Fermi

معمـــاری جـــدید nVIDIA

هومن سیاری
Sayyari@ComputerNews.ir

آخرین اخبار منتشـره از شـرکت nVIDIA حاکی از آن است که این شرکت هـدف اصلی خود را روی راندمان بـالای پردازش، مخصوصاً در گیم قرار داده و این هدف را از طریق معماری جدید Fermi دنبال می‌کند. البته ویژگی‌های جدید و مشـخصات بهبود یافته در Fermi، سـرنخ‌هایی را از آنچه که باید در GPUهای آینده در انتظار آن باشیم، به ما می‌دهد.

شـرکت nVIDIA، معماری Fermi را در اواخر سـال ۲۰۰۹ ارایه کرد و بیشـتر توصیف‌های اولیه از آن با محوریت Tesla انجام می‌شد که شامل محصولات پرقدرت nVIDIA است. GPUهای Tesla، به صورت اختصاصی برای سوپرکامپیوترها به کار می‌روند، چرا که پردازش‌های سنگین با استفاده از آنها انجام می‌شود و در واقع کمک حال پردازنده CPU خواهند بود. هدف از GPUهای Tesla، انجام کارهای گرافیکی مثل گیم نیسـت، بلکه هدف انجام محاسـبات سـنگین و پیچیده است. nVIDIA قصد دارد GPUهای Fermi را براسـاس اطلاعاتی که در نمایشـگاه CES 2010 ارایه کرد، مخصوص کار روی گیم و گرافیک‌های سـه‌بعدی تعریف کند. بنابراین مجبوریم به همین اطلاعات اولیه‌ای که از Fermi و Tesla و HPC (پردازش‌های با راندمان بالا) داریم، اکتفا کرده و به توضیح آنها بپردازیم.

Fermi

GPUهای Fermi قصد دارند مسـیری را که با GPUهای G80 شـروع شد و با GPUهای GT200 توسعه پیدا کرد، ادامه دهند. GPUهای GT200 را می‌توان GPUهایی فرض کرد که قدرت پردازشی آنها تقریباً برابر با قدرت گرافیکی آنهاست. GPUهای Fermi دارای مشخصات زیر هستند:

- داشتن حدود ۳ میلیارد ترانزیستور
- توانایی پشتیبانی از حداکثر ۶ گیگابایت GDDR5
- پشتیبانی از DirectX11
- دارای ۵۱۲ هسته پردازشی CUDA
- پشتیبانی از زبان برنامه‌نویسی C++
- پشتیبانی از کدهای تصحیح خطا (ECC)

برخـی از این ویژگی‌هـا برای آنچه که nVIDIA «پـردازش با GPU» می‌نامد، بسیار مفیدند. معماری جدید Fermi در چند سال اخیر و پس از معماری GT200 ظهـور پیـدا کرد. وقتی به تاریـخ نوآوری‌های nVIDIA در سـال‌های اخیر نگاه می‌کنیم، می‌بینیم که تقریباً هر سه سال یک معماری جدید معرفی کرده است.

ریاضیات جدید

شـاید جذاب‌ترین جنبه معماری جدید Fermi برای HPCها آن باشـد که امکان اسـتفاده از آخرین اسـتاندارد اعداد اعشاری که توسـط IEEE تهیه شده را فراهم می‌کند. اسـتاندارد جدید ۲۰۰۸-۷۵۴ اسـت که جایگزین اسـتاندارد ۱۹۸۵-۷۵۴ شـده است. اسـتاندارد جدید شامل دسـتورالعمل FMA اسـت. FMA، عملیات ضرب و جمع را در یک دسـتورالعمل اجرا می‌کند که موجب گرد شدن (Round) دقیق‌تر حاصل عملیات می‌شود؛ در حالی که تا پیش از این از دستورالعمل MAD برای این کار استفاده می‌شد که گاهی باعث از بین رفتن دقت می‌شد.

