CURRENT TREND IN PARALLEL COMPUTING

GPU COMPUTING AND CUDA

Source: NVIDIA websites

WHAT IS GPU-ACCELERATED COMPUTING?

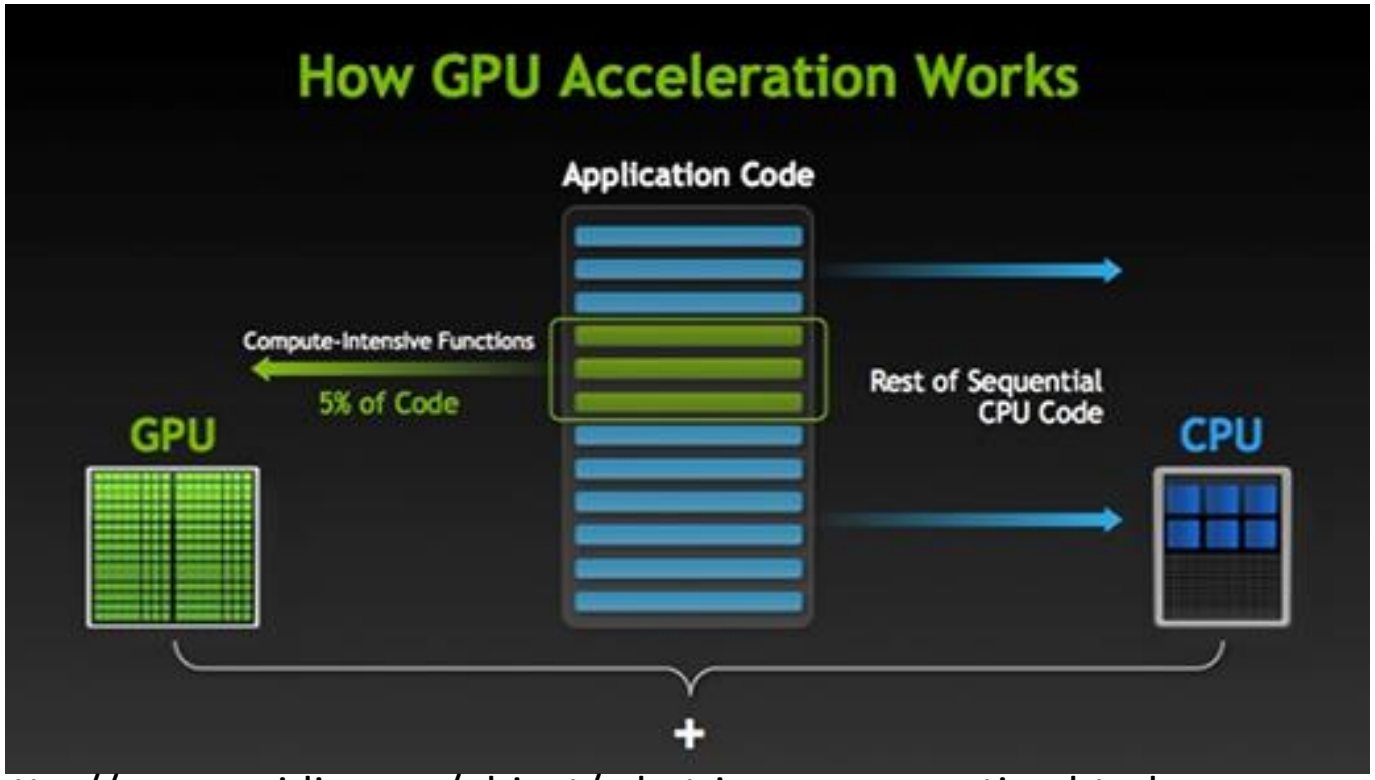
GPU-accelerated computing is the use of a graphics processing unit (GPU) together with a CPU to accelerate deep learning, analytics, and engineeringapplications.

Pioneered in 2007 by NVIDIA, GPU accelerators now power energy-efficient data centers in government labs, universities, enterprises, and small-and-medium businesses around the world.

They play a huge role in accelerating applications in platforms ranging from artificial intelligence to cars, drones, and robots.

http://www.nvidia.com/object/what-is-gpu-computing.html

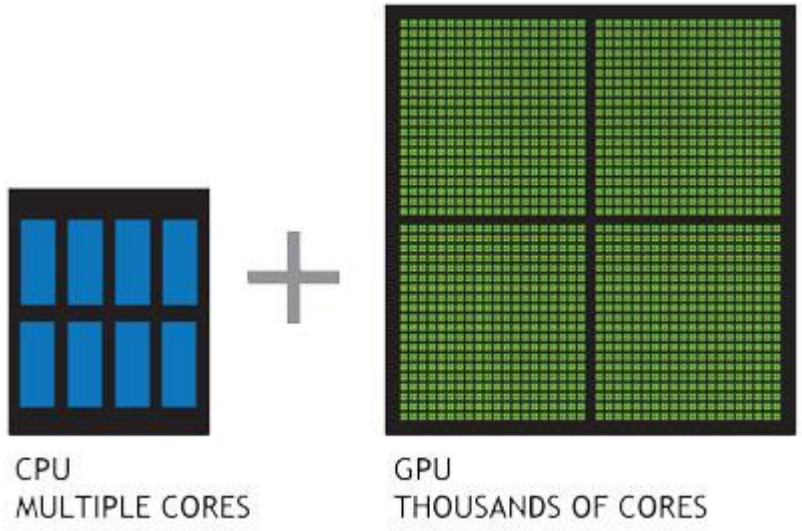
HOW GPUs ACCELERATE SOFTWARE APPLICATIONS



http://www.nvidia.com/object/what-is-gpu-computing.html

GPU-accelerated computing offloads compute-intensive portions of the application to the GPU, while the remainder of the code still runs on the CPU. From a user's perspective, applications simply run much faster.

GPU vs CPU Performance



GPUs have thousands of cores to process parallel workloads efficiently

http://www.nvidia.com/object/what-is-gpu-computing.html

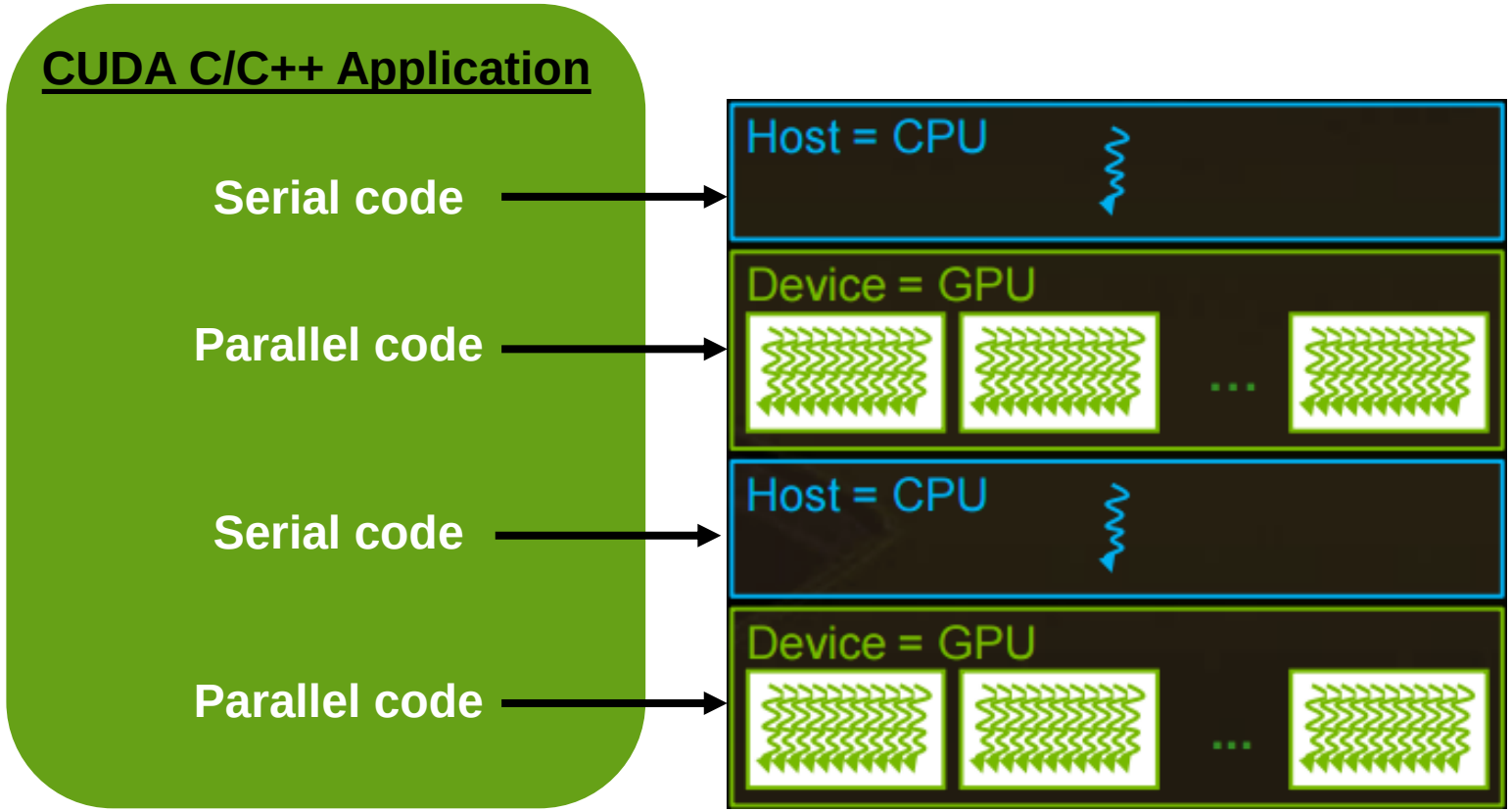
What is CUDA?

- *CUDA* stands for Compute Unified Device Architecture.
- *CUDA* is a parallel computing platform and programming model invented by NVIDIA. It enables dramatic increases in computing performance by harnessing the power of the graphics processing unit (GPU).
- C/C++ extension.
- The CUDA programming model enables you to execute applications on heterogeneous computing systems

Courtesy: Prof. Rosni Abdullah (USM)

Structure of a CUDA C/C++ Application

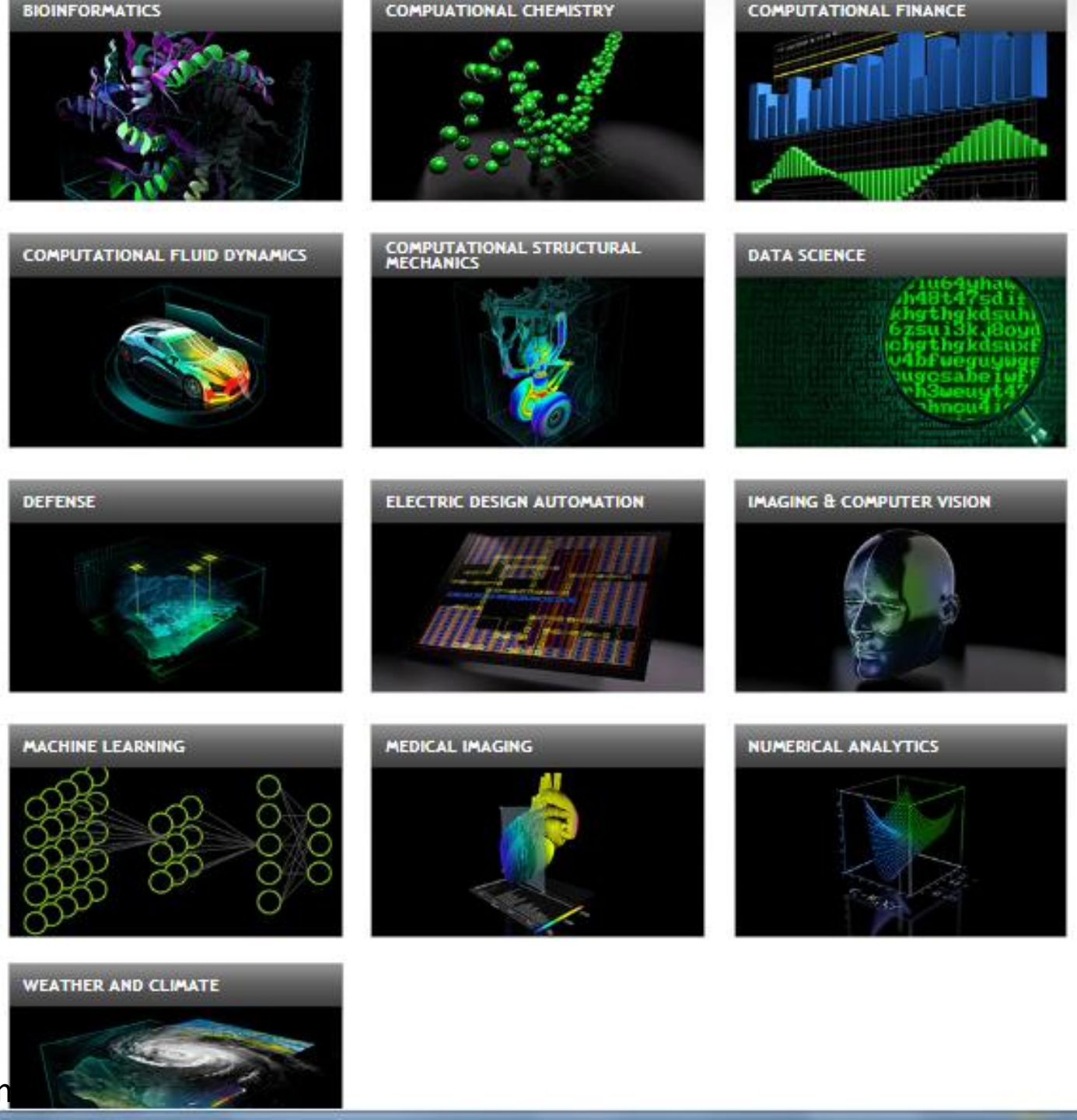
- Serial code executes in a Host (CPU) thread
- Parallel code executes in many Device (GPU) threads across multiple processing cores



Courtesy: Prof. Rosni Abdullah (USM)

Example Applications

Astronomy, biology, chemistry, physics, data mining, manufacturing, finance, weather modelling etc…



http://www.nvidia.com/object/gpu-applications-domain.htm