Nghiên cứu và tìm hiểu kiến trúc GPU Nvidia

# Tổng quan về điện toán GPU (GPU computing)

## GPU là gì?

GPU viết tắt cho Graphics Processing Unit được gọi là Đơn vị xử lý đồ họa ( đôi khi được gọi VPU – Visual Processing Unit) là một bộ vi xử lý chuyên dụng nhận nhiệm vụ tăng tốc, xử lý đồ họa cho bộ vi xử lý trung tâm CPU.

Các GPU hiện đại có năng suất rất cao trong xử lý đồ họa máy tính. Với cấu trúc mang tính xử lý song song mạnh mẽ của mình, GPU cho thấy nó hiệu quả hơn CPU rất nhiều trong nhiều thuật toán phức tạp.

GPU được sử dụng trong các hệ thống nhúng, điện thoại di động, máy tính cá nhân, máy trạm, máy chơi game v…v.. Trong máy tính cá nhân, một GPU có thể xuất hiện ở card đồ họa, hoặc nó cũng có thể được gắn trên mainboard.

## Lịch sử phát triển của điện toán GPU (GPU computing)

GPU được đưa ra và giới thiệu lần đầu tiên bởi NVIDIA vào 31-8-1999 được xem như bộ xử lý song song phổ biến nhất ngày nay. Dựa trên những mong muốn chưa được thỏa mãn về khả năng đồ họa như thật, thời gian thực, GPU đã phát triển trở thành bộ xử lý có thể biểu diễn các tính toán dấu chấm động (floating-point) và khả năng lập trình chưa từng có tiền lệ trước đây. Những GPU ngày nay đã thật sự vượt qua các CPU trong khả năng tính toán số học và băng thông bộ nhớ, khiến chúng trở thành bộ xử lý lý tưởng để tăng tốc đa dạng các ứng dụng xử lý dữ liệu song song.

Các nỗ lực để khai thác GPU cho các ứng dụng không mang tính đồ họa đã được tiến hành trong khoảng thời gian 1999-2000. Đặc biệt trong khoa học máy tính, cùng với các nhà nghiên cứu trong các lĩnh vực như hình ảnh y khoa và điện từ đã bắt đầu sử dụng GPU để chạy các ứng dụng tính toán mục đích chung (general purpose). Họ đã nhận thấy sự xuất sắc trong khả năng biểu diễn các tính toán dấu chấm động của các GPU dẫn đến tăng hiệu suất rất lớn cho 1 loạt ứng dụng khoa học. Đây chính là sự ra đời của GPGPU – General Purpose computing trên các GPU.

Nhưng bên cạnh đó cũng nảy sinh một số vấn đề, đó là GPGPU đòi hỏi sử dụng các ngôn ngữ lập trình đồ họa như OpenGL, Cg, DirectX và nhiều thuật toán xử lý dữ liệu song song khác nhau để lập trình GPU. Các lập trình viên phải làm cho những ứng dụng khoa học của mình giống như các ứng dụng đồ họa và ánh xạ các chúng thành các vấn đề

Graphics chips started as fixed function graphics pipelines. Over the years, these graphics chips became increasingly programmable, which led NVIDIA to introduce the first GPU or Graphics Processing Unit. In the 1999-2000 timeframe, computer scientists in particular, along with researchers in fields such as medical imaging and electromagnetics started using GPUs for running general purpose computational applications. They found the excellent floating point performance in GPUs led to a huge performance boost for a range of scientific applications. This was the advent of the movement called **GPGPU** or General Purpose computing on GPUs.  
  
The problem was that GPGPU required using graphics programming languages like OpenGL and Cg to program the GPU. Developers had to make their scientific applications look like graphics applications and map them into problems that drew triangles and polygons. This limited the accessibility of tremendous performance of GPUs for science.  
  
NVIDIA realized the potential to bring this performance to the larger scientific community and decided to invest in modifying the GPU to make it fully programmable for scientific applications and added support for high-level languages like C, C++, and Fortran. This led to the **CUDA architecture** for the GPU.

## Những lĩnh vực ứng dụng của GPU

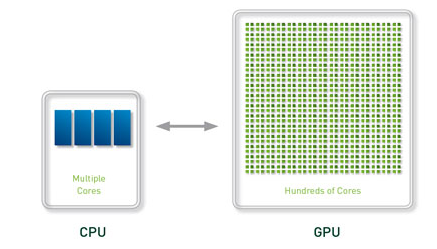
# Kiến trúc GPU

## Kiến trúc G80

## Kiến trúc GT200

## Kiến trúc Fermi

# GPU và CPU



# Các thuật toán song song áp dụng trên GPU

Chúng em sẽ nghiên cứu thuật toán **Parallel Prefix Sum (Parallel Scan)** và demo một bài toán trên GPU có áp dụng thuật toán này.

Hướng nghiên cứu của chúng em như vậy có ok không thầy? Mong thầy cho chúng em thêm ý kiến để chúng em có thể tiếp tục đề tài của mình. Cám ơn thầy.