علاوه بر این، عمل ضرب اعداد صحیح ۳۲ بیتی به طور کامل در ALU واحدهای پـردازش گرافیکـی Fermi انجام خواهد شـد، در حالی که پیـش از این ALU واحدهـای پـردازش گرافیکی ۲۴ بیت بـود و امکان پـردازش ۳۲ بیت را در یک دستورالعمل نداشت.

معمـاری Fermi به خصوص بـرای کاربردهایی با دقت مضاعف از قبیل شـیمی کوانتومـی و جبر خطی طراحی شـده اسـت که البتـه آنها هم بـه HPC تمایل دارنـد. بـا Fermi تـا ۱۶ عملیـات FMA با دقـت مضاعف را می‌تـوان در هر SM (Streaming Processor) و در هـر کلاک انجام داد که این حدود ۴ برابر قدرت معماری GT200 است.

سیاست جدید nVIDIA

بعضی از کارشناسـان نسـبت به تصمیم nVIDIA برای تأخیر در اعلام مشخصات و توانایی‌هـای گرافیک‌هـای Fermi و تمرکز شـدید روی Tesla و HPC اظهار نگرانی و ناراحتی می‌کنند. nVIDIA با GPUهایی که برای گیم و گرافیک طراحی می‌کند، تجارت خود را تضمین کرده اسـت، در حالی که HPCها سـهم کوچکی از درآمد شرکت را تأمین می‌کنند. nVIDIA امیدوار است HPC را به قدری رشد دهد که سهم بزرگی از فروش شرکت را در برگیرد. این شرکت اعتقاد دارد HPC فناوری جدیدی اسـت کـه در آینده همه به آن نیاز پیدا خواهند کرد، به همین دلیل بیشـتر تمایل دارد GPUهای خود را برپایه HPC بسازد. از طرفی فناوری‌های به کار رفته در HPCها هزینه سربار زیادی را برای کاربردهای عمومی ایجاد خواهند کرد، چرا که مثلاً برای گیم و خیلی از کاربردهای عمومی نیازی به ویژگی ECC نیست.

همچنیـن nVIDIA در آینـده دوباره مجبور به طراحـی دو GPU متفاوت برای HPCهـا و کاربران معمولی خواهد شـد. طراحی جداگانه پروژه‌ها بسـیار گران‌تر خواهند بود و اثر استفاده از بازار بزرگ مصرف‌کنندگان را خنثی خواهند کرد.

GPU در مقابل CPU

حتی اگر انجمن HPC به طور کامل GPUهای Fermi و پلتفرم Tesla را بپذیرد، این مهاجرت به فناوری جدید به هیچ وجه به معنای پایان عصر CPU نخواهد بود. هنـوز خیلی مانده تا GPU بتواند تمام کارهای CPU را انجام دهد، هر چند برخی از کارهای آن را بهتر انجام می‌دهد. CPU همچنان سلطان سیستم است و تا آینده قابل پیش‌بینی بر آن حکومت خواهد کرد!

قدم بعدی، ترکیب GPU و CPU در یک چیپ است. این فناوری در سیستم‌های کوچک مثل موبایل‌ها اتفاق افتاده و به زودی در سیسـتم‌های دسـکتاپ نیز شاهد آن خواهیم بود. اینتل با معماری Larrabee و AMD هم با معماری Fusion در این فناوری جدید به رقابت خواهند پرداخت.

CPU مزایـای زیادی دارد. مثلاً بعضی از انواع برنامه‌هـا مثل برنامه‌هایی که نیاز بـه threadهای محـدودی دارند یـا عبارت‌هایی که ترکیب زیـادی از عملگرها (مثـل جمع و تفریق و ...) دارند را به خوبی پردازش می‌کند. همچنین CPUها در برنامه‌هایـی که نیازمند به چندین thread با یک توالی طولانی از دسـتورالعمل‌ها هستند، بهتر عمل می‌کنند.

GPU برای اینکه بتواند جای CPU را بگیرد، باید چندین اشـکال ساختاری مانند اندازه نسبتاً کوچک حافظه و ناتوانی در دسترسی مستقیم (Direct I/O) به حافظه را برطرف کند.

