

GPU Computing

E. Carlinet, J. Chazalon {firstname.lastname@lrde.epita.fr}

May 21

EPITA Research & Development Laboratory (LRDE)



GPU and architectures

Scientific Computing

```
!include agenda.inc.md
```

GPU and architectures

Why using GPU ?

We want to have things *done quickly*.



Mobile dev.



Big data

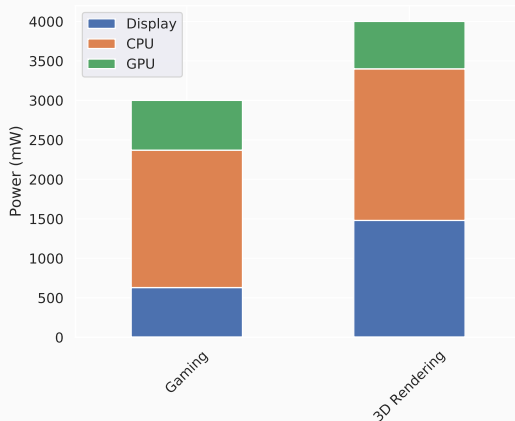


Real time computing

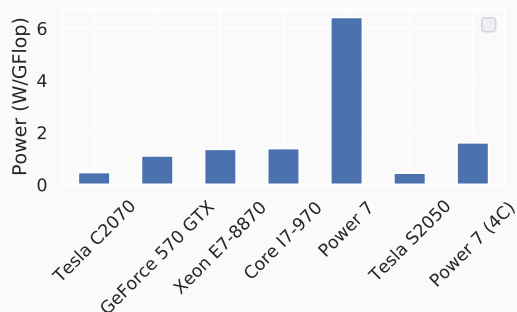
- Mobile development: limited battery
- Big data analysis: huge data volume
- Real time system: has to provide a response in a bounded time

Power Consumption on Smartphones

CPU is a major source of power consumption in smartphones (even with graphical-oriented app)



Power Consumption of Some Processors



Fabricant	Type	Modèle	Gflops	Prix	Watt
Nvidia	1x GPU (448 cœurs)	Tesla C2070	515	2500 \$	238 W
Nvidia	1x GPU (448 cœurs)	GeForce 570 GTX	198	350 \$	218 W
Intel	1x CPU (10 cœurs)	Xeon E7-8870	96	4616 \$	130 W
Intel	1x CPU (6 cœurs)	Core I7-970	94	583 \$	130 W
IBM	CPU (8 cœurs)	Power 7	265	34 152 \$	1700 W
Nvidia	4xGPU (1792 cœurs)	Tesla S2050	2060	12 000 \$	900 W
IBM	4xCPU (32 cœurs)	Power 7	1060	101 952 \$	1700 W

Scientific Computing

A bit of history - The first GPU

- Back in 70's GPU were for Image Synthesis
- First GPU: Ikonas RDS-3000
- N. England & M. Whitton foundend Ikonas Graphics Systems
- Tim Van Hook wrote microcode for ray tracing (SIGGRAPH'86)
- "All computation is taking place in the Adage 3000 Display" (1)

The GRAPHICS SYSTEM for the 80's

RDS-3000 Graphics Processor and Raster Display System

If your graphics and imaging applications are demanding, the IKONAS RDS-3000 series is the system that can meet your need. The RDS-3000 offers:



POWER

- High Speed Architecture designed for computer graphics and image processing
- Fast 32 bit processor for graphics data generation
- Hardware Matrix Multiplier for 3-D transformations, vector products, and filtering operations
- Real Time Video Processing Module for image processing applications
- Video Input Module for real time "frame grabbing"



FLEXIBILITY

- Software selectable 512' or 1024' display format
- Variable frame and line rates: 200-2000 lines/frame
- Pan and scroll in pixel increments, zoom in integer ratios
- Full Window and Viewport Control



PROGRAMMABILITY

- Graphics Processor is completely user micro-programmable and executes the highly parallel code needed for real time and near real time applications
- IDL, the IKONAS DISPLAY LANGUAGE, is a high level command language which makes the IKONAS package of standard graphics routines easy to use.

