1/6/2026

پردازنده گرافیکی چیست

هر آنچه باید درباره GPU بدانید

محمدرضا پورمحمد روح افزا

درس معماری کامپیوتر

استاد رشادت جو

آذر ۱۴۰۴

پردازنده گرافیکی (GPU) چیست:

اگر واحد پردازش مرکزی یا پردازنده (CPU) را به‌عنوان مغز کامپیوتر در نظر بگیریم که تمامی محاسبات و دستورها منطقی را مدیریت می‌کند، واحد پردازش گرافیکی یا پردازنده‌ی گرافیکی (GPU)را می‌توان به‌عنوان واحدی برای مدیریت خروجی بصری و گرافیکی محاسبات و دستورها و اطلاعات مرتبط با تصاویر دانست که ساختار موازی آن‌ها برای پردازش الگوریتم بلوک‌های بزرگ داده، بهینه‌تر از واحدهای پردازش مرکزی یا همان پردازنده‌ها عمل می‌کند؛ در واقع GPU، رابطی گرافیکی برای تبدیل محاسبات صورت‌گرفته توسط پردازنده به شکلی قابل فهم برای کاربر به حساب می‌آید و می‌توان با اطمینان گفت هر دستگاهی که به نحوی خروجی گرافیکی را نمایش می‌دهد، به نوعی از پردازنده‌ی گرافیکی مجهز است.

واحد پردازش گرافیکی در یک کامپیوتر، می‌تواند روی کارت گرافیک یا روی مادربرد تعبیه شده باشد یا همراه با پردازنده در تراشه‌ی مجتمع(برای مثالAPU های AMD) عرضه شود. تشخیص مدل کارت گرافیک در ویندوز با سریعترین روش نیز امکان پذیر است کافی است به مقاله لینک شده مراجعه کرده و آن را مطالعه کنید.

تراشه‌های مجتمع نمی‌توانند خروجی گرافیکی آنچنان چشمگیری تولید کنند و قطعاً خروجی آن‌ها هیچ گیمری را راضی نمی‌کند؛ برای بهره‌مندی از جلوه‌های بصری با کیفیت‌تر باید کارت گرافیکی (در ادامه با تفاوت‌های پردازنده‌ی گرافیکی و کارت گرافیک بیشتر آشنا خواهیم شد) مجزا با قابلیت‌هایی فراتر از یک پردازنده‌ی گرافیکیِ ساده تهیه کرد. در ادامه با چند مفهوم اولیه‌ی پرکاربرد در بحث گرافیک‌ها، به‌طور خلاصه آشنا می‌شویم.

تصویر سه‌بعدی

به تصویری که علاوه بر طول و عرض، عمق هم داشته باشد، تصویری سه‌بعدی گفته می‌شود که در مقایسه با تصاویر دوبعدی مفاهیم بیشتری را به مخاطب منتقل می‌کند و اطلاعات بیشتری دارد. برای مثال اگر به مثلثی نگاه کنید، تنها سه خط و سه زاویه مشاهده می‌کنید، اما اگر جسمی هرمی‌شکل داشته باشید، ساختاری سه‌بعدی خواهید دید که از چهار مثلث، پنج خط و شش زاویه تشکیل شده است.

گرافیک بیت‌مپ شده(.BMP)

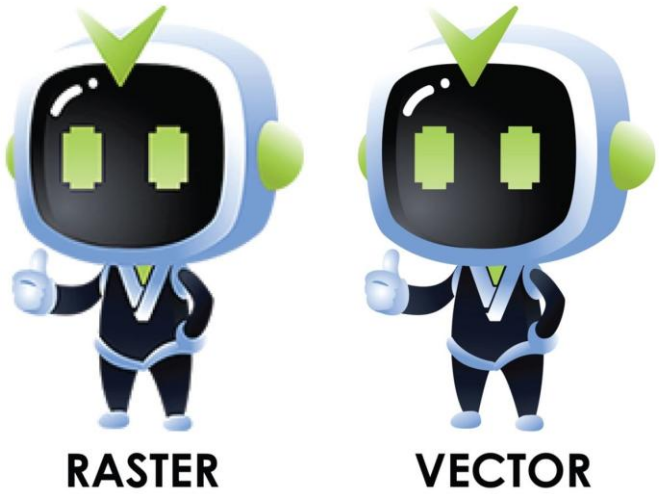
گرافیک بیت‌مپ‌شده یا همان گرافیک شطرنجی‌شده (Rasterized)، تصویری دیجیتالی است که در آن هر پیکسل با تعدادی بیت نمایش داده می‌شود؛ این گرافیک با تقسیم تصویر به چهارخانه‌های کوچک یا پیکسل ساخته می‌شود که هرکدام حاوی اطلاعاتی مانند کنترل شفافیت و رنگ هستند؛ بنابراین در گرافیک شطرنجی هر پیکسل مربوط به یک ارزش محاسبه‌شده و از پیش تعیین‌شده است که می‌تواند با دقت زیاد مشخص شود.

وضوح تصویر گرافیک شطرنجی به وضوح تصویر وابستگی دارد، بدین معنی که مقیاس تصاویر تولیدشده با این گرافیک را نمی‌توان بدون از دست دادن کیفیت ظاهری، افزایش داد.

گرافیک برداری

گرافیک برداری (فرمت‌های .ai یا .eps یا .pdf یا.svg) نیز تصویری است که مسیرهایی با نقطه‌ی شروع و پایان را ایجاد می‌کند. این مسیرها همگی براساس عبارات ریاضی بوده و از اشکال هندسی پایه‌ای مانند خطوط، چندضلعی‌ها و منحنی‌ها تشکیل شده‌اند. مزیت اصلی استفاده از گرافیک برداری به جای گرافیک بیت‌مپ شده (شطرنجی)، توانایی آن‌ها در مقیاس‌بندی بدون از دست دادن کیفیت است. مقیاس تصاویر تولیدشده با گرافیک برداری را می‌توان به‌راحتی، بدون افت کیفیت و به اندازه‌ی توانایی دستگاهی که آن‌ها را رندر می‌کند، افزایش داد.

همان‌طورکه گفته شد برخلاف گرافیک‌های برداری که به کمک فرمول‌های ریاضی به هر اندازه مقیاس می‌شوند، گرافیک بیت‌مپ‌شده با مقیاس‌بندی کیفیت خود را از دست می‌دهد. پیکسل‌های یک گرافیک بیت‌مپ‌شده هنگام افزایش ابعاد، باید درون‌یابی شوند که این امر تصویر را تار می‌کند و هنگام کاهش ابعاد نیز باید دوباره نمونه‌برداری شوند، که این کار باعث از دست دادن داده‌های تصویر می‌شود.



به‌طور کلی، گرافیک‌های برداری برای خلق آثار هنری تشکیل‌شده از اشکال هندسی، مانند لوگو یا نقشه‌های دیجیتال، حروف‌چینی یا طراحی‌های گرافیکی بهترین گزینه هستند و گرافیک‌های شطرنجی نیز بیشتر با عکس‌ها و تصاویر واقعی سروکار دارند و برای تصاویر عکاسی مناسب هستند.

از گرافیک برداری می‌توان برای ساخت بنر یا لوگو استفاده کرد؛ چراکه تصاویر با این روش هم در ابعاد کوچک هم در ابعاد بزرگ با کیفیتی یکسان نمایش داده می‌شوند. یکی از محبوب‌ترین برنامه‌هایی که برای مشاهده و ایجاد تصاویر برداری استفاده می‌شود Adobe Illustrator است.

رندرینگ

به فرایند تولید تصاویر سه‌بعدی از نرم‌افزاری بر پایه‌ی مدل‌های محاسباتی و نمایش آن به‌عنوان خروجی روی نمایشگر دو بعدی، رندرینگ (Rendering) گفته می‌شود.

APIگرافیکی

رابط برنامه‌نویسی نرم‌افزاری (Application Programming Interface) یاAPI ، پروتکلی برای ارتباط میان بخش‌های مختلف برنامه‌های کامپیوتری و ابزاری مهم برای تعامل نرم‌افزار با سخت‌افزار گرافیکی به‌حساب می‌آید؛ این پروتکل ممکن است مبتنی بر وب، سیستم‌عامل، مرکز داده، سخت‌افزار یا کتابخانه‌های نرم‌افزاری باشد. امروزه برای تصویرسازی و رندرینگ مدل‌های سه‌بعدی، ابزارها و نرم‌افزارهای فراوانی توسعه داده شده‌اند و یکی از کابردهای مهم-API های گرافیکی نیز آسان‌کردن فرایند تصویرسازی و رندرینگ برای توسعه‌دهنگان به‌شمار می‌رود. درواقع-API های گرافیکی دسترسی مجازی به برخی پلتفرم‌ها را برای توسعه‌دهندگان برنامه‌های گرافیکی خود و تست آن‌ها فراهم می‌کنند. در ادامه برخی از شناخته‌شده‌ترینAPI های گرافیکی را معرفی می‌کنیم:

OpenGL (مخفف Open Graphics Library) کتابخانه‌ای از توابع مختلف برای ترسیم تصاویر سه‌بعدی است که استاندارد بین پلتفرمی و واسط برنامه‌نویسی کاربردی (API) برای گرافیک‌ها و رندرهای دو بعدی و سه بعدی و شتاب‌دهنده‌ی گرافیکی در بازی‌های ویدئویی، طراحی، واقعیت مجازی و سایر برنامه‌ها به حساب می‌آید. این کتابخانه بیش از ۲۵۰ تابع فراخوانی مختلف برای ترسیم تصاویر سه‌بعدی دارد و در دو نوع Microsoft (اغلب در ویندوز یا نرم‌افزار نصب کارت گرافیک) و Cosmo (برای سیستم‌هایی که شتاب‌دهنده‌ی گرافیکی ندارند) طراحی شده است.

رابط گرافیکی OpenGL برای اولین بار توسط Silicon Graphics در سال ۱۹۹۱ طراحی شد و در سال ۱۹۹۲ به‌بازار آمد؛ جدیدترین سری این‌API ، یعنی OpenGL ۴.۶ نیز در جولای ۲۰۱۷ معرفی شد.

مجموعه‌ای از رابط‌های برنامه‌نویسی کاربردی (API) که توسط مایکروسافت برای فراهم‌سازی امکان ارتباط دستورالعمل‌ها با سخت‌افزارهای صوتی و تصویری توسعه داده شده است. بازی‌هایی که به DirectX مجهز هستند، این قابلیت را دارند که از ویژگی‌های چندرسانه‌ای و شتاب‌دهنده‌های گرافیکی به‌طور کارآمدتری استفاده کنند و عملکرد کلی بهبودیافته‌تری داشته باشند.

زمانی که مایکروسافت در اواخر سال ۱۹۹۴، خود را برای انتشار ویندوز ۹۵ آماده می‌کرد، الکس سنت جان، یکی از کارمند مایکروسافت، درباره‌ی توسعه‌ی بازی‌های سازگار با MS-DOS تحقیق کرد. برنامه‌نویسان این بازی‌ها اغلب امکان انتقال آن‌ها را به ویندوز ۹۵ رد کردند و توسعه‌ی بازی‌ها را برای محیط ویندوز دشوار خواندند. به همین منظور تیم سه‌نفره‌ای تشکیل شد و این تیم در عرض چهار ماه توانست اولین مجموعه از رابط‌های برنامه‌نویسی کاربردی (API) را به نام DirectX برای حل این مشکل توسعه دهد.

اولین نسخه‌ی DirectX سپتامبر ۱۹۹۵ با عنوان Windows Games SDK منتشر شد و جایگزین Win32برای DCI و-API های WinG برای ویندوز ۳.۱ بود DirectX .برای ویندوز ۹۵ و همه نسخه‌های ویندوز مایکروسافت بعد از آن، امکان داد که محتوای چند رسانه‌ای با کارایی بالا را در خود جای دهند .

مایکروسافت برای پذیرش هرچه بیشتر DirectX از سوی توسعه‌دهندگان، به جان کارمَک (John Carmack)، توسعه‌دهنده‌ی بازی‌های Doom و Doom 2 پیشنهاد داد که این دو بازی را از MS-DOS به صورت رایگان و با DirectX به ویندوز ۹۵ منتقل و شناسه تمامی حقوق انتشار بازی را نیز حفظ کند. کارمک موافقت کرد و اولین نسخه از بازی‌ها به نام Doom 95 در آگوست ۱۹۹۶ به‌عنوان اولین بازی توسعه داده شده روی DirectX منتشر شد. DirectX 2.0 با انتشار نسخه‌ی بعدی Windows 95 و Windows NT 4.0 در اواسط سال ۱۹۹۶ به یکی از اجزای خود ویندوز تبدیل شد.

ازآنجاکه در آن زمان ویندوز ۹۵ هنوز در ابتدای راه خود بود و بازی‌های منتشرشده‌ی کمی برای آن وجود داشت، مایکروسافت برای این رابط برنامه‌نویسی دست به تبلیغات گسترده زد و در طی رویدادی برای اولین بار Direct3D و DirectPlay را در دموی آنلاین بازی چندنفره‌ی MechWarrior 2 معرفی کرد. تیم توسعه‌دهنده‌ی DirectX با چالش آزمایش هر نسخه از این رابط برنامه‌نویسی برای هر مجموعه سخت‌افزار و نرم‌افزار کامپیوتر روبه‌رو شد و در همین راستا نیز انواع کارت‌های گرافیک مختلف، کارت‌های صوتی، مادربردها، پردازنده‌ها، ورودی‌ها، بازی‌ها و سایر برنامه‌های چندرسانه‌ای با هر نسخه‌ی بتا و نهایی آزمایش شدند و حتی آزمایش‌هایی تولید و توزیع شد تا صنعت سخت‌افزار، سازگاری طراحی‌های جدید و نسخه‌های درایور خود با DirectX را بررسی کند.

