‘

Rendering Pipeline

مقدمه

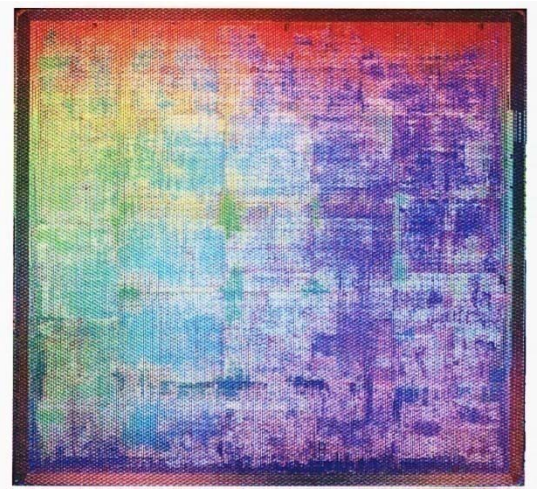
زمانی که معماری پردازندههای گرافیکی تغییر کرد و به معماری Streaming تبدیل شد، شکل محاسبه و نحوه ارتباطات بر روی بردها تغییر زیادی یافت و پس از آن پردازنده‌هایگرافیکی به انجام چند صد گیگافلاپ عملیات Single Precision Floating-Point (به شکلی از دقت نمایش اعداد اعشاری گفته می‌شود که اعداد توسط یک عبارت کامپیوتری قابل نمایش هستند) قادرشدند، این درحالی بود که قوی‌ترین پردازنده‌های آن زمان تنها درحدود 12 گیگافلاپ قدرت داشتند. البته این اختلاف هر روز زیادتر شده و می‌شود و امروز به موقعیتی رسیده‌ایم که کارت‌گرافیکیGTX 295 در حدود 2هزارگیگافلاپ قدرت دارد این در حالی است که قدرت پردازنده‌ی گران‌قیمت شرکت اینتل به نام Intel Core i7 965 Extreme تنها در حدود 70 گیگافلاپ است. در این شماره قصد داریم بخشی از مقاله نوشته شده توسط Emmet Kilgariff و Randima Fernando از شرکت انویدیا در مورد معماری پردازنده‌های گرافیکی جی‌فورس سری شش را مورد بررسی قرار دهیم، دراین مقاله به بررسی برترین محصول این سری یعنی Geforce 6800 خواهیم پرداخت. اگرچه این سری از پردازنده‌های گرافیکی کمی قدیمی هستند اما با مطالعه‌ی این مقاله به درک بهتری از معماری کارت‌های گرافیکی خواهید رسید.

نحوه‌ی قرارگیری یک پردازنده‌ی گرافیکی در سیستم کامپیوترها

در کامپیوترهای مدرن امروزی، پردازنده مرکزی با پردازنده گرافیکی از طریق اتصال PCI-Express یا AGP موجود برروی مادربرد ارتباط برقرار می‌کند. از آنجایی که این اتصال مسئول انتقال داده‌های مربوط به فرمان‌ها، Textureها و Vertexها از پردازنده مرکزی به پردازنده گرافیکی است، این پهنای باند آن در سال‌های اخیر در کنار پردازنده گرافیکی پیشرفت کرده است. اسلات AGP ، 66 مگاهرتز سرعت دارد و داده‌ها را با طول 32 بیت جا به جا می‌کند و در مجموع 264 MB/sec انتقال داده را فراهم می‌آورد. پس از AGP، نسل‌های بعدی این اسلات با نام‌های AGP 2x, 4x, 8x عرضه شدند که با ارائه هر کدام از این نسل‌ها پهنای باند اتصال دوبرابر شد تا این که درسال 2004

،

اسلاتی با حداکثر پهنای باند ممکن از لحاظ تئوری، با نام PCI-Express عرضه گشت. اسلاتی با پهنای باند، 4GB/sec است.