EXPANDABILITY

- RDS-3000 components are modular allowing easy expansion of systems
- Small frame buffer systems can be upgraded at a later time by adding processing modules and image memories up to 1024' x 32'

Photo credits: Terrain Model: H. Weiss
B. Marshall, Computer Graphics
Research Group, Ohio State University
Mountain: Owen Carpenter, Browning
Computer Services

IKONAS

IKONAS GRAPHICS SYSTEMS, INC.
4033 Glenwood Ave.
Raleigh, NC 27603, 919/833-5401

Reader Service Number 31



First programmable GPU:

- Vertex Shaders – programmable vertex transforms, 32-bit float
- Data-dependent, configurable texturing + register combiners



First programmable GPU:

- Vertex Shaders – programmable vertex transforms, 32-bit float
- Data-dependent, configurable texturing + register combiners

Enabled early GPGPU results:

- Hoff (1999) – Voronoi diagrams on NVIDIA TNT2
- Larsen & McAllister (2001): first GPU matrix multiplication (8-bit)
- Rumpf & Strzodka (2001): first GPU PDEs (diffusion, image segmentation)
- NVIDIA SDK Game of Life, Shallow Water (Greg James, 2001)

GPGPU for physics simulation on Geforce 3

Approximate simulation of natural phenomena:

- Boiling liquid,
- fluid convection,
- chemical reaction-diffusion



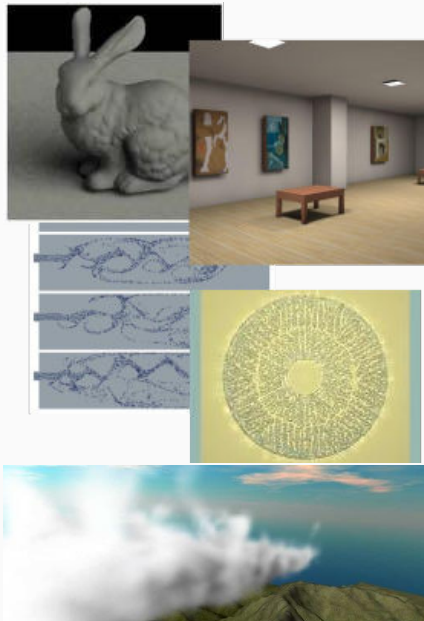
"Physically-Based Visual Simulation on Graphics Hardware". Harris, Coombe, Scheuermann, and Lastra. Graphics Hardware 2002

At that time, limited by computing precision (mostly integers).

A bit of history. GEFORCE FX (2003) : floating point

True programmability enabled broader simulation research:

- Ray Tracing (Purcell, 2002), Photon Maps (Purcell, 2003)
- Radiosity (Carr et al., 2003 & Coombe et al., 2004)
- PDE solvers
 - Red-black Gauss-Seidel (Harris et al., 2003)
 - Conjugate gradient (Bolz et al. 2003, Krueger et al. 2003)
 - Multigrid (Goodnight et al. 2003)
- Physically-based simulation
 - Fluid and cloud simulation [(Krueger et al. 2003, Harris et al. 2003)]
 - Cloth simulation (Green, 2003)
 - Ice crystal formation (Kim and Lin, 2003)
 - Thermodynamics (latent heat, diffusion)
 - Water condensation / evaporation
- FFT (Moreland and Angel, 2003)
- High-level language: Brook for GPUs (Buck et al. 2004)



Two factors for the massive surge in GPGPU dev:

- **Architecture Nvidia G80** GPU arch. and software platform designed for computing
 - Dedicated computing mode – threads rather than pixels/vertices
 - General, byte-addressable memory architecture
- **Software support.** C and C++ languages and compilers for GPUs (spoiler... it's **CUDA**)

Nvidia's G80 commercial:

A programmer will be able to treat G80 like a hugely parallel data processing engine. Applications that require massively parallel compute power will see huge speed up when running on G80 as compared to the CPU. This includes financial analysis, matrix manipulation, physics processing, and all manner of scientific computations.