جدیدترین نسخه‌ی DirectX، یعنی DirectX 12 در سال ۲۰۱۴ رونمایی شد و یک سال بعد از آن نیز همراه با نسخه‌ی ۱۰ ویندوز به‌طور رسمی به بازار آمد. این API گرافیکی از آداپتور

چندگانه‌ی خاصی پشتیبانی کرده و امکان استفاده‌ی هم‌زمان از چند گرافیک را روی یک سیستم فراهم می‌کند.

قبل از DirectX، مایکروسافت OpenGL را در پلتفرم ویندوز NT خود گنجانده بود و حالا Direct3D قرار بود جایگزینی برای OpenGL تحت کنترل مایکروسافت باشد که در ابتدا روی گیمینگ متمرکز بود. در این مدت OpenGL هم توسعه داده شده بود و تکنیک‌های برنامه‌نویسی برای برنامه‌های چندرسانه‌ای تعاملی مانند بازی‌ها را بهتر پشتیبانی می‌کرد، اما ازآنجاکه OpenGL در مایکروسافت توسط تیم DirectX پشتیبانی می‌شد، کم‌کم از میدان رقابت کناره گرفت.

Vulkan

Vulkan یک API گرافیکی کم‌هزینه و چندپلتفرمی است که برای کاربردهای گرافیکی مانند گیمینگ و تولید محتوا به کار می‌رود. وجه تمایز این API گرافیکی با DirectX و OpenGL، توانایی آن در رندرینگ گرافیک‌های دوبعدی و مصرف برق کم‌تر است.

در ابتدا بسیاری تصور می‌کردند که Vulkan می‌تواند OpenGL بهبودیافته‌ی آینده و ادامه‌دهنده‌ی مسیر آن باشد، اما گذشت زمان نشان داد که این پیش‌بینی درست نبود. جدول زیر تفاوت‌های عملکرد این دو API گرافیکی را نشان می‌دهد.

Vulkan	OpenGL
مبتنی بر شیء است و فاقد global state	تنها یک ماشین global state دارد
مفهوم تمامی stateها در بافر دستورها قرار گرفته است	state تنها به یک محتوا منحصر می‌شود
قابلیت برنامه‌نویسی چندرشته‌ای دارد	عملکردها فقط به‌صورت ترتیبی انجام می‌شوند

Vulkan	OpenGL
کنترل و مدیریت همگام‌سازی و حافظه مقدور است	حافظه و همگام‌سازی GPU معمولاً مخفی است
درایورها حین اجرا، بررسی خطا انجام نمی‌دهند. در عوض برای سازندگان، یک لایه اعتبارسنجی در نظر گرفته شده است.	**بررسی خطا به‌صورت مداوم انجام می‌شود**

Mantle

API گرافیکی Mantle، رابطی ارزان‌قیمت برای رندر بازی‌های ویدیویی سه‌بعدی است که اولین بار توسط AMD و شرکت تولیدکننده‌ی بازی‌های ویدیویی DICE در سال ۲۰۱۳ طراحی شد. هدف از این مشارکت رقابت با Direct3D و OpenGL در کامپیوترهای خانگی بود، بااین‌حال Mantle در سال ۲۰۱۹ رسما متوقف شد و API گرافیکی Vulkan جای آن را گرفت. Mantle می‌توانست به‌صورت بهینه بار کاری پردازنده را کاهش داده و گره‌های ایجاد شده در فرایند پردازش را از بین ببرد.

Metal

Metal رابط گرافیکی اختصاصی اپل است که مبتنی بر زبان C++ نوشته شده و اولین‌بار در iOS 8 به‌کار گرفته شد. Metal را می‌توان ترکیب رابط گرافیکی OpenGL و فریم‌ورک OpenCL دانست که هدف از طراحی آن شبیه‌سازی APIهای گرافیکی دیگر پلتفرم‌ها مانند Vulkan و DirectX 12 برای سیستم‌عامل iOS، Mac و tvOS بود. در سال ۲۰۱۷ دومین نسخه‌ی API گرافیکی Metal با پشتیبانی از سیستم‌های عامل macOS High Sierra، iOS 11 و tvOS 11 منتشر شد. این نسخه در مقایسه با نسخه‌ی قبلی کارایی بالاتر و بهینه‌تری داشت.

GDDR چیست

به حافظه‌ی DDR که در واحد پردازش گرافیکی قرار دارد، GDDR یا **رم پردازنده‌ی گرافیکی** گفته می‌شود. DDR (مخفف Double Data Rate) یا نرخ انتقال دوگانه، نسخه‌ی پیشرفته‌ی رم داینامیک هم‌زمان (SDRAM) است و از فرکانس‌های مشابه با آن استفاده می‌کند. تفاوت DDR با SDRAM در تعداد دفعات ارسال داده در هر چرخه است؛ DDR داده‌ها را دو بار در هر چرخه انتقال می‌دهد و سرعت حافظه را دو برابر می‌کند، درحالی‌که SDRAM سیگنال‌ها را تنها یک بار در هر چرخه ارسال می‌کند. DDRها خیلی سریع محبوبیت پیدا کردند، چراکه علاوه بر سرعت انتقال دوبرابری، از SDRAM ارزان‌تر بوده و همچنین انرژی کمتری نسبت به ماژول‌های SDRAM قدیمی مصرف می‌کنند.

تاریخچه گرافیک سه بعدی

اصطلاح GPU برای اولین بار در دهه‌ی ۱۹۷۰، به‌عنوان مخففی برای واحد پردازش گرافیکی (Graphic Processor Unit)معرفی شد و یک واحد پردازش قابل‌برنامه‌ریزی را توصیف می‌کرد که عملکردی مستقل از واحد پردازنده‌ی مرکزی یا همان پردازنده داشت و مسئولیت تنظیم و خروجی گرافیکی را عهده‌دار بود؛ البته در آن زمان این اصطلاح آن‌گونه که امروزه تعبیر می‌شود، تعریف نشده بود.

IBM در سال ۱۹۸۱ برای اولین بار دو کارت گرافیک خود از نوع MDA (آداپتور نمایشگر تک‌رنگ) و CGA (آداپتور گرافیک رنگی) را توسعه داد. MDA از چهار کیلوبایت حافظه‌ی ویدئویی بهره‌مند بود و تنها از نمایش متنی پشتیبانی می‌کرد؛ این گرافیک امروزه دیگر کاربردی ندارد، اما ممکن است در برخی از سیستم‌های قدیمی یافت شود.