نگاهی دقیق‌تر به معماری Fermi

nVIDIA هنوز مشـخصات نهایی Fermi را منتشر نکرده، اما برخی از مهم‌ترین ویژگی‌های آن را معرفی کرده است. بخشی از آن چیزی که درباره معماری Fermi می‌دانیم را به همراه پیش‌بینی خود براساس دانسته‌هایمان در این جا آورده‌ایم.

GPUهای Fermi حداکثر دارای ۵۱۲ هسته پردازشی CUDA هستند. هر هسته CUDA در هر کلاک یک دستورالعمل اعشاری با دقت تکی یا یک دستورالعمل صحیح را به ازای هر thread اجرا می‌کند. اگرچه nVIDIA هنوز فرکانس کاری Fermi را اعلام نکرده، اما پیش‌بینی می‌شود حدود ۱,۵ گیگاهرتز باشد که در نتیجه می‌تواند تا ۱,۵ میلیارد عملیات ممیز شناور را در هر ثانیه انجام دهد.

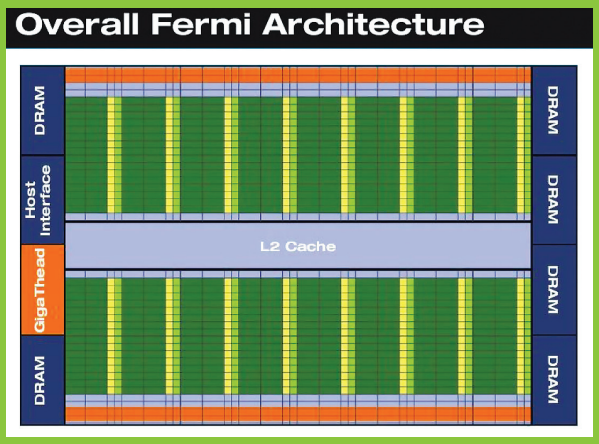
این ۵۱۲ هسته CUDA به ۱۶ گروه تقسیم می‌شوند که به هر گروه SM یا Streaming Processor گفته می‌شود. بنابراین هر SM شامل ۳۲ هسته CUDA خواهد شد (که به آن پردازنده CUDA نیز می‌گویند). در شکل ۱ SMها به صورت مستطیل‌های سبز رنگ نمایش داده شده‌اند و هر کدام شامل ۳۲ هسته هستند. این مقدار ۴ برابر هسته‌های CUDA در هر SM در GPUهای قبلی است، یعنی SMها در GPUهای قبلی حداکثر دارای ۸ هسته بودند. داخل هر هسته CUDA، یک ALU صحیح و یک ALU واحد ممیز شناور قرار دارد.

این ۱۶ عدد SM، یک کش L2 با ظرفیت ۷۶۸ کیلوبایت را مطابق شکل احاطه می‌کنند. هر SM دارای بخش‌های کش، رجیستر (بخش سرمه‌ای رنگ)، Dispatch و Warp (بخش نارنجی رنگ) می‌شود که این دو بخش نهایی، وظیفه آماده شدن برای پردازش بعدی را به عهده دارند.