شکل1- یک پردازنده‌ی Geforce 6800

در جدول زیر قصد داریم تفاوت میان پهنای باند واسط حافظه پردازنده گرافیکی را با پهنای باند قسمت‌های دیگر سیستم مقایسه کنیم:

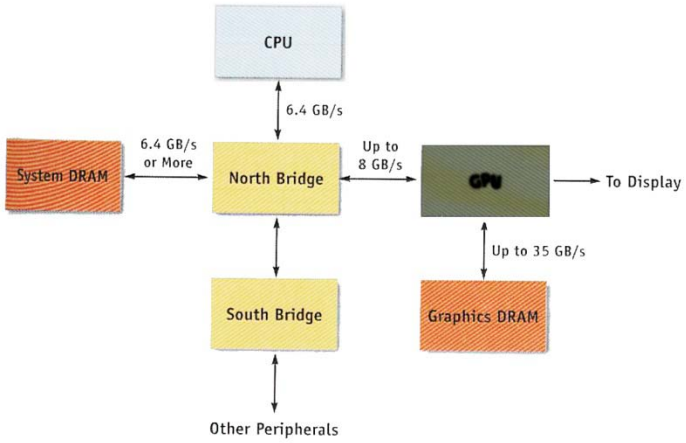
Component	Bandwidth(GB/sec)
واسط حافظه پردازنده گرافیکی	35 GB/sec
PCI Express Bus(x16)	8 GB/sec
واسط حافظه پردازنده مرکزی (800 Mhz FSB)	6.4 GB/sec

این جدول به شکل ویژه، تفاوت محسوس پهنای باند موجود در داخل پردازنده گرافیکی با پهنای باند قسمت‌های دیگر سیستم را نمایش می‌دهد و اهمیت الگوریتم‌هایی را که بر روی پردازنده گرافیکی اجرا می‌شوند و از این پهنای باند عظیم بهره می‌برند را نشان می‌دهد.

نحوه قرار گرفتن پردازنده گرافیکی در معماری سیستم

‘

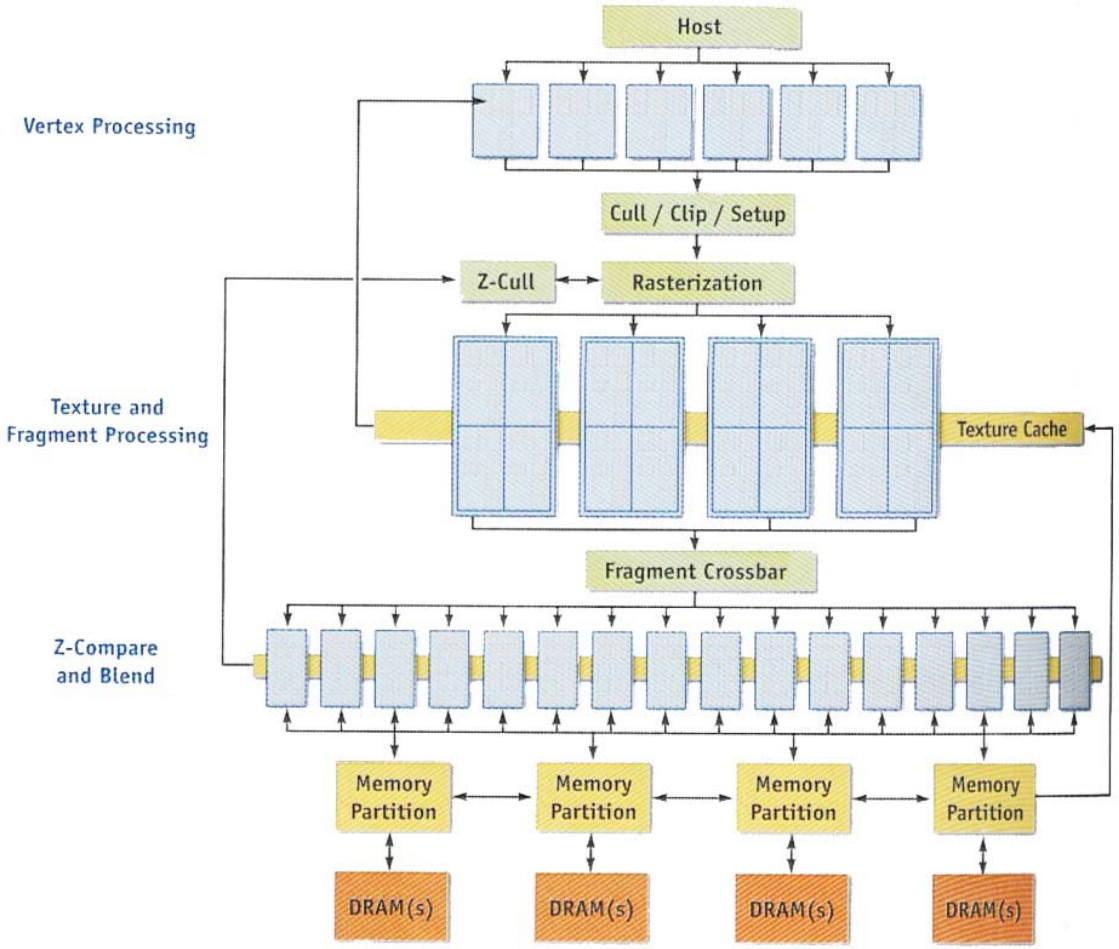
شکل 2 نحوه قرار گرفتن پردازنده گرافیکی درسیستم را نشان می‌دهد:



شکل 2-نحوه قرار گرفتن پردازنده گرافیکی درمعماری سیستم

بلاک دیاگرام نشان دهنده عملیات‌های گرافیکی که انجام می‌شوند

شکل 3 معماری پردازنده گرافیکی Geforce 6 را نمایش می‌دهد. دراین قسمت قصد داریم به بررسی Graphics Pipeline بپردازیم. این Pipeline ورودی‌هایی که توسط CPU ارسال می‌شود را دریافت کرده و پیکسل‌های خروجی را بر روی FrameBuffer نقش می‌بندد.



شکل 3-بلاک دیاگرام معماری Geforce6

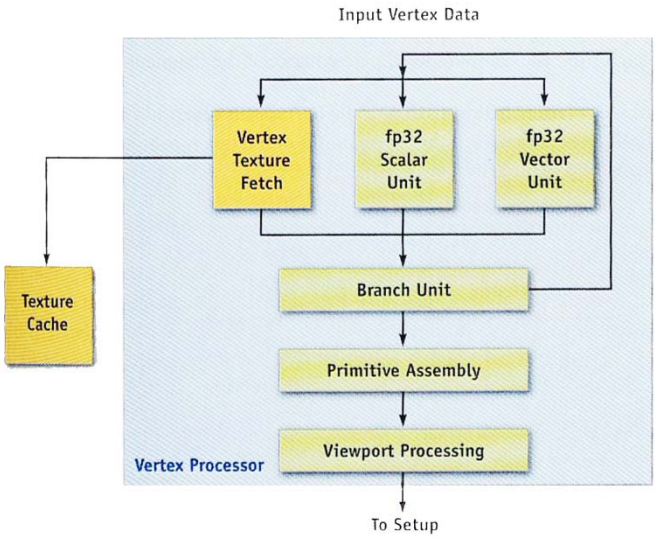
در ابتدا فرمان‌ها ، Textureها و داده‌های مربوط به هر Vertex از طریق بافر مشترک بین پردازنده مرکزی و پردازنده گرافیکی موجود در حافظه سیستم یا حافظه مربوط به فریم‌بافر محلی، توسط پردازنده گرافیکی دریافت می‌شوند. زنجیره‌ای از دستورات توسط پردازنده مرکزی در این حافظه مشترک نوشته می‌شود که وظیفه مقداردهی اولیه و ویرایش state مربوط به عملیات‌های گرافیکی ، ارسال فرمان‌های لازم جهت رندر کردن و ارجاع‌هایی به داده‌های مربوط به Vertexها و Textureها را دارد. فرمان‌ها به اجزا کوچکتر تقسیم می‌شوند و از واحد Vertex Fetch برای خواندن و ذخیره کردن اطلاعات مربوط به Vertexهایی که در فرمان‌های مربوط به عملیات رندر، به آن‌ها ارجاع داده شده است، استفاده می‌شود. فرمان‌ها، Vertexها و Stateها هنگام عبور از Pipeline Stageهای متوالی، تغییر می‌کنند. پردازنده