CGAنیز اولین گرافیک برای کامپیوترها محسوب می‌شد که تنها به شانزده کیلوبایت حافظه ویدئویی مجهز بود و قابلیت تولید ۱۶ رنگ با وضوح ۱۶۰ در ۲۰۰ پیکسل را داشت. یک سال پس از این اتفاق شرکت فناوری کامپیوتری هرکول (Hercules Graphics) برای پاسخ به کارت‌های گرافیکیIBM ، گرافیک HGC (کارت گرافیکی هرکول) را با ۶۴ کیلوبایت حافظه‌ی ویدئویی توسعه داد که ترکیبی از MDA با گرافیک بیت‌مپ شده بود.

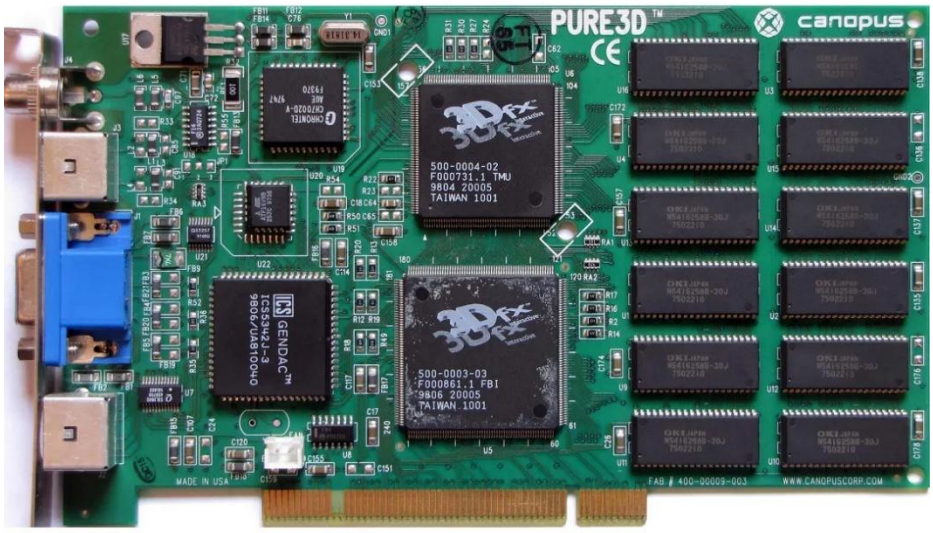
در سال ۱۹۸۳ اینتل با معرفی گرافیک iSBX 275 Video Graphics Multimodule وارد بازار کارت گرافیک شد. این کارت می‌توانست هشت رنگ را با وضوح ۲۵۶ در ۲۵۶ نمایش دهد . IBMایک سال پس از این اتفاق گرافیک‌های PGC یا کنترل‌کننده‌ی گرافیک حرفه‌ای (Professional Graphic Controller)و EGA یا آداپتور گرافیکی پیشرفته (Enhanced Graphic Adapter)را معرفی کرد که ۱۶ رنگ را با وضوح ۶۴۰ در ۳۵۰ پیکسل نمایش می‌دادند.

استاندارد VGA یا آرایه‌ی گرافیکی ویدئویی (Video Graphics Array) در سال ۱۹۸۷ معرفی شد، این استاندارد وضوح تصویر ۶۴۰ در ۴۸۰ را با ۱۶ رنگ و حداکثر ۲۵۶ کیلوبایت حافظه‌ی ویدئویی ارائه می‌داد. ATI همان سال اولین کارت گرافیک VGA خود را با نام ATI VGA Wonder معرفی کرد؛ برخی از مدل‌های این کارت گرافیک حتی به پورتی برای اتصال ماوس مجهز بودند. تا اینجا کارت‌های ویدئویی حافظه‌های کمی داشتند و پردازنده‌ها، پردازش‌های گرافیکی را به این حافظه‌های ویدئویی انتقال می‌دادند و پس از انجام محاسبات و تبدیل سیگنال، آن‌ها را روی دستگاه خروجی نمایش می‌دادند.

پردازنده‌ی گرافیکی مکملی برای پردازنده و سبک‌تر کردن بار کاری

۴ سال بعد Voodoo اولین کارت گرافیک خود را توسط شرکتی به نام ۳dfx معرفی کرد. این گرافیک Voodoo1 نام داشت و برای رندر گرافیک سه‌بعدی به نصب کارت گرافیک دوبعدی نیاز داشت و خیلی سریع در میان گیمرها محبوب شد.

انویدیا در سال ۱۹۹۷ در پاسخ به Voodoo شتاب‌دهنده‌ی گرافیکی RIVA 128 را عرضه کرد. RIVA 128 مانند Voodoo1 امکان استفاده از شتاب‌دهنده‌های گرافیکی را همراه با گرافیک‌های دو بعدی برای سازندگان کارت‌های ویدئویی فراهم می‌کرد، اما در مقایسه با Voodoo1 رندر گرافیکی ضعیف‌تری داشت.



پس از RIVA 128 شرکت ۳dfx گرافیک Voodoo2 را به‌عنوان جایگزینی برای Voodoo1 عرضه کرد. این اولین کارت گرافیکی بود که از SLI پشتیبانی می‌کرد و امکان اتصال دو یا چند گرافیک را برای تولید یک خروجی فراهم می‌کرد SLI .یا رابط پیوند مقیاس‌پذیر (Scalable Link Interfaceنام تجاری فناوری منسوخ‌شده‌ای است که توسط انویدیا برای پردازش موازی و به منظور افزایش قدرت پردازش گرافیکی توسعه داده شد.

اصطلاح GPU در سال ۱۹۹۹ هم‌زمان با عرضه‌ی جهانی GeForce 256 به‌عنوان اولین پردازنده‌ی گرافیکی جهان توسط انویدیا رواج داده شد. انویدیا این GPU را به‌عنوان پردازنده‌ای تک تراشه با تبدیل یکپارچه‌ی نمای دو بعدی از صحنه‌ای سه‌بعدی، نورپردازی و تغییر رنگ سطوح و قابلیت ترسیم بخش‌هایی از تصویر پس از رندر معرفی کرد ATI Technologies .نیز برای رقابت با انویدیا در سال ۲۰۰۲ گرافیک Radeon 9700 را با اصطلاح واحد پردازش بصری یا VPU منتشر کرد.

با گذشت زمان و پیشرفت تکنولوژی، GPUها به قابلیت برنامه‌ریزی مجهز شدند و همین امر باعث شد انویدیا و ATI نیز به صحنه‌ی رقابت وارد شوند و اولین پردازنده‌های گرافیکی خود را (GeForce برای انویدیا و Radeon برای ATI) معرفی کنند.

