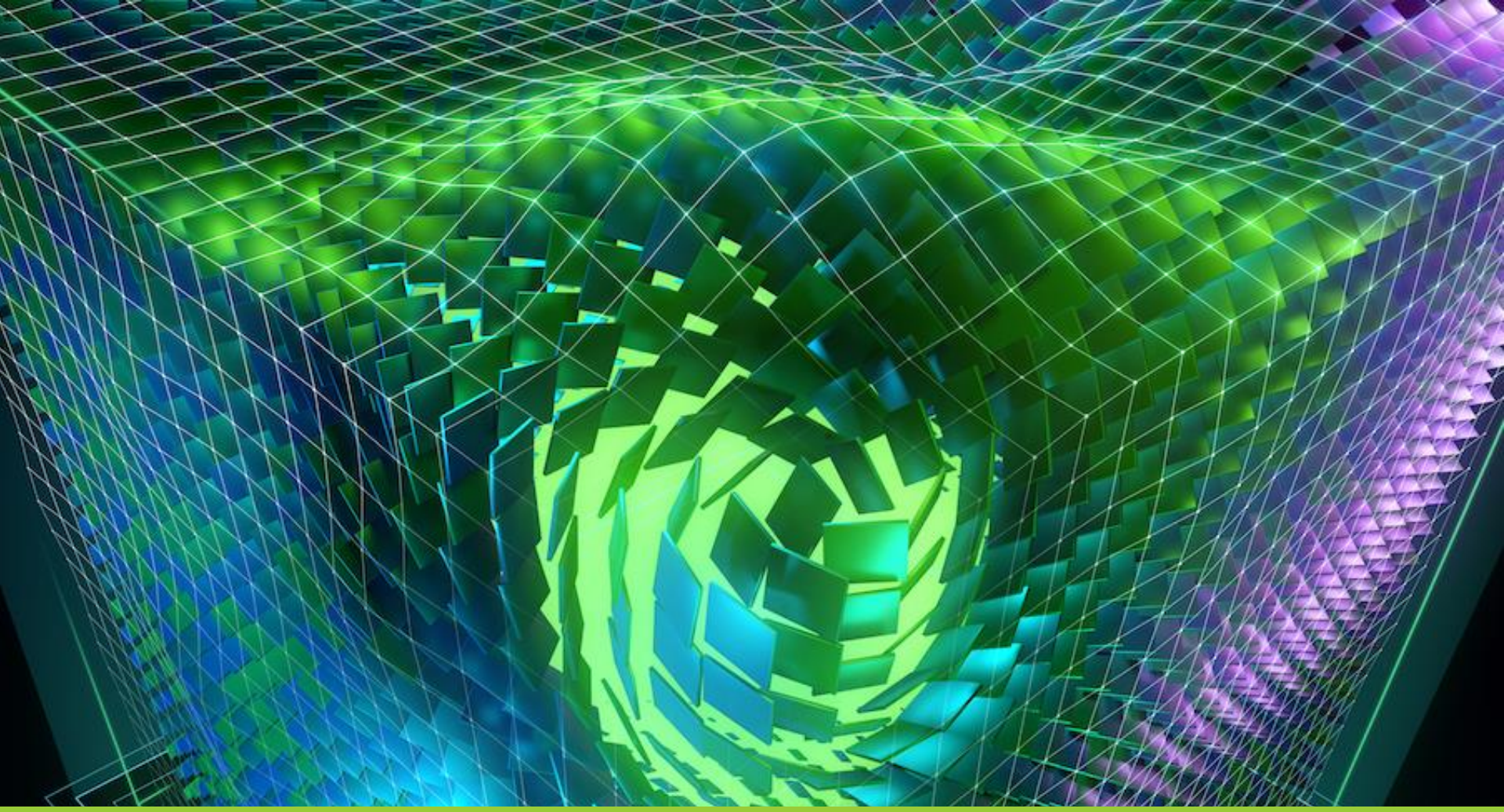


INF351 – Computación de Alto Desempeño

PROF. ÁLVARO SALINAS

Presentación del curso

Este curso cubre tópicos de programación paralela en dispositivos gráficos NVIDIA mediante el uso de CUDA.

