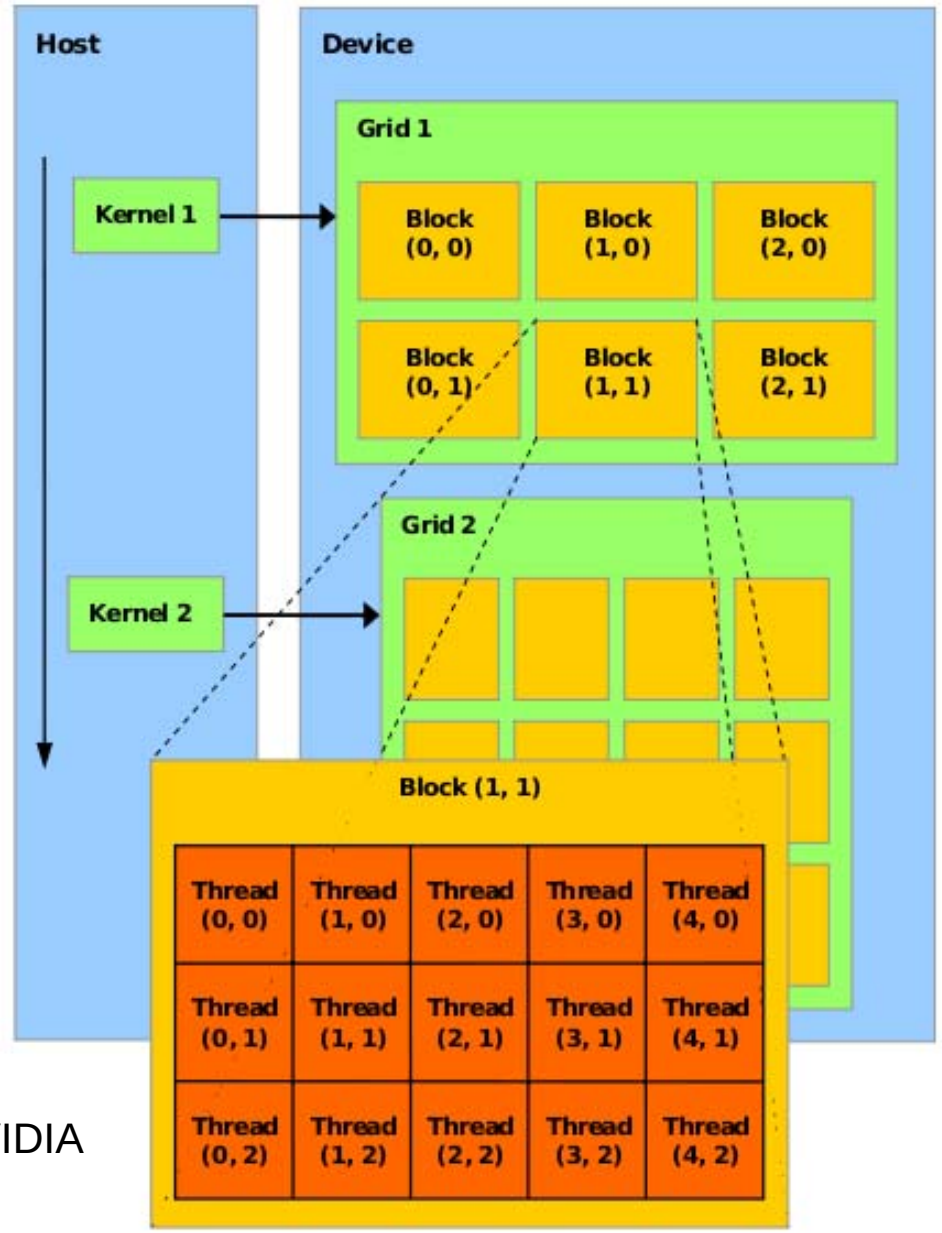
- **Stanford University's GPGPU library**
- **A derivative of ANSI C**
- **Backends: OpenGL 1.3+, DirectX 9+, CTM**
- **Runs on Linux, Windows, Mac OS X; BSD license**
- **410 GFLOPS cited (DX9, ATI HD 2900 XT)**
- **Development picked up again in 2007**

INSIDE THE HARDWARE

A peek into commodity gaming gear of today and tomorrow

NVIDIA G80

- **Stream processor developed by NVIDIA**
- **Moved away from traditional GPU design**
  - Uniform shader model
  - DirectX 10 support
- **128 stream processors**
- **330 GFLOPS peak**
- **Second generation: G92**