انویدیا در سال ۱۹۹۹ با عرضه‌ی گرافیک GeForce 256 رسماً به بازار کارت گرافیک وارد شد. این گرافیک اولین پردازنده‌ی گرافیکی واقعی در جهان شناخته می‌شود که ۳۲ مگابایت حافظه‌ی DDR (همان GDDR) داشت و به‌طور کامل از DirectX 7 پشتیبانی می‌کرد.

هم‌زمان با تلاش‌ها برای سرعت بخشیدن به انجام محاسبات و پردازش‌های گرافیکی کامپیوترها و بهبود کیفیت آن‌ها، شرکت‌های تولیدکننده‌ی بازی‌های ویدئویی و کنسول‌های گیمینگ نیز هرکدام به نحوی (سِگا با Dreaنحوه تولید گرافیک سه‌بعدی

فرایند تولید گرافیک سه بعدی به سه مرحله‌ی اصلی تقسیم می‌شود:

مدل‌سازی سه بعدی

فرایند توسعه‌ی آرایه‌ای مبتنی بر مختصات ریاضی از رویه یا سطح جسمی (بی‌جان یا جان‌دار) به صورت سه‌بعدی است که ازطریق نرم‌افزارهای تخصصی با دستکاری اضلاع، رئوس و چندضلعی‌هایی که در فضای سه‌بعدی شبیه‌سازی شده‌اند، انجام می‌گیرد.

اجسام فیزیکی با استفاده از مجموعه‌ای نقاط در فضای سه‌بعدی نشان داده می‌شوند که توسط عناصر هندسی مختلف مانند مثلث‌ها، خطوط، سطوح منحنی و غیره به هم متصل می‌شوند. اساساً مدل‌های سه‌بعدی در ابتدا با اتصال نقاط و تشکیل چندضلعی ایجاد می‌شوند. چندضلعی، ناحیه‌ای است که از حداقل سه رأس (مثلث) تشکیل شده باشد و یکپارچگی کلی مدل و مناسب بودن آن برای استفاده در انیمیشن به ساختار این چند ضلعی‌ها بستگی دارد.

مدل‌های سه‌بعدی (D۳) از دو روش مدل‌سازی چندضلعی (Vertex) و با اتصال خطوط شبکه‌ای از بردارها یا مدل‌سازی منحنی (Pixel) با وزن‌دهی به هر نقطه ساخته می‌شوند؛ امروزه به دلیل انعطاف‌پذیری بیشتر و امکان رندر سریع‌تر فرایند مدل‌سازی سه‌بعدی در روش اول، اکثریت

قریب به اتفاق مدل‌های سه‌بعدی، به روش چندضلعی و بافت‌دار تولید می‌شوند. یکی از اصلی‌ترین وظایف کارت‌های گرافیکی نگاشت بافت (Texture Mapping) است که به یک تصویر یا مدل سه‌بعدی، بافت اضافه می‌کند. برای مثال، با اضافه کردن بافت سنگی به یک مدل آن را به تصویر سنگی واقعی شبیه می‌کند یا با اضافه کردن بافتی شبیه به صورت انسان، برای مدل سه‌بعدی اسکن‌شده‌ای، چهره طراحی می‌کند. mcast، سونی با PS1 و نینتندو با Nintendo 64) سعی کردند تا در این حوزه به رقابت بپردازند.

در روش دوم نیز مدل‌سازی با کنترل وزنی نقاط منحنی به دست می‌آیند، البته نقاط درون‌یابی نمی‌شوند، بلکه تنها می‌توان سطوح منحنی را با استفاده از ازدیاد نسبی چندضلعی‌ها ایجاد کرد. در این روش افزایش وزن برای یک نقطه، منحنی را به آن نقطه نزدیک می‌کند.

چیدمان و انیمیشن

پس از مدل‌سازی باید نحوه‌ی قرار دادن و تعیین حرکت اشیا (مدل‌ها، نورها و غیره) در یک صحنه پیش از پرداخت اجسام و ایجاد تصویر مشخص شود؛ بدین معنی که قبل از رندر شدن تصاویر، اجسام باید طرح‌بندی و درون صحنه چیده شوند. در واقع با تعریف مکان و اندازه‌ی هر جسم، رابطه‌ی فضایی بین اجسام شکل می‌گیرد. حرکت یا انیمیشن نیز به توصیف زمانی یک جسم اشاره دارد (نحوه‌ی حرکت و تغییر شکل در طول زمان). روش‌های رایج طرح‌بندی و

انیمیشن عبارت‌اند از: فریم‌بندی (قاب‌بندی) کلیدی، حرکت‌شناسی معکوس و ضبط حرکت. البته این تکنیک‌ها اغلب به صورت ترکیبی استفاده می‌شوند.

رندرینگ

در مرحله‌ی آخر براساس نحوه‌ی قرارگیری نور، انواع سطوح و سایر عوامل مشخص‌شده، محاسبات کامپیوتری برای تولید و پرداخت تصویر انجام می‌شود. در این قسمت متریال‌ها و بافت‌ها داده‌هایی هستند که برای رندر کردن استفاده می‌شوند.

میزان انتقال نور از سطحی به سطح دیگر و میزان پخش و تعامل آن روی سطوح، دو عمل اساسی در رندرینگ هستند که اغلب با استفاده از نرم‌افزارهای گرافیکی سه‌بعدی اجرا می‌شوند. در واقع رندرینگ فرایند نهایی ایجاد تصویر یا انیمیشن دوبعدی از مدلی سه‌بعدی و صحنه‌ای آماده به کمک چندین روش مختلف و اغلب تخصصی است که شاید تنها کسری از ثانیه یا گاهی تا چند روز برای یک تصویر/فریم منفرد طول بکشد.

تکنیک سایه‌زنی

پس از توسعه‌ی پردازنده‌های گرافیکی برای کم کردن حجم کاری پردازنده‌ها و فراهم کردن بستری برای تولید تصاویر با کیفیتی بسیار چشمگیرتر از قبل، انویدیا و ATI کم‌کم به بازیگرهای اصلی دنیای گرافیک‌های کامپیوتری تبدیل شدند. این دو رقیب برای پیشی گرفتن از یکدیگر سخت تلاش می‌کردند و هر کدام سعی داشتند تا با افزایش تعداد سطوح در مدل‌سازی و رندرینگ و بهبود تکنیک‌ها با هم رقابت کنند. تکنیک سایه‌زنی را می‌توان زاده‌ی رقابت آن‌ها دانست.

در صنعت گرافیک کامپیوتری، سایه‌زنی به فرایند تغییر رنگ جسم/سطح/چندضلعی در صحنه‌ای سه‌بعدی، براساس مواردی مانند فاصله‌ی آن از نور، زاویه‌ی آن نسبت به نور یا زاویه سطح نسبت به نور اشاره دارد.

تفاوت GPU وCPU

پردازنده‌ی گرافیکی از ابتدا به‌عنوان مکملی برای پردازنده و سبک‌تر کردن بار کاری این واحد، تکامل پیدا کرد. امروزه عملکرد پردازنده‌ها با دستاوردهای جدید در معماری ساخت آن‌ها، افزایش فرکانس و تعداد هسته‌ها، روزبه‌روز قدرتمندتر می‌شود، درمقابل پردازنده‌های گرافیکی به‌طور خاص برای سرعت بخشیدن به پردازش‌های گرافیکی توسعه داده شده‌اند.

پردازنده‌ها به صورتی برنامه‌ریزی شده‌اند که بتوانند علاوه بر اینکه یک کار را با کمترین تأخیر و بالاترین سرعت انجام می‌دهند، خیلی سریع هم بین عملیات جابه‌جا شوند. در واقع نحوه‌ی پردازش درCPU ها، سریالی است.

درمقابل، پردازنده‌ی گرافیکی به‌طور خاص برای بهینه‌سازی توان عملیاتی پردازش‌های گرافیکی توسعه داده شده است و امکان انجام کارها به‌طور هم‌زمان و موازی را فراهم می‌کند. در تصویر زیر تعداد هسته‌های یک پردازنده و تعداد هسته‌های یک پردازنده‌ی گرافیکی را مشاهده می‌کنید؛ این تصویر نشان می‌دهد که تفاوت اصلی بین CPU و GPU در تعداد هسته‌های آن‌ها برای پردازش یک وظیفه است.

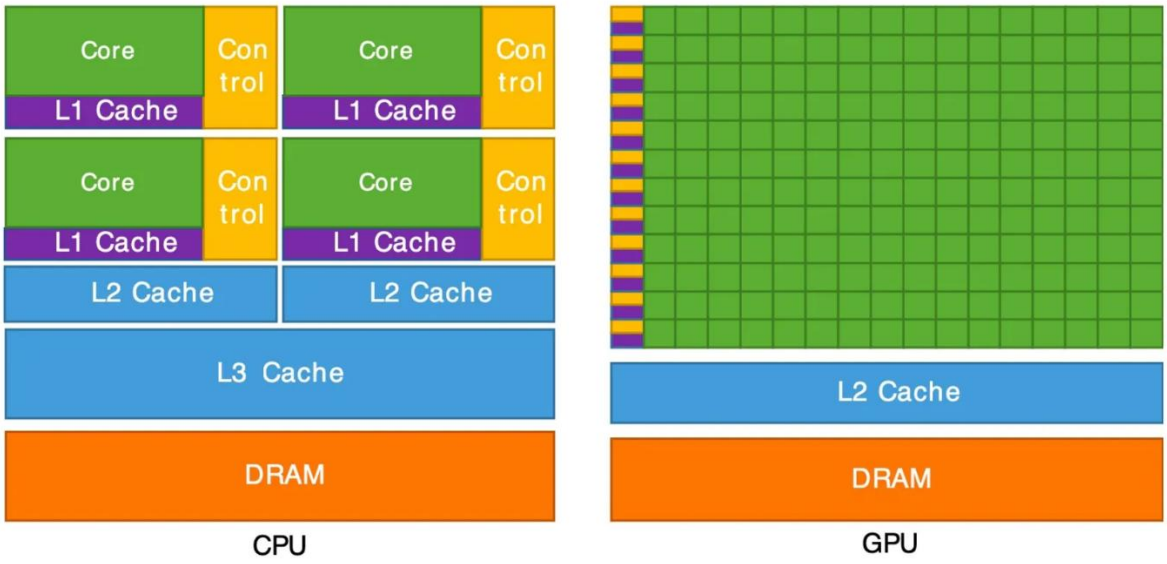
تفاوت GPU وCPU

پردازنده‌ی گرافیکی از ابتدا به‌عنوان مکملی برای پردازنده و سبک‌تر کردن بار کاری این واحد، تکامل پیدا کرد. امروزه عملکرد پردازنده‌ها با دستاوردهای جدید در معماری ساخت آن‌ها، افزایش فرکانس و تعداد هسته‌ها، روزبه‌روز قدرتمندتر می‌شود، درمقابل پردازنده‌های گرافیکی به‌طور خاص برای سرعت بخشیدن به پردازش‌های گرافیکی توسعه داده شده‌اند.

پردازنده‌ها به صورتی برنامه‌ریزی شده‌اند که بتوانند علاوه بر اینکه یک کار را با کمترین تأخیر و بالاترین سرعت انجام می‌دهند، خیلی سریع هم بین عملیات جابه‌جا شوند. در واقع نحوه‌ی پردازش درCPU ها، سریالی است.

درمقابل، پردازنده‌ی گرافیکی به‌طور خاص برای بهینه‌سازی توان عملیاتی پردازش‌های گرافیکی توسعه داده شده است و امکان انجام کارها به‌طور هم‌زمان و موازی را فراهم می‌کند. در تصویر زیر تعداد هسته‌های یک پردازنده و تعداد هسته‌های یک پردازنده‌ی گرافیکی را مشاهده می‌کنید؛ این تصویر نشان می‌دهد که تفاوت اصلی بین CPU و GPU در تعداد هسته‌های آن‌ها برای پردازش یک وظیفه است.

۱



از مقایسه‌ی معماری کلی پردازنده‌ها و پردزانده‌های گرافیکی می‌توان شباهت‌های زیادی بین این دو واحد نیز پیدا کرد. هر دو از ساختارهای مشابهی در لایه‌های کش بهره می‌برند و هردو از کنترلری برای حافظه و یک رم اصلی استفاده می‌کنند. نمای کلی از معماری پردازنده‌های مدرن حاکی از آن است که در این واحد با تمرکز بر حافظه و لایه‌های کش، دسترسی به حافظه با تأخیر

کم مهم‌ترین عامل در طراحی پردازنده‌ها است (چیدمان دقیق به فروشنده و مدل پردازنده بستگی دارد).

هر پردازنده از چندین لایه کش تشکیل شده است:

حافظه‌ی کش سطح یک (L1) سریع‌ترین، کم‌ظرفیت‌ترین و نزدیک‌ترین حافظه به پردازنده است و مهم‌ترین داده‌های مورد نیاز برای پردازش را در خود ذخیره می‌کند.

لایه‌ی بعدی حافظه‌ی کش سطح دو (L2) یا حافظه‌ی کش خارجی است که نسبت به L1 سرعت کمتر و حجم بیشتری دارد.