Objetivos:

- Comprender los fundamentos básicos de la programación paralela en GPU.
- Diferenciar las arquitecturas y características de los dispositivos gráficos NVIDIA.
- Paralelizar un código secuencial a través de CUDA kernels.
- Administrar el uso de los distintos tipos de memoria disponibles en una GPU.
- Aplicar estrategias básicas de optimización al trabajar con CUDA.

Software:

- C/C++ - CUDA
- NVIDIA Visual Profiler

Evaluaciones

Durante el semestre se realizarán:

- Laboratorios: 4 laboratorios prácticos para aplicar los conocimientos adquiridos en clases.
- Actividades en clases: Problemas pequeños en el segundo bloque de cada semana.
- Proyecto: Al finalizar el curso, se deberá desarrollar y presentar una aplicación paralelizada en CUDA que resuelva un problema de su propio interés (previa aprobación del profesor).

$$NL = \frac{\sum_{i=1}^{4} L_i - \min_i L_i}{3} \qquad NF = \begin{cases} 0.2NA + 0.4NL + 0.4NP & \text{si } NP \geq 55 \\ NP & \text{si } NP < 55 \end{cases}$$

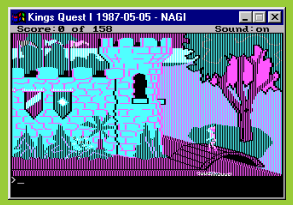
Donde L_i corresponde a la nota del i-ésimo laboratorio, NL es la nota de laboratorios, NA es la nota de las actividades en clases, NP es la nota del proyecto y NF es la nota final. El requisito de aprobación es $NF \geq 55$.

Un poco de historia

1981
IBM desarrolla la primera tarjeta gráfica: MDA.
4KB de RAM
Solo texto

1981
Llegan los colores y gráficos con CGA.
16KB de RAM
4 colores en 320x200

1985
IBM lanza EGA.
256KB de RAM
16 colores en 640x350



1990
Aparecen los gráficos de alta calidad con VGA a manos de IBM.
2MB de RAM
256 colores en 1024x768

1993
Se funda NVIDIA Corporation.

1995
NVIDIA lanza su primer producto, NV1 (2D/3D).

1995
Aparecen las tarjetas 2D/3D a manos de Matrox, S3, ATI y 3dfx.

1999
NVIDIA inventa la GPU con la línea GeForce.

Un poco de historia

2000
Debido al éxito de la GPU, Matrox y S3 caen, mientras 3dfx es absorbido por NVIDIA. ATI lanza la línea Radeon para competir.

2003
INTEL, IBM, DELL y HP crean PCI Express, el cual es el estándar hasta la actualidad.

2006
ATI es comprada por AMD, convirtiéndose en la principal competencia de NVIDIA e INTEL.

2006
NVIDIA presenta la arquitectura CUDA.

2007
NVIDIA lanza su serie Tesla de GPU para uso profesional.

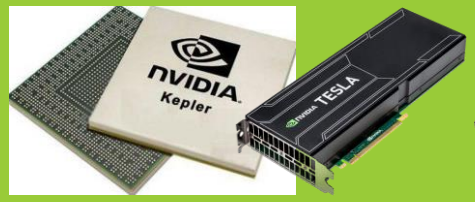
2008
NVIDIA lanza la línea Tegra.