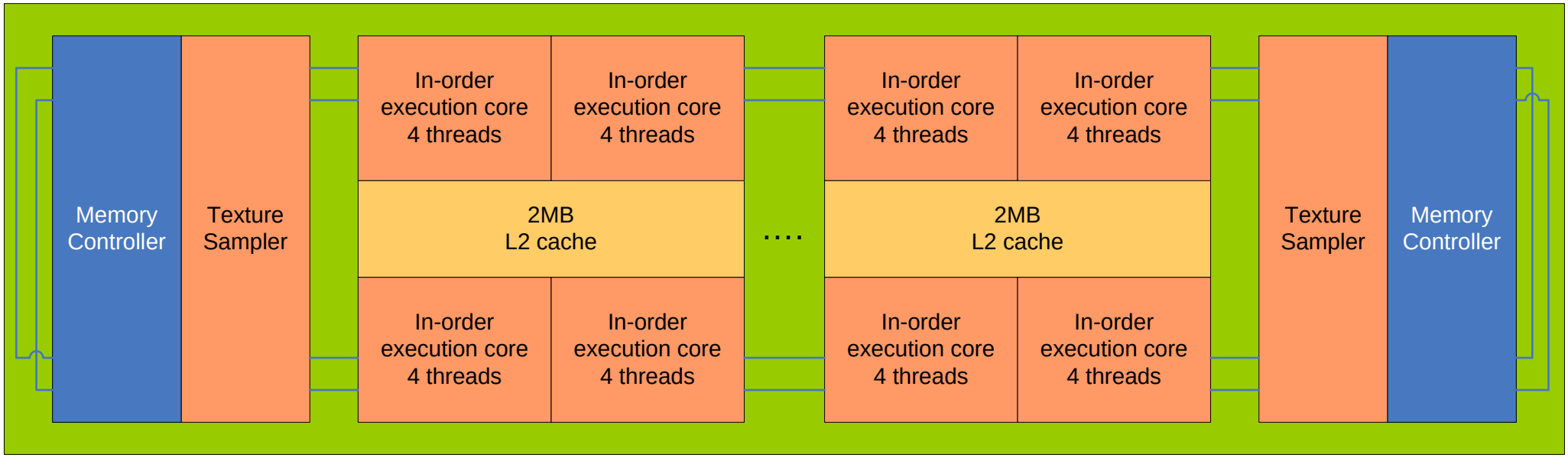
Graphics: NVIDIA

AMD FireStream

- **Stream processor developed by ATI**
- **Targets not only gamers, but the HPC community as well**
- **A FireStream general purpose extension card exists**
  - Can be used as a floating point coprocessor
- **Specs:**
  - 48 pixel shaders
  - 600 MHz clock
- **Part of AMD Torrenza**

Intel Larrabee

- **45nm process, 1.7 – 2.5 GHz, > 150W**
- **16-24 in order cores for pixel/vertex shading**
  - 4 threads per core, capable of 2 double-precision FP ops per cycle



SPECULATIVE INFORMATION. Source: ArsTechnica

ClearSpeed cards

- **Attached to the PCI bus (or PCIe)**
- **Central point: the CSX600 chip**
  - 96 compute engines
  - 64-bit floating point capability
  - Full IEEE floating point compliance
- **2 chips, 80 GFLOPS per board**
- **They claim to have the highest FLOP/Watt (2GFLOP/Watt)**
  - 30 Watts per board
- **Toolkit available**
  - C-based compiler
  - Development tools – assembler, debugger, profiling tools
  - BLAS, LAPACK available

AMD Torrenza

- **An initiative to link coprocessors with AMD Opteron systems**
  - Hyper Transport
  - PCIe
- **Related to the AMD Fusion platform project**
- **Example conforming products:**
  - Qlogic Infinipath network adapters
  - DRC coprocessor modules (Xilinx Virtex-4 FPGA)
  - XtremeData coprocessor modules (Altera Stratix II FPGA)
- **IBM Roadrunner supercomputer will link 16’000 Opteron systems and 16’000 CELL systems to reach 1 petaflop**

Other mainstream accelerators

- **EMU10k1 (1998)**
  - DSP processor for audio applications (SB Live)
  - 1000 MIPS
  - 2.5 M transistors
- **EMU20k1 (2005)**
  - DSP processor for audio applications (SB X-FI)
  - 10’000 MIPS
  - 50 M transistors
- **KillerNIC**
  - Network acceleration card
  - Offloads common network operations from the CPU

Possible future scenarios

- **CPUs will feature more and more functionality integrated on a single chip**
- **The evolution of FPGAs will facilitate the delivery of multi-purpose reconfigurable accelerators**
- **GPUs will become more versatile, with double-precision floating point support**
- **As sophisticated technologies become more available and faster interconnects settle in for good, general purpose accelerators will enter the mainstream**
- **We will see more accelerator hardware from startups**

Predictions for the future

- **The graphics accelerator market will continue to grow and evolve at a rapid pace due to consumer demand**
- **Programming graphics accelerating devices will become easier with time, as hardware manufacturer's incorporate GPGPU friendly changes into their products**
- **Shrinking manufacturing processes will ensure rapid hardware evolution – better logic, more logic on a single chip**

Q&A