حافظه‌ی کش L3 در پردازنده بین تمام هسته‌ها مشترک است و از لحاظ ظرفیت نسبت به حافظه‌ی کش L1 و L2 حجم بیشتری و سرعت پایین‌تری دارد؛ حافظه‌ی کش L4 هم مانندL3 ، نسبت به L1 و L2 حجم بیشتر و سرعت کمتری دارد؛ این دو معمولاً به‌صورت اشتراکی استفاده می‌شوند. اگر داده‌ها در لایه‌های کش قرار نگرفته باشند از رم اصلی (DDR) فراخوانی می‌شوند.

با نگاه به نمای کلی معماری پردازنده‌ی گرافیکی (چیدمان دقیق به تولیدکننده و مدل بستگی دارد) متوجه می‌شویم که ماهیت این واحد به جای دسترسی سریع به حافظه کش یا کاهش تأخیر، روی به کار انداختن هسته‌های موجود تمرکز دارد. در واقع پردازنده‌ی گرافیکی از چندین گروه هسته تشکیل شده است که در حافظه‌ی کش سطح یک قرار دارند.

پردازنده‌ی گرافیکی در مقایسه با پردازنده، لایه‌های حافظه‌ی کش کمتر و کم‌ظرفیت‌تری دارد، این واحد به ترانزیستورهای بیشتر مختصِ محاسبات مجهز است و کمتر به بازیابی داده‌ها از حافظه اهمیت می‌دهد؛ پردازنده‌ی گرافیکی با رویکرد انجام محاسبات موازی توسعه داده شده است.

محاسبات با کارایی بالا (High Performance Computing) یکی از موارد استفاده‌ی مؤثر و قابل‌اعتماد پردازش‌های موازی برای اجرای برنامه‌های کاربردی پیشرفته است؛ دقیقاً به همین دلیل پردازنده‌های گرافیکی برای این دست از محاسبات مناسب هستند.

به زبان ساده، فرض کنید برای انجام نوعی از محاسبات سنگین، دو راه پیش رو داشته باشید:

استفاده از تعداد کمی هسته‌ی قدرتمند که پردازش‌ها را به صورت سریالی انجام می‌دهند.

استفاده از تعداد بالای هسته‌هایی نه‌چندان قدرتمند که می‌توانند چندین پردازش را به صورت هم‌زمان انجام دهند.

در سناریوی اول اگر یکی از هسته‌ها را از دست بدهیم با مشکلی جدی روبه‌رو خواهیم شد؛ عملکرد دو هسته‌ی دیگر تحت تأثیر قرار خواهد گرفت و قدرت پردازشی به‌شدت کاهش خواهد یافت، درمقابل اگر در سناریوی دوم هسته‌ای را از دست بدهیم، تغییر محسوسی در روند پردازش به وجود نمی‌آید و باقی هسته‌ها به کار خود ادامه می‌دهند.

آشنایی با معماری پردازنده گرافیکی

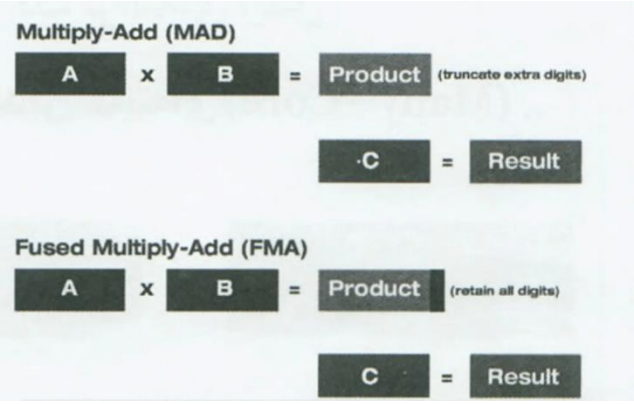
در نگاه اول، پردازنده‌ی مرکزی نسبت به پردازنده‌ی گرافیکی، واحدهای محاسباتی بزرگ‌تر اما کمتری دارد. البته در نظر داشته باشید که یک هسته در پردازنده، سریع‌تر و هوشمندتر از یک هسته در پردازنده‌ی گرافیکی عمل می‌کند.

در طول زمان، فرکانس هسته‌های پردازنده برای بهبود عملکرد به تدریج افزایش پیدا کرد و برخلاف آن، فرکانس هسته‌های پردازنده‌ی گرافیکی برای بهینه کردن مصرف و تطبیق نصب در گوشی‌ها یا دستگاه‌های دیگر، کاهش داده شد.

قابلیت انجام غیرمنظم پردازش‌ها را می‌توان گواهی بر هوشمند بودن هسته‌های پردازنده دانست. همان‌طورکه گفته شد، واحد پردازش مرکزی می‌تواند دستورالعمل‌ها را با ترتیبی متفاوت از آنچه برایش تعریف شده، اجرا کند یا دستورالعمل‌های مورد نیاز در آینده‌ای نزدیک را پیش‌بینی کرده و عملوندها را برای بهینه‌سازی هرچه بیشتر سیستم و صرفه‌جویی در زمان، قبل از اجرا، آماده کند.

درمقابل، هسته‌ی یک پردازنده‌ی گرافیکی مسئولیت پیچیده‌ای بر عهده ندارد و برای پردازشی خارج از دستور و برنامه، آنچنان کار زیادی انجام نمی‌دهد. به‌طور کلی، تخصص اصلی هسته‌های پردازنده‌های گرافیکی انجام عملیات ممیز شناور مانند ضرب دو عدد و جمع کردن عددی سوم (A x B + C = Result)با گرد کردن نتیجه به عددی صحیح بود که به اختصار آن را multiply-add یا MAD می‌نامند یا همان نتیجه را با دقت کامل (بدون کوتاه‌سازی) در مرحله‌ی ضرب استفاده می‌کند که به آن Fused Multiplay-Add یا FMA می‌گویند.

جدیدترین ریزمعماری‌های پردازنده‌های گرافیکی دیگر امروزه به FMA هم محدود نیستند و عملیات پیچیده‌تری مانند رهگیری پرتو یا پردازش‌های هسته‌های تنسور را انجام می‌دهند. هسته‌های تنسور (Tensore Cores) و هسته‌های رهگیری پرتو نیز برای ارائه‌ی رندرهایی بیش از حد واقع‌گرایانه طراحی شده‌اند.



هسته‌های تنسور

انویدیا در سال ۲۰۲۰، پردازنده‌های گرافیکی مجهز به هسته‌های افزوده‌ای را تولید کرد که علاوه بر قابلیت سایه‌زنی(Shader) ، برای پردازش‌های هوش مصنوعی، یادگیری عمیق و شبکه‌ی عصبی نیز کاربرد داشتند. این هسته‌ها، تنسور (Tensor) نام دارند. تنسور مفهومی ریاضیاتی است که کوچک‌ترین واحد تصورپذیر آن، صفر بُعد (ساختار صفر در صفر) دارد و تنها یک مقدار را شامل می‌شود. با افزایش تعداد بُعدها، ساختارهای دیگر تنسور عبارت‌اند از:

تنسور یک‌بُعدی: برداری (Vectorبا ساختار صفر در یک)

تنسور دوبُعدی: ماتریسی (Matrixبا ساختار یک در یک)

هسته‌های تنسور در دسته‌ی SIMD یا «دستورالعمل تکی برای چندین داده» قرار می‌گیرند و استفاده از آن‌ها در پردازنده‌های گرافیکی با فراهم کردن تمامی نیازهای محاسباتی و پردازشی موازی، تراشه‌ای بسیار باهوش‌تر از ماشین‌حساب برای جلوه‌های گرافیکی به وجود آورد. انویدیا در سال ۲۰۱۷ گرافیکی را با معماری کاملاً جدید به‌نام ولتا (Volta) معرفی کرد که با هدف‌گیری بازارهای حرفه‌ای طراحی و ساخته شده بود؛ این گرافیک به هسته‌هایی مخصوص محاسبه‌های تنسور مجهز بود، اما پردازنده‌های گرافیکی GeForce از آن بی‌بهره بودند.

در سال ۲۰۲۰ معماری امپر در پردازنده‌های گرافیکی A100 برای دیتاسنترها معرفی شد؛ در این معماری بهره‌وری و قدرت هسته‌ها افزایش پیدا کرده، تعداد عملیات در هر چرخه چهار برابر شد و فرمت‌های داده‌ای جدیدی هم به مجموعه‌ی پشتیبانی‌شده، اضافه شده بود. امروزه هسته‌های تنسور قطعات سخت‌افزاری خاص و محدودی هستند که در تعداد کمی از گرافیک‌های مختص مصرف‌کننده استفاده می‌شوند. اینتل و AMD (دو بازیگر دیگر در دنیای گرافیک‌های کامپیوتری) در پردازنده‌های گرافیکی خود هسته‌های تنسور ندارند؛ اما شاید در آینده فناوری مشابهی عرضه کنند.

هسته‌های تنسور در فیزیک و مهندسی و در ریاضیات کاربرد فراوانی دارند: می‌توانند محاسبات پیچیده الکترومغناطیس و نجوم و مکانیک سیالات را انجام دهند.

هسته‌های تنسور می‌توانند وضوح تصاویر را افزایش دهند: این هسته‌ها تصاویر را در سطح گرافیک پایین‌تری (یا رزولوشن پایین‌تر) استخراج کرده و بعد از اتمام رندرگیری کیفیت تصاویر را بالا می‌برند.

هسته‌های تنسور نرخ فریم را بالا می‌برند: هسته‌های تنسور می‌توانند بعد از فعال کردن قابلیت رهگیری پرتو در بازی‌ها، نرخ‌فریم را در بازی افزایش دهند.

موتور رهگیری پرتو

پردازنده‌های گرافیکی علاوه بر هسته‌ها و لایه‌های حافظه‌ی کش ممکن است شامل سخت‌افزاری برای تسریع رهگیری پرتو (Ray Tracing) نیز باشند، که تابیدن منبع نور روی اجسام را شبیه‌سازی کرده و منطقه‌بندی‌های مختلفی را از لحاظ تابش نور ایجاد می‌کند. رهگیری پرتوهای سریع در بازی‌های ویدئویی می‌تواند تصاویر واقعی‌تر و باکیفیت‌تری را به نمایش بگذارد.

قابلیت رهگیری پرتو یکی از بزرگ‌ترین پیشرفت‌های سال‌های اخیر در گرافیک کامپیوترها و صنعت گیمینگ است. این قابلیت در ابتدا تنها در صنعت فیلم‌سازی، تولید تصاویر کامپیوتری و در انیمیشن و افکت‌های بصری به‌کار گرفته می‌شد، اما امروزه کنسول‌های گیمینگ PS5 و XBOXسری X نیز از قابلیت رهگیری پرتو پشتیبانی می‌کنند.

در دنیای واقعی هر آنچه می‌بینیم نتیجه برخورد نور به اجسام و بازتاب آن به چشم ما است؛ رهگیری پرتو همین کار را به صورت برعکس و با شناسایی منابع نور، مسیر پرتوهای نور، متریال، نوع سایه و میزان انعکاس هنگام برخورد با اجسام انجام می‌دهد. الگوریتم رهگیری پرتو بازتاب نور از اجسام با جنس‌های متفاوت را به شکل‌های متفاوت و واقعی‌تری نمایش می‌دهد، سایه‌ی اجسامی را که در مسیر پرتو نوری قرار دارند بسته به شفاف یا نیمه‌شفاف

بودن آن‌ها ترسیم می‌کند و از قوانین فیزیک پیروی می‌کند. به همین دلیل تصاویر تولیدشده با این قابلیت تا حد زیادی به واقعیت نزدیک هستند.

۱۱

انویدیا برای اولین بار قابلیت رهگیری پرتو را در سال ۲۰۱۸ و در گرافیک‌های سری RTX تحت معماری Turing منتشر کرد و پس از آن هم درایور جدیدی معرفی کرد که پشتیبانی از رهگیری پرتو را برای برخی از گرافیک‌های سری GTX فراهم می‌کرد که عملکردی ضعیف‌تر نسبت به سری RTX دارند.

AMDنیز با معرفی معماری RDNA 2 رهگیری پرتو را به کنسول‌های PS5 و Xbox سری XS وارد کرد. فعال شدن این قابلیت در بازی‌ها به دلیل بار پردازشی سنگین، نرخ‌فریم را کاهش می‌دهد؛ برای مثال اگر یک بازی در حالت عادی با نرخ ۶۰ فریم‌برثانیه روی سیستمی اجرا شود ممکن است با قابلیت رهگیری پرتو تنها ۳۰ فریم‌برثانیه ارائه دهد.

نرخ فریم که بر حسب فریم‌برثانیه (FPS) اندازه‌گیری می‌شود، معیاری مناسب برای نشان دادن عملکرد پردازنده‌ی گرافیکی به حساب می‌آید که نشان‌دهنده‌ی تعداد تصاویر تکمیل‌شده‌ای است که در هر ثانیه می‌توان نمایش داد؛ برای مقایسه، چشم انسان می‌تواند حدود ۲۵

فریم‌برثانیه را پردازش کند، بااین‌حال بازی‌های اکشن سریع باید حداقل ۶۰ فریم‌برثانیه پردازش کنند تا یک جریان بازی به شکل روان نمایش داده شود.

با تشکر از توجه شما

محمدرضا پورمحمد روح افزا