2009
NVIDIA presenta la arquitectura Fermi.

2009
Aparece OpenCL a manos de Apple y Khronos con colaboración de AMD, IBM, Intel y NVIDIA.

Un poco de historia

2012
NVIDIA lanza la arquitectura Kepler. K20m es uno de sus productos.



2013
NVIDIA lanza la serie TITAN.

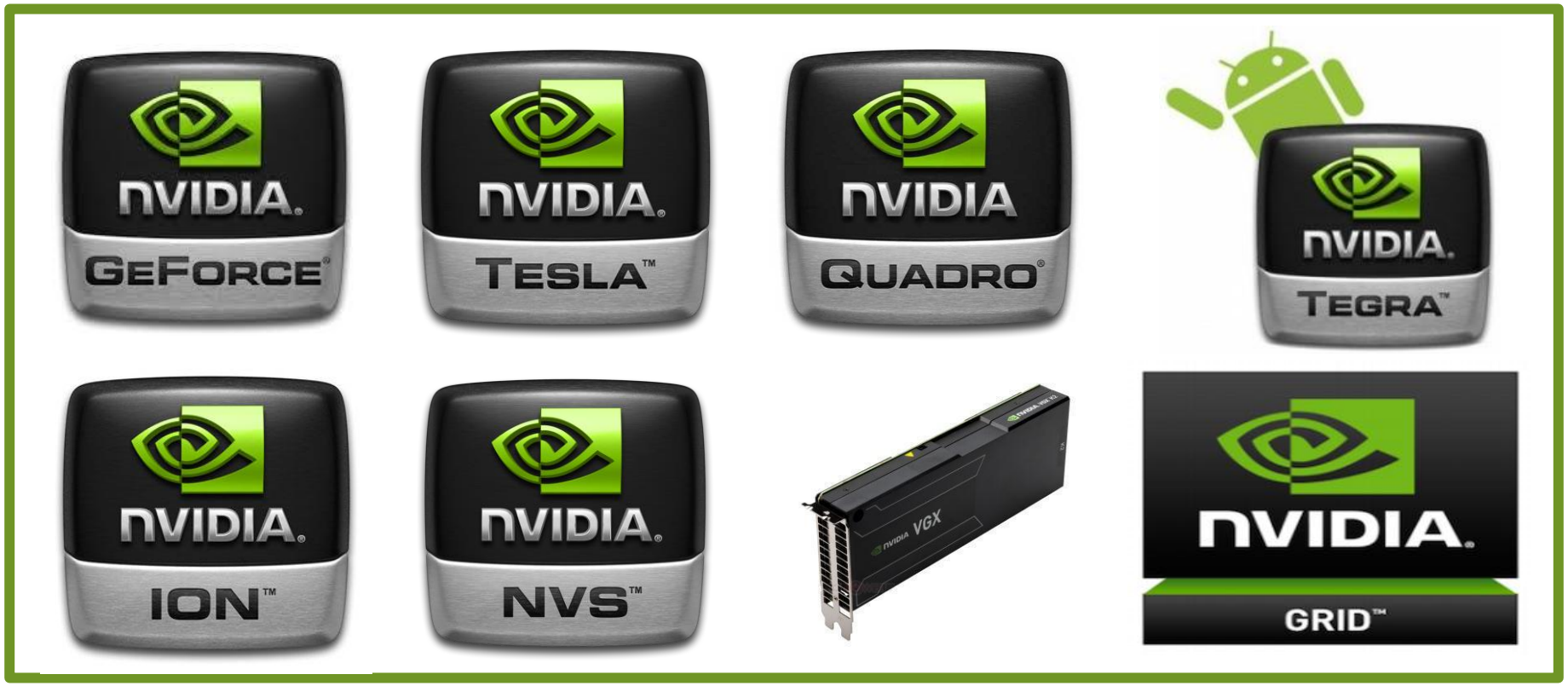
2014
NVIDIA presenta la arquitectura Maxwell.

2016
NVIDIA presenta la arquitectura Pascal.

2017
NVIDIA lanza la arquitectura Volta:
- Tesla V100
- Titan V
- Quadro GV100

2018
NVIDIA presenta la arquitectura Turing.

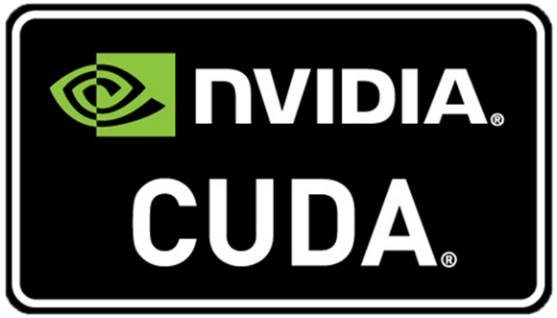
Familias de procesadores



CUDA

NVIDIA:

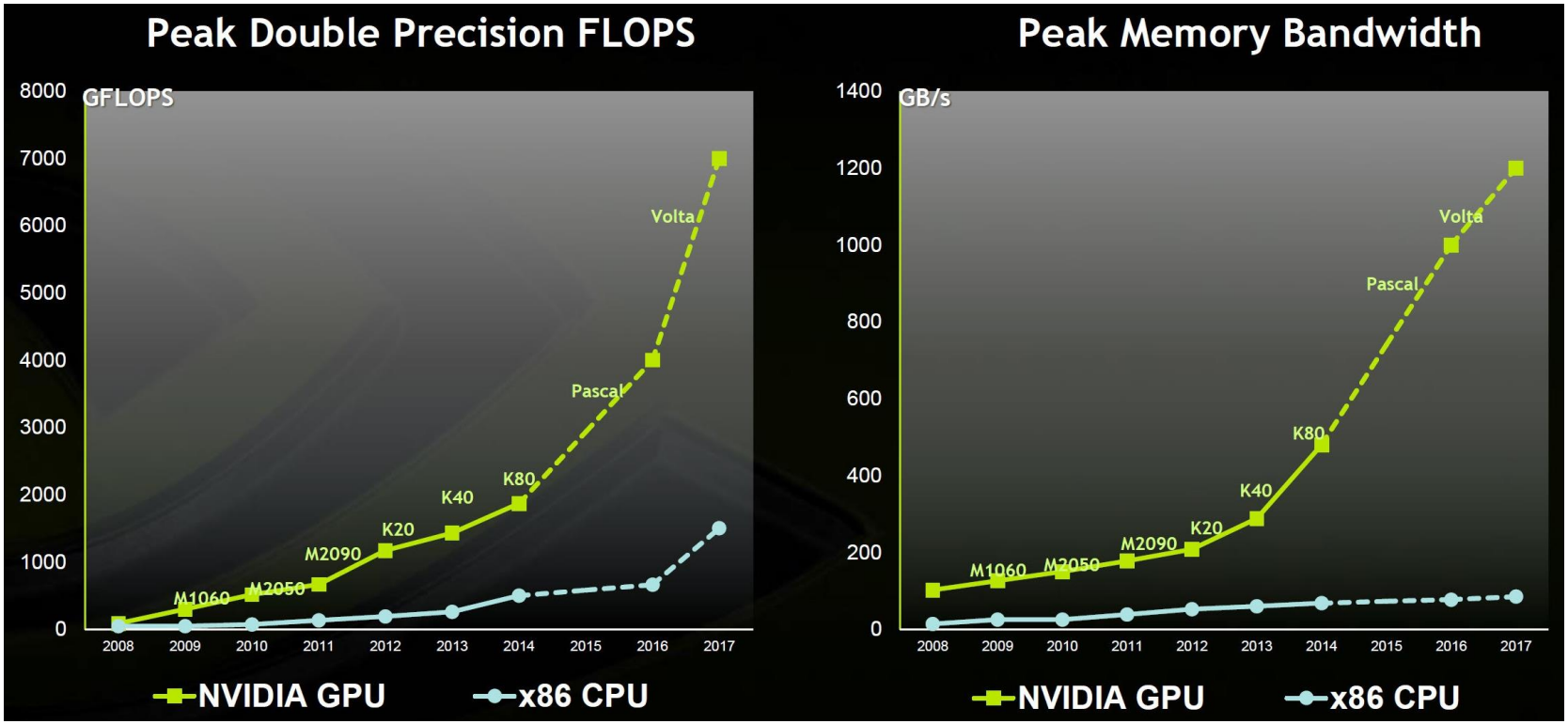
"CUDA es una arquitectura de cálculo paralelo de NVIDIA que aprovecha la gran potencia de la GPU (unidad de procesamiento gráfico) para proporcionar un incremento extraordinario del rendimiento del sistema."