مخصوصVertexها که Vertex Shader نامیده می‌شود (شکل4)، این امکان را فراهم می‌کند که یک برنامه برروی هر کدام از Vertexهای مربوط به یک شی اجرا شود و تبدیلات، عملیات Skinning و دیگر عملیات‌هایی که برروی Vertexها انجام می‌شوندرا اعمال کند. برای اولین بار در کارت گرافیک‌های سری شش شرکت انویدیا، این امکان فراهم شده است تا برنامه‌هایی که بر روی Vertexها اجرا می‌شوند، بتوانند داده‌های مربوط به Textureها را واکشی(Fetch) کنند. همه عملیات‌ها به صورت عملیات‌های 32 بیت، Floating-Point(fp32) موجود برای هر جز انجام می‌شوند. تعداد پردازنده‌های Vertex که همراه این معماری استفاده می‌شوند قابل تغییر است و مدل‌های دارای کارآیی بیشتر و گرانتر از 6 واحد پردازش Vertex بهره می‌برند و مدل‌های ارزان قیمت‌تر از 2 واحد استفاده می‌کنند.

از آنجایی که پردازنده‌های Vertex، می‌توانند به Textureها دسترسی پیدا کنند، موتورهای Vertex به حافظه‌ی کش مربوط به Textureها متصل هستند. همچنین حافظه کش دیگری وجود دارد که داده‌های مربوط به Vertexها را قبل و بعد از انجام پردازش توسط پردازنده‌های Vertex ذخیره می‌کند تا تعداد عملیات واکشی و محاسبات کاهش یابند، به این ترتیب که اگر اندیس یک Vertex در استقاده از تابع Draw دوبار تکرارشود(مثلا در هنگام استفاده از روش اتصال Strip، برای ساخت یک سه ضلعی)، برنامه‌ای که برروی Vertex عمل می‌کند، نیازی به انجام دوباره عملیات برروی Vertex نخواهد داشت د و از نتیجه موجود در حافظه کش استفاده می‌کند.

سپس Vertexها گروه‌بندی شده و باعث شکل گیری اشکال ابتدایی نقطه، خط و سه ضلعی می‌شوند. سپس بلاک‌های Cull، Clip وSetup، عملیات‌هایی را برروی هر شکل ابتدایی انجام می‌دهند. بلاکی که مسئولیت عملیات Culling را برعهده دارد، اشکال ابتدایی که قابل رویت نیستند را حذف می‌کند، بلاک Clipping، اشکال ابتدایی که با هرم‌ناقص شکل مشاهده صحنه، برخورد می‌کنند را مشخص می‌نماید و بلاک Setup، معادلات مربوط به لبه‌ها و صفحه‌ها را به گونه‌ای سازماندهی می‌کند که داده‌ها برای عملیات Rasterization آماده شوند. بلاک Rasterization محاسبه می‌کند کدام پیکسل و یا Sample (اگر MultiSampling فعال باشد) توسط کدام شکل ابتدایی

پوشانده شده است و به سرعت با استفاده از بلاک z-cull برای دور انداختن پیکسل‌ها و یا Sampleهایی که توسط شی دیگری که دارای عمق کمتری است، پوشانده شده‌اند، اقدام می‌کند. پیکسلی که عملیات برروی آن انجام خواهدشد به عنوان Fragment شناخته می‌شود. این Fragment، از Fragment Processor و تست‌های مختلف عبور می‌کند و اگر از تمام آن‌ها با موفقیت خارج شود، حاوی رنگ و عمق مربوط به پیکسل موجود برروی فریم‌بافر و یا Render Target خواهد بود.