... everywhere



**Computational
Geoscience**



**Computational
Chemistry**



**Computational
Medicine**



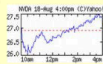
**Computational
Modeling**



**Computational
Physics**



**Computational
Biology**



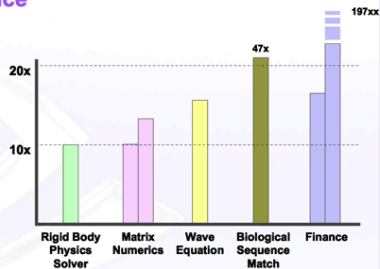
**Computational
Finance**



**Image
Processing**

Performance

CUDA
Advantage



GeForce 8800 vs. 2.66 GHz Dual Core Conroe

GPGPU provides the computing power...

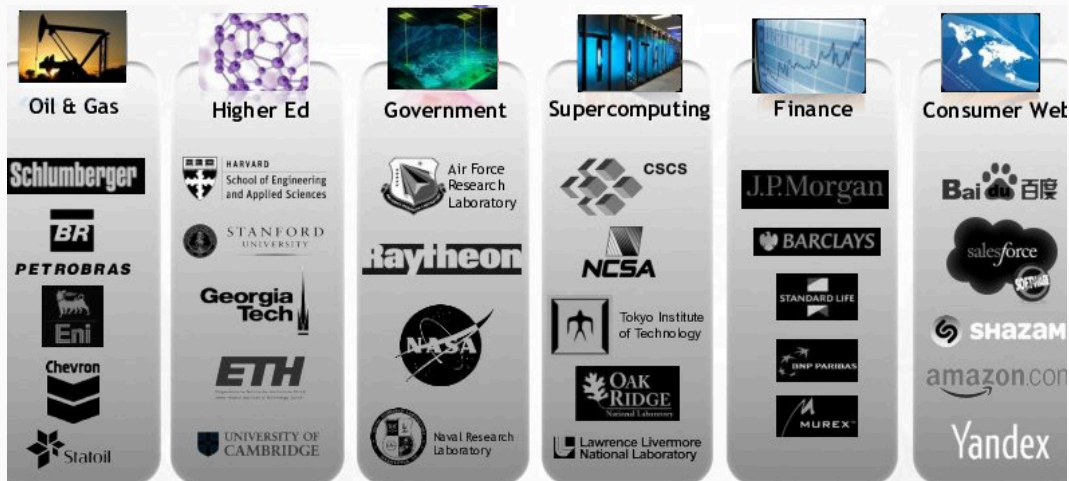


Accelerating Discoveries

Using a supercomputer powered by 3,000 tesla processors, university of illinois scientists performed the first all-atom simulation of the hi virus and discovered the chemical structure of it capsid — “the perfect target for fighting the infection.”