Wikipedia:

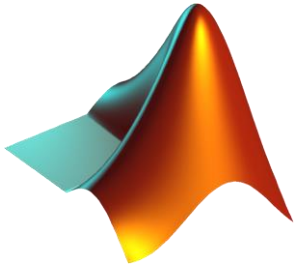
"CUDA son las siglas de **C**ompute **U**nified **D**evice **A**rchitecture (Arquitectura Unificada de Dispositivos de Cómputo) que hace referencia a una plataforma de computación en paralelo incluyendo un compilador y un conjunto de herramientas de desarrollo creadas por nVidia que permiten a los programadores usar una variación del lenguaje de programación C para codificar algoritmos en GPU de nVidia."

Evolución según microarquitecturas



Lenguajes soportados

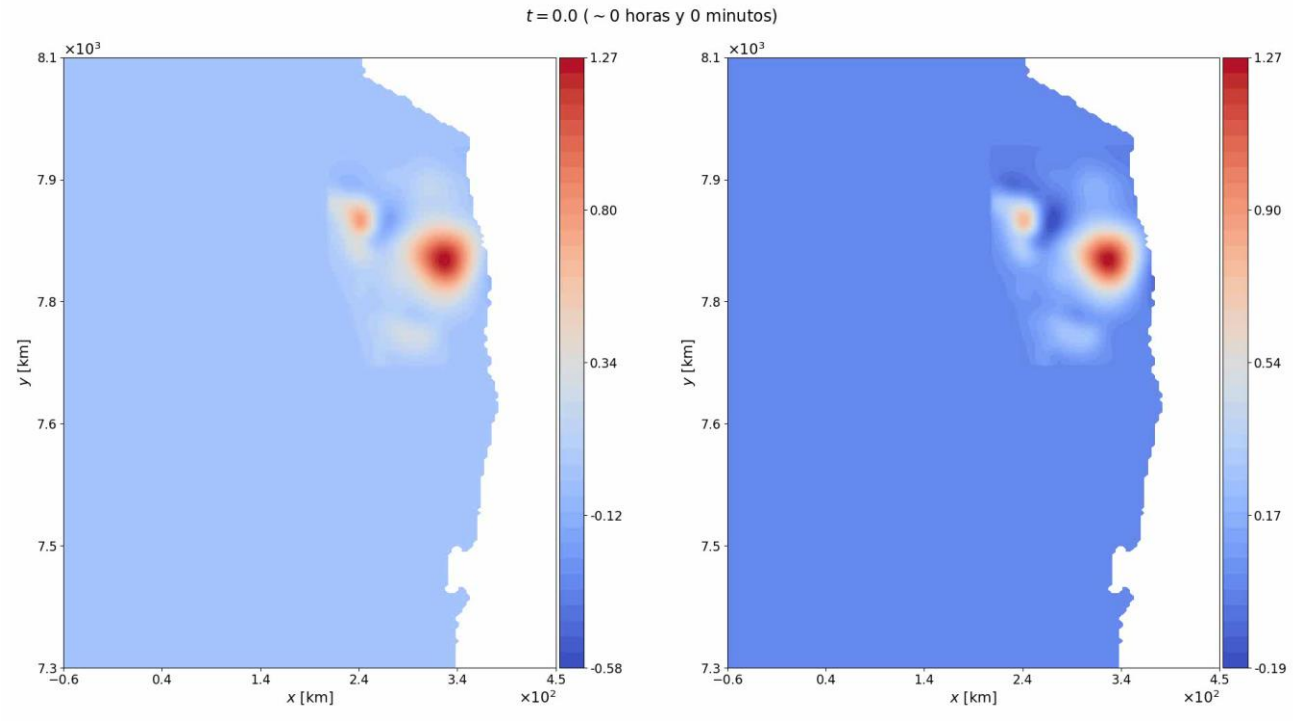
Fortran



MATLAB

¿Por qué CUDA en vez de OpenCL?

Motivación: aplicación propia



Tiempos por iteración:

Python secuencial:
$\sim 100[s]$

Python (NumPy):
$\sim 2[s]$

CUDA (Naive):
$\sim 10[ms]$

CUDA (Optimizado):
$\sim 500[\mu s]$

¡200000 veces más rápido!