

تحلیل مقاله

نام و نام خانوادگی: **سحر بهمن پور - سحر فائزی - سپیده این یوسف** درس: **رایانش ابری**
مقطع تحصیلی: **دکتری / ارشد**

GPU (Graphic Processing Unit)

GPU به معنای واحد پردازش گرافیکی می باشد. کار آن تضمین نمایش محتویات و ارائه آن ها بر روی صفحه نمایش است. به طور کلی کارهایی که به نحوی به گرافیک ربط دارند، مانند بازی ها، برنامه ها، صفحات وب و... که توسط GPU به شما نمایش داده می شود. کار اصلی GPU رندر کردن و نمایش است و به طور کلی کارهایی است که به نحوی از گرافیک بالایی بر خوردار هستند. در درون پردازنده ها یا CPU نیز یک GPU ضعیفی برای انجام رندهای گرافیکی محدودی مانند نمایش دسکتاپ و... وجود دارد. اما برای انجام کارهای سنگین تر مانند گرفتن رندر های سنگین از برنامه های طراحی و یا انجام بازی هایی که گرافیک بالایی دارند از GPU جداگانه و قوی استفاده می شود. یک GPU معمولاً بر روی کارت های گرافیکی قرار می گیرد، اگرچه کارت های گرافیکی غیر حرفه ای مستقیماً بر روی مادربرد به صورت OnBoard قرار می گیرند. GPU ابزاری است شامل تعدادی عملگر ابتدایی گرافیکی، که باعث می شود نسبت به CPU در خلق تصاویر بر صفحه نمایشگر بسیار سریعتر عمل کنند. رایج ترین عملگرها برای گرافیک دو بعدی کامپیوترها شامل عملگر بیت بلیت (Bit Blit) است که معمولاً در سخت افزارهای مخصوص یک "Biltter" نامیده می شود. این عملگرها برای کشیدن مستطیل، مثلث، دایره و قوس بکار می روند. پردازنده های گرافیکی جدید، پردازش گرافیک سه بعدی را نیز در رایانه ها انجام می دهند.

کاربردهای GPU

کاربرد اصلی این سخت افزار، تقویت گرافیک ۳ بعدی است؛ اما پس از مدتی، برنامه نویس ها دریافتند که امکان استفاده بیشتر از توان آن وجود دارد. کارهایی مثل:

- ✓ **صنعت بازی سازی:** همان طور که گفتیم، اولین کاربرد GPU در ساخت بازی بود و همچنان یکی از اصلی ترین کارآمدی هایش در این زمینه است.
- ✓ **ادیت ویدیو و تولید محتوا:** در پردازش موازی که جزو ویژگی های GPU است، رندر گرفتن از ویدیوها و انجام کارهای مختلف گرافیکی، بسیار سریع تر از گذشته انجام می شود.
- ✓ **Machine Learning:** از آنجایی که پردازشگر گرافیکی توان پردازشی بالایی دارد، بسیاری از تکنولوژی های Deep Learning و AI (هوش مصنوعی) به آن وابسته هستند. شاید فکر کنید GPU همان کارت گرافیکی است که از گذشته روی دستگاه های مان داشتیم؛ البته تفاوت هایی (هرچند کوچک) بین این دو وجود دارد.

انواع GPU

✓ GPU یکