nVidia cho rằng tiềm năng của điện toán song song là vô cùng to lớn, đặc biệt là khi các tác vụ xử lý ngày càng phức tạp mà việc xử lý bằng một vài nhân vật lý đơn lẻ của CPU khó lòng đáp ứng được. Tốc độ tính toán song song của các GPU đã vượt ra khỏi giới hạn tính toán thông thường, nâng cao khả năng hoạt động cho các chương trình điện toán, đồ họa và kỹ thuật ở cấp độ thời gian thực. <http://vozforums.com/showthread.php?t=1398907>

<http://my.opera.com/incom/blog/show.dml/353454>

<http://planet64bit.de/NVIDIA_reveals_first_details_of_their_next_generation_GPUs>

|  |
| --- |
| **Nvidia ra mắt cấu trúc GPU Fermi - Công nghệ GPGPU đã có chỗ đứng - phần cuối** |
| **Xét từ nhiều góc độ thì SM ( Streaming Multiprocessors ) chính là trái tim của Fermi. SM có khả năng lấy những câu lệnh, vì thế chúng là những “nhân xử lý” thực sự của GPU**  Xét từ nhiều góc độ thì SM ***( Streaming Multiprocessors )*** chính là trái tim của Fermi. SM có khả năng lấy những câu lệnh, vì thế chúng là những “nhân xử lý” thực sự của GPU. Fermi có tất cả 16 SM, và chúng có khả năng xử lý song song nội bộ mạnh hơn những lõi của CPU.  Các nhóm tuyến ***( Thread Group )*** được xử lý song song bởi đơn vị thực thi GPU là thành phần quan trọng trong GPU. Nvidia vẫn giữ độ rộng 32-Thread trong Fermi, nhưng giờ đây một SM có hai bộ lập trình ***( Scheduler )*** nhóm tuyến và những bộ phận  gửi câu lệnh ( ***Dispatch*** ) .  /  Sơ đồ khối của một SM  Vậy là SM có tổng cộng 4 đơn vị thực thi chính. 2 trong số đó là các nhóm thành hai cặp ***16 “lõi CUDA” vô hướng*** ,theo ngôn ngữ của Nvidia, và được gọi là “lõi – ***Core***” trong biểu đồ khối. Ngoài ra còn có 16 Bộ phận tải/lưu trữ ***( Load/Store Unit – LSU )*** và 04 nhóm các đơn vị chức năng đặc biệt ***( Special Func Unit – SFU )*** . Các ***SFU*** xử lý các dạng toán đặc biệt , với số lượng tăng gấp đôi so với GT200 vốn chỉ có 2 SFU trên một SM.  Hai khối ***Scheduler / Dispatch*** đều nối trực tiếp tới những 04 bộ phận thực thi ( những Core ) vì thế mà mỗi***Scheduler / Dispatch*** có thể gửi một Thread Group tới bất kỳ đơn vị nào trong số 4 đơn vị thực thi trong một chu kỳ xung nhịp. Điều này khiến Fermi trở thành một thiết kế xử lí kép  thực sự chứ không như GT200 là mô phỏng xử lí kép . Trường hợp ngoại lệ duy nhất là khi dùng phép tính đòi hỏi có độ chính xác gấp đôi ( Double-Precision ) .  Khả năng chia sẻ dữ liệu địa phương trong SM của Fermi cũng lớn hơn, 64KB so với 16KB của GT200. Đây cũng là một hướng đi khá thông minh.  /  CUDA Core  Tuy nhiên đầu tiên hãy xem xét cái gọi là “***nhân CUDA – CUDA Core ”*** Mỗi trong số các nguồn thực thi vô hướng này đều có đường dẫn dữ liệu dấu phẩy động riêng và đường dữ liệu số nguyên riêng. Đơn vị xử lí những phép tính số nguyên đứng ***( INT Unit )*** riêng chứ không còn được gộp vào với đơn vị MAD như trong các thiết kế trước. Và mỗi Đơn vị xử lí các phép tính dấu phẩy động ***( FP Unit )*** giờ đây cũng có khả năng tạo ra kết quả FP có độ chính xác gấp đôi tương thích IEEE 754-2008 trong 2 chu kỳ xung nhịp, tức bằng một nửa tốc độ xử lí của phép tính có độ chính xác đơn ( Single – Precision ) .  Đây là một bước tiến lớn so với một đơn vị DP ( ***Dispatch Port*** ) duy nhất trên mỗi SM của GT200 – do đó có thể tạm thời đánh giá Fermi có DP với hiệu suất gấp 10 lần so với GT200 . Lại một lần nữa, việc đưa khả năng tính toán có độ chính xác gấp đôi vào thực sự là một lời cam kết của Nvidia, bởi sự chính xác như vậy thường chẳng có tác dụng gì trong đồ họa thời gian thực, và chỉ hữu dụng với các ứng dụng tận dụng sức mạnh GPU để tính toán mà thôi – hỗ trợ GPGPU.  Fermi vẫn giữ lại mô hình tính toán kiểu SIMT ( Single instruction – Multiple thread ) . Mỗi tuyến trong ***Thread Group*** sẽ thực thi liên tiếp trên “nhân  CUDA,” trong khi 15 tuyến còn lại cũng xử lý tương tự như vậy và song song với nhau .  Với đồ họa, mỗi pixel được coi như một tuyến, và thành phần màu pixel được xử lý lần lượt như sau: đỏ, xanh lá, xanh và alpha. Do mỗi ***Thread Group*** gồm 32 tuyến nên hoạt động của chúng mất tối thiểu 2 chu kỳ xung nhịp trong Fermi.  Như vậy mỗi SM có thể xử lí được 48 ***Thread Group*** cùng một lúc mà một ***Thread Group*** lại có 32-Thread , trong Fermi có tới 16SM vì thế nó có thể xử lí tới 24.576 Thread cùng một thời gian .  **Độ chính xác và khả năng lập trình được cải thiện**  /Fermi đã đưa thêm rất nhiều cải tiến để nâng cao độ chính xác của các phép toán, trong đó có hỗ trợ phép toán nhân-cộng hỗn hợp (FMA) với cả 2 phép tính có độ chính xác đơn và độ chính xác đôi. FMA đã cải thiện độ chính xác bằng cách tránh làm tròn giữa phép tính nhân và phép tính cộng, trong khi vẫn giữ được kết quả tức thời chính xác hơn nhiều. Về mặt này Fermi khá giống với Cypress của AMD, cả hai đều tương thích với chuẩn IEEE 754-2008. Cũng như Cypress, Fermi còn hỗ trợ được cả những phép tính với giá trị gần bằng 0.  Tập lệnh bên trong Fermi cũng được mở rộng theo rất nhiều cách với hỗ trợ phần cứng cho cả ***OpenCL***và***DirectCompute***. Các thay đổi này khiến PTX – ISA Nvidia được tạo ra để xây dựng cho ứng dụng tính toán CUDA -- cũng phải cập nhật theo. PTA là mức thấp của ISA , nhưng không hẳn ở mức mã máy. Tuy nhiên các ứng dụng CUDA vẫn có thể gộp vào PTX.  Nvidia vẫn liên tục hỗ trợ OpenCL và DirectCompute. Trong số các thay đổi trong PTX 2.0 có không gian địa chỉ 40-bit 1TB. Không gian địa chỉ đơn này bao gồm các không gian bộ nhớ từng tuyến, từng SM (hoặc xung nhịp) tích hợp vào mô hình lập trình CUDA, với một tập lệnh tải và lưu trữ. Các lệnh này hỗ trợ địa chỉ 64-bit, đủ chỗ cho nhu cầu tương lai. Theo Nvidia, các cải tiến này sẽ cho phép xử lý chính xác C++ và PTX 2.0 cũng thực hiện nhiều cải tiến khác để tăng cường hỗ trợ C++.  **Phân tầng bộ nhớ**  /Như đã nói, mỗi SM có 64KB SRAM liên kết với nhau . Điều thú vị là Fermi đặt bộ nhớ này giữa phân lưu trữ dữ liệu truyền thống và Cache L1, có thể là 16KB bộ nhớ chia sẻ và 48KB cache hoặc ngược lại. Để thay đổi mode phải tiến hành khi chip ở trạng thái nghỉ. Phần lưu trữ cấu hình cục bộ hoạt đọng như Cache có chức năng như một cache L1 thực sự, gắn kết từng SM nhưng không phải toàn bộ, để củng cố thêm mô hình lập trình CUDA.  Hỗ trợ cache L1 trong Fermi là một cache L2 768KB, kết nối với tất cả các SM. Toàn bộ các truy cập bộ nhớ đều đi qua bộ nhớ Cache này, và chip sẽ đi qua DRAM trong trường hợp lỡ cache. Do đó cache này có tác dụng như một nơi chia sẻ dữ liệu tốc độ cao.  Cache L2 đặc biệt hữu dụng khi các tuyến từ nhiều SM khác nhau cùng truy cập một loại dữ liệu, khi đó Cache có thể cung cấp băng thông cực lớn vốn đã lớn sẵn trong cấu trúc. Nvidia cũng đưa ra vài ví dụ về những thuật toán được lợi từ cache nhờ khả năng truy cập bộ nhớ bất thường và ngẫu nhiên, các thuật toán này rất phong phú từ ứng dụng cá nhân cho đến tính toán tốc độ cao như: ray tracing, kernel vật lý và phép nhân ma trận rải rác. Phép tính cơ bản cũng trở nên nhanh hơn – Nvidia ước lượng chúng nhanh hơn GT200 từ 5 đến 20 lần – một phần nhờ vào cache L2. (Fermi cũng có nhiều đơn vị phần cứng hơn.)  Ngoài ra toàn bộ phân tầng bộ nhớ, từ file thanh ghi đến cache L1 L2 và mạch điều khiển bộ nhớ 64 bit, đều được bảo vệ bởi ECC. Hỗ trợ ECC mạnh đáp ứng nhu cầu đối với nhóm nhiều máy tính lớn, giống như các máy tính trên thị trường HPC, và đây lại là một ví dụ cho việc Nvidia sử dụng transistor vào các tính năng chuyên về tính toán. Trên thực tế, việc hỗ trợ ECC không hoàn toàn có ý nghĩa với các GPU nhỏ hơn dành cho thị trường đồ họa cá nhân.  Fermi hỗ trợ ECC sửa lỗi một lớp và phát hiện lỗi hai lớp cho cả GDDR5 và DDR3. Nhưng hiện vẫn chưa biết Nvidia sử dụng cơ chế sửa lỗi nào. Công ty này từ chối cho biết liệu các giao diện bộ nhớ này có phải rộng 72 bit hay không, mà chỉ nói giao diện bộ nhớ là “64 bit chức năng.” Tuy nhiên Fermi cũng có lớp bảo vệ thực sự cho các lỗi nhỏ trong bộ nhớ, vì thế đây không chỉ là một hệ thống sửa lỗi nền tảng CRC tích hợp vào giao thức chuyển GDDR5.  /Như đã nói, không gian địa chỉ ảo và vật lý của Fermi là 40-bit, nhưng giới hạn vật lý thực sự cho kích thước bộ nhớ của loại chip này lại tùy vào số lượng các thiết bị sử dụng nó. Giới hạn thực tế là 6GB cho bộ nhớ 2Gb và 12GB cho thiết bị 4Gb.  Tất nhiên CPU cũng sẽ liên lạc với phần còn lại của hệ thống. Fermi cũng thừa nhận thực tế đó với một giao diện cải tiến trong hệ thống chủ bao gồm các hệ thống độc lập, chuyên biệt dành cho việc chuyển dữ liệu qua lại với GPU, cho phép truyền dữ liệu đồng thời GPU-hệ thống chủ và ngược lại.  **Tiếp theo sẽ là gì?**  Nvidia vẫn đang tiếp tục phát triển phần mềm CUDA. Tại hội thảo công nghệ GPU vừa rồi, Nvidia cũng ra mắt nền tảng phát triển Nexus với một plug-in Microsoft Visual Studio dành cho CUDA. Với Fermi, việc sửa lỗi bằng các công cụ như vậy dễ dàng hơn nhiều.  */*  Sự đầu tư của Nvidia vào các công cụ phần mềm cho GPU thực sự bỏ xa AMD. Tuy rằng điều này dẫn đến nhiều tranh cãi về tiêu chuẩn hỗ trợ nhưng có vẻ quan tâm hàng đầu của Nvidia là đảm bảo rằng mọi điều cần thiết cho việc lập trình GPU của họ đều phải được hỗ trợ tốt, cho dù đó là PhysX và C cho CUDA hay  OpenCL và DirectCompute.  Đây cũng là một phần trong chiến lược khai thác thị trường mới về GPU, và công ty này hy vọng sẽ giành được nhiều thành công về lĩnh vực này. Theo ước tính của Nvidia, toàn bộ thị trường GPU có giá trị hơn $1.1 tỉ, cả về thị trường HPC truyền thống , giáo dục và quốc phòng. Tuy các dự án này có thể gây tranh cãi nhưng chúng cũng cho biết nhiều điều về động cơ của Nvidia phía sau cấu trúc Fermi.  Vẫn còn rất nhiều điều chúng ta chưa biết về GPU tiếp theo của Nvidia, trong đó có những thông tin quan trọng về tính năng đồ họa và tốc độ thực tế. Hồi đầu tháng 9 năm nay, những con chip đầu tiên đang trong quá trình hoàn thành. Tùy vào tốc độ thực hiện quy trình, có thể chúng ta sẽ có được những sản phẩm này vào cuối năm nay hoặc đầu năm sau.  Và khi Fermi ra mắt, ít nhất nó cũng đuổi kịp Cypress của AMD về hỗ trợ OpenCL và DirectCompute API, cùng độ chính xác trong các phép toán tương thích IEEE 754-2008. Đối với một số lĩnh vực trong thế giới GPU, có thể Fermi rất đáng để chờ đợi nhờ vào sự vượt trội trong tốc độ tính toán chính xác bội, băng thông bộ nhớ, cache và hỗ trợ ECC \_ cùng nhiều công cụ phần cứng cũng như phần mềm khiến cho sản phẩm này có được khả năng lập trình tuyệt vời chưa từng có trong GPU. |

<http://www.tuvantinhoc1088.com/print.php?id=827>