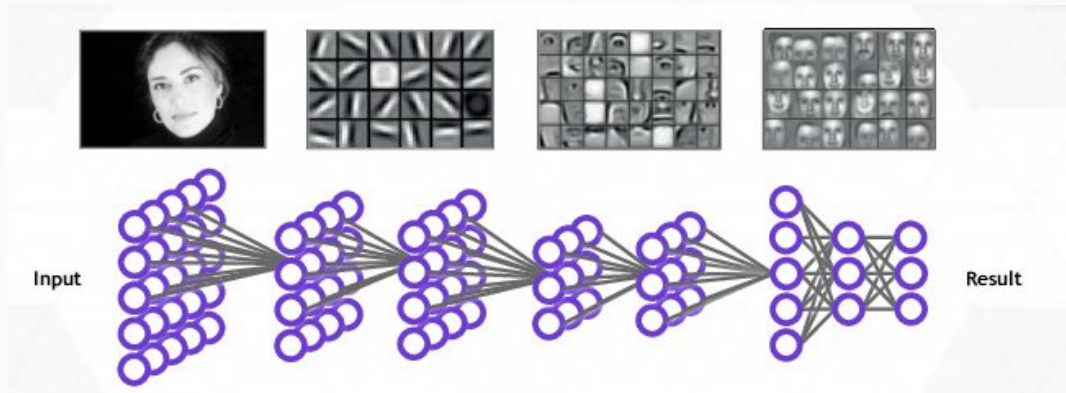
Without gpus, the supercomputer would need to be 5x larger for similar performance.

“High Performance Computing” (HPC) gives birth to Enterprise Datacenters



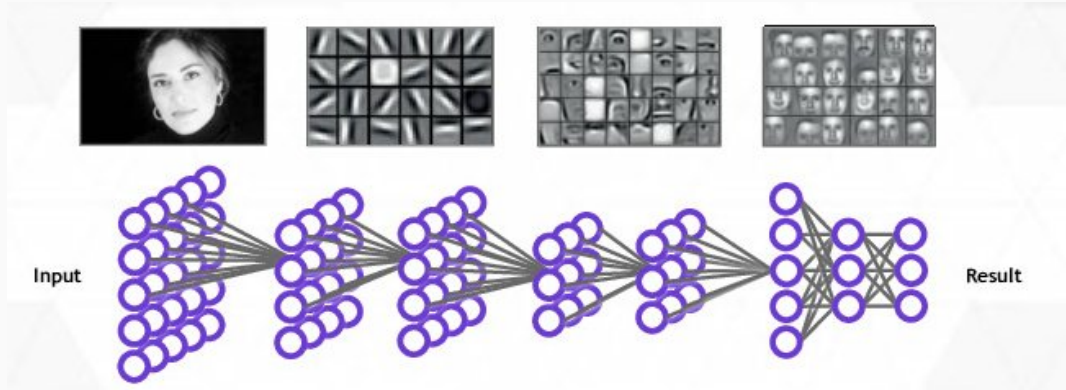
A bit of history - 2010's (3/3)

And data center gave birth to Deep-Learning (... *)



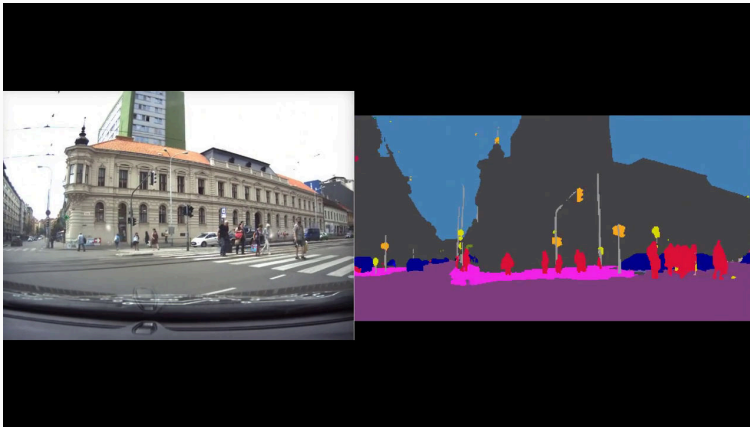
A bit of history - 2010's (3/3)

And data center gave birth to Deep-Learning (... *)



(*...) and made image processing experts useless :'(

Embedded systems - The *real*-time constraints



Clement Farabet, Camille Couprie, Laurent Najman and Yann LeCun: Learning Hierarchical Features for Scene Labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, August, 2013

Need both the two worlds:

- Need ultra-performance computing
- With limited resources