Heterogeneous Computing
(GPU+CPU Power) Joaquín Obregón-Cobo – joaquin.obregon@gmail.com

ABSTRACT

YESTERDAY
Basic concepts you need to know to see what your CUDA GPU can do will be easy to understand with a brief historic introduction. That includes concepts as SIMD an some references to Silicon Graphics and the NVIDIA steps to CUDA.

Today
Present situation is described here with references to NVIDIA and AMD. Clearly stating that, in my opinion, CUDA is leading the technology. Mainly for his SIMT architecture which has better performance and much easier and powerful full programming interface (for allowing execution divergence with SIMT instead of SIMD).

It is also explained here why we consider real parallel programming the GPU as opposite to multi-core CPUs, and the different roles of them in a “heterogeneous computing” environment. Advantages of GPU for highly parallel problems and energy consumption are mentioned.

Now
There are two main concepts to understand what heterogeneous computing is and offers to us:
A. Difference between parallel and serial programming. As we consider CUDA the best option, we will use it to illustrate the concept, implementing the sum of a vector:

A.1. Serial way. We write a C program which relevant part is:

```c
vectorAdd( int *a, int *b, int *c, int n){for (i = 0 ; i < n ; i++) c[i] = a[i] + b[i]; return;};
```

A.2. Parallel way. Our equivalent CUDA C program is:

```c
__global__ vectorAdd( int *a, int *b, int *c){int i = threadIdx.x; c[i] = a[i] + b[i]; return;};
```

B. Performance will be shown with vector sum: the addition of every element in a N sized vector. This is done in a wide spectrum of computations, for instance in matrix operations. While explaining how to do it and analyzing results we will see strong points of CUDA.

Tomorrow
We see here what we can expect from the future, as the main role to be played by GPUs due to the energy consumption/dissipation limitations to be faced or the focus on memory management in the future development of GPUs and interaction between CPU and GPU.

Tomorrow has come
The release of CUDA 4.0 confirmed the focus on memory management... new “flat” memory model.

Conclusions
Even for non-optimal problems speedup achieved is very important.
We have easy access to the tools we need, most with an affordable price, and it is probable that your computer has it inside yet.

We have the opportunity for optimizing our programs to achieve two targets:

I. Execution time reduction.

II. Interactivity, “instant” interaction, a more responsive interface.

We have the opportunity to use this technology to give to the user a better experience, to face bigger or more complex problems, to solve problems with higher precision... to obtain an advantage from our competitors.

NB: Written year 2011