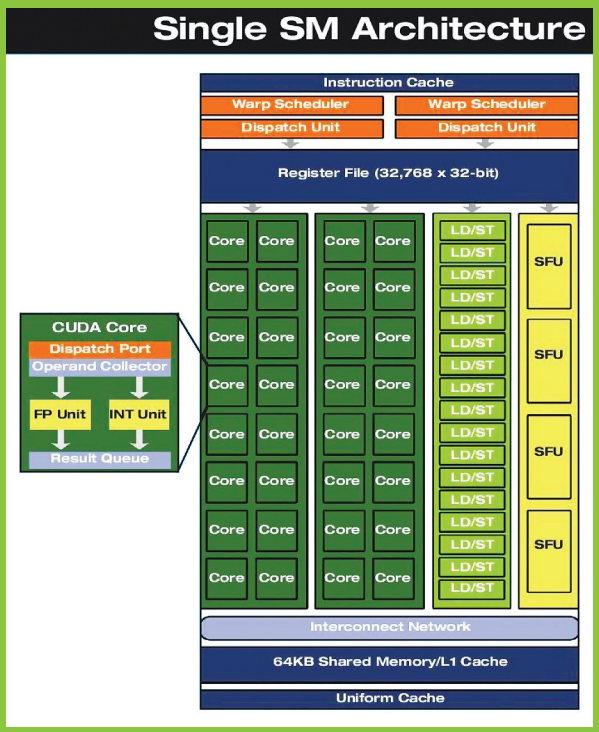
داخل هر SM چهار ستون وجود دارد که شامل ۳۲ هسته CUDA، ۱۶ واحد لود و ذخیره (LD/ST)، و ۴ واحد توابع خاص (SFU) می‌شود (شکل ۲). واحد لود و ذخیره، آدرس‌های مبدأ و مقصد را برای شانزده thread در هر کلاک محاسبه می‌کند. واحد SFU انواع خاصی از دستورالعمل‌ها مثل کسینوس و جذر را محاسبه می‌کند. هر SFU نیز می‌تواند یک دستورالعمل را در هر کلاک اجرا کند.

حداکثر ۳۲ تا thread موازی در هر SM زمان‌بندی می‌شوند تا اجرا شوند که به این کار warp می‌گویند. این زمان‌بندی در واحدی به همین نام یعنی warp انجام می‌گیرد. هر SM دارای ۲ واحد زمان‌بندی warp و ۲ واحد dispatch است و این پیکربندی می‌تواند دو warp را در یک زمان اجرا کند.

GPUهای Fermi دارای ۶ کنترلر حافظه ۶۴ بیتی هستند که در شکل به صورت DRAM نمایش داده شده‌اند. این کنترلرها، داده‌ها را به هسته‌های CUDA تحویل می‌دهند. کنترلرهای حافظه از GDDR5 پشتیبانی می‌کنند و این امر باعث می‌شود Fermi اولین GPU رده بالایی باشد که می‌تواند از GDDR5 استفاده کند. پیش‌بینی می‌شود پهنای باند حافظه Fermi حدود ۱۹۲ گیگابایت بر ثانیه باشد که حدود ۳۰ درصد از Geforce GTX280 بیشتر است. Geforce GTX280 دارای ۸ کنترلر حافظه GDDR3 است. مزیت دیگر حرکت از GDDR3 به GDDR5 این است که GDDR5 پهنای باند بالاتری را به ازای هر پین فراهم می‌کند و این امر موجب می‌شود که Fermi بتواند از رابط حافظه کوچک‌تری استفاده کند و باز پیش‌بینی می‌شود Fermi از یک رابط ۳۸۴ بیتی استفاده کند، در حالی که Geforce GTX280 دارای رابط ۵۱۲ بیتی است. رابط حافظه کوچک‌تر منجر به استفاده از توان مصرفی پایین‌تر و استفاده از تعداد پین‌های کمتر خواهد شد. ارتباط اصلی بین GPU و CPU نیز از طریق یک مسیر PCI-Express 16X انجام می‌گیرد که حداکثر به سرعت ۸ گیگابایت بر ثانیه می‌رسد. ■



شکل ۱: ساختار کلی معماری Fermi



شکل ۲: ساختار هر SM در معماری Fermi

جدول ۱: مقایسه معماری Fermi با معماری‌های موجود

GPU	G80	GT200	Fermi
Transistors	681 million	1.4 billion	3.0 billion
CUDA Cores	128	240	512
Double Precision Floating Point Capability	None	30 FMA ops/clock	256 FMA ops/clock
Single Precision Floating Point Capability	128 MAD ops/clock	240 MAD ops/clock	512 MAD ops/clock
Warp Schedulers / SM	1	1	2
SFUs / SM	2	2	4
Shared Memory / SM	16 KB	16 KB	16 KB or 48 KB
L1 Cache	None	None	16 KB or 48 KB
L2 Cache	None	None	768 KB
ECC Memory Support	No	No	Yes
Concurrent Kernels	No	No	Up to 16