شکل 4- پردازنده Vertex مربوط به Geforce6

شکل5 نشان دهنده یک Fragment Processor (که معمولا با نام Pixel Shader شناخته می‌شود) و یک Texture Pipeline است.واحدهای Fragment Processing و Texture Processing، در کنار یکدیگر عمل می‌کنند تا یک برنامه shader را برروی هر Fragment به صورت مجزا اجرا کنند. معماری Geforce6 به گونه‌ای است که می‌توان به آن تعداد متغیری Fragment Processor متصل کرد یعنی پردازنده های گرافیکی سری Geforce 6 دارای تعداد متفاوتی Fragment Pipeline یا Pixel Pipeline می‌باشند. همانند Vertex Processorها، داده‌های مربوط به Textureها نیز برروی تراشه، در درون کش ذخیره می‌شوند تا از پهنای باند خارجی کمتری استفاده نمایند و کارآیی افزایش یابد.

واحدهای پردازشTextureو Fragment در هر لحظه برروی چهارضلعی‌هایی که از چهار پیکسل تشکیل شده‌اند(که Quad نامیده می‌شوند) عمل کرده و مشتقاتی از آن را محاسبه می‌کنند که برای تعیین سطح

جزییات مربوط به آن بخش از Texture، قابل استفاده است. همچنین Fragment Processor در هر لحظه برروی گروه‌هایی متشکل از صدها پیکسل عمل می‌کند و با یک دستور همزمان پردازش را برروی چندین داده انجام می‌دهد به این پردازنده‌ها SIMD(Single Instruction Multiple Data) گفته می‌شود (که درآن یک موتور پردازش Fragment در هر لحظه فقط برروی یک Fragment کار می‌کند) و با این کار تاخیر مربوط به عملیات واکشی، Texture برکارآیی محاسباتی پردازنده Fragment تاثیری نخواهد داشت.

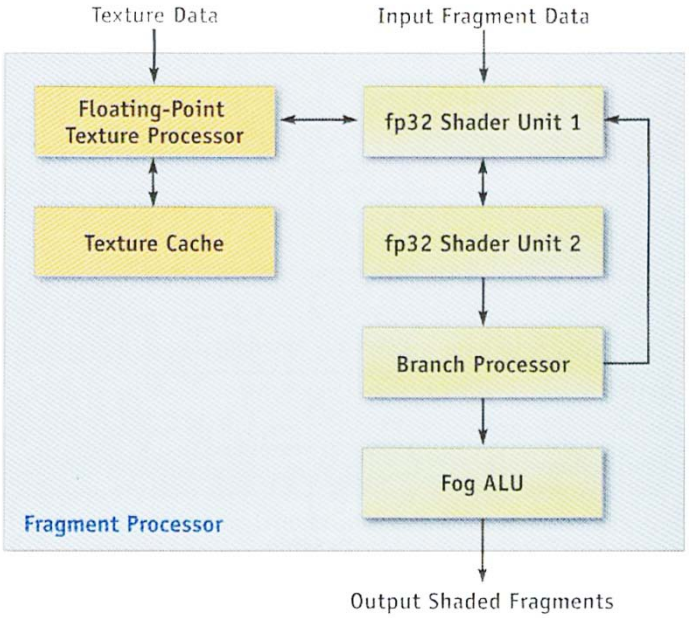
پردازنده Fragment، از واحد Texture برای واکشی داده از حافظه استفاده می‌کند. واحد Texture از فرمت‌های مختلفی پشتیبانی می‌کند (فرمت تعیین کننده شکل ذخیره شدن جزییات هر پیکسل یا نحوه کاربرد آن می‌باشد)، سپس این داده‌ها می‌توانند توسط روش فیلترینگ bilinear، trilinear یا Anisotropic، فیلتر شوند. همه داده‌ها در نهایت به شکل fp16 یا fp32 به پردازنده Fragment ارسال می‌شوند. یک Texture را می‌توان به صورت آرایه‌ای دوبعدی یا سه بعدی از داده‌ها که توسط یک واحد Texture در محل‌های دلخواه حافظه قرار می‌گیرند فرض کرد و تحت عملیات فیلترینگ به Texture دیگری تبدیل کرد که مجددا می‌تواند توسط همین بخش عملیاتی به Texture دیگری تبدیل شود. این بخش عملیاتی به صورت متوالی می‌تواند برروی یک Texture عمل کند. سری Geforce6 قادر است عملیات فیلترینگ را برروی Textureهایی با فرمت fp16 به صورت سخت‌افزاری انجام دهد.

واحد پردازش Fragment دارای دو واحد fp32 shader در هر Pipeline است و Fragmentها قبل از این که مجددا به Pipeline برگردند از این دو واحد fp32 shader و Branch Processor عبور می‌کنند و سپس مجددا به Pipeline برگشته تا سری بعدی دستورات بر روی آن‌ها اعمال شود. این عملیات بازگشت در هر کلاک فقط یک بار انجام می‌شود. همچنین می‌توان از واحدfp32 shader اول، برای اصلاح مختصات Texture مربوط به هر vertex جهت ایجاد پرسپکتیو در صحنه استفاده نمود یا برای انجام عملیات‌های ضرب از آن بهره برد. می‌توان هشت یا تعداد بیشتری عملیات ریاضی را در یک کلاک با Pixel Shader انجام داد و اگر عملیات واکشی texture در واحد اول shader رخ دهد، می‌توان 4 عملیات ریاضی را در هر کلاک انجام داد.

در آخرین مرحله از Pixel Shader Pipeline، واحد ایجاد مه برای ترکیب مه و تصویر استفاده می‌شود، بدون این که کارآیی پردازنده گرافیکی، ذره‌ای کاهش پیدا کند. عملیات ترکیب مه، و تصویر حاصل شده تا این مرحله، در نرم‌افزارهای گرافیکی به وسیله تابع زیر انجام می‌شود:

رنگ مه × میزان رنگ مه + رنگ پیکسل حاصل شده تا این مرحله × (میزان رنگ مه - 1) = رنگ پیکسل خروجی

Fragment هایی که از واحد پردازش Fragment خارج می‌شوند به همان ترتیبی که Rasterize شده‌اند به واحد مقایسه z و Blend فرستاده می‌شوند که درآن‌جا عملیات تست عمق (مقایسه مقدار z و به روزرسانی آن)، عملیات‌های Stencil، ترکیب آلفا و نوشتن رنگ نهایی در صفحه مقصد (یک render target یا بافر رندر که به صورت Off-Screen است و بعد از این که کاملا رندر شد on-Screen می‌شود- برروی صفحه نمایش به تصویر درمی‌آید-) می‌باشد انجام می‌شود.



شکل 5- پردازنده‌ی Fragment و Texel Pipeline مربوط به Geforce6

حافظه سیستم حداکثر به چهارقسمت تقسیم شده است و هرکدام از آن‌ها به DRAMهایی متصل هستند. پردازنده‌های گرافیکی به جای استفاده از RAMهای متعارف در بازار از DRAMهای استاندارد بهره می‌برند. این کار باعث صرفه‌جویی و کاهش

قیمت کارت گرافیکی می‌شود. استفاده از قطعات کوچکتر ومستقل حافظه، باعث می‌شود تا سیستم حافظه در هر صورت چه درحالت انتقال بلاک‌های بزرگ و چه بلاک‌های کوچک از حافظه بهینه‌تر عمل کند. همه صفحه‌های رندر شده در DRAMها ذخیره شده اند در حالی که داده‌های مربوط به Textureها و ورودی‌ها می‌توانند در DRAMهای پردازنده گرافیکی یا سیستم ذخیره شوند. حافظه چهار تکه، برای پردازنده گرافیکی پهنای‌باندی معادل 256-bit فراهم می‌کند. سیستم مربوط به این حافظه، امکان ارسال داده‌های نسبتا کوچک 32 بیتی به صورت زنجیره‌ای با محدودیت سخت‌افزاری 35 GB/sec را فراهم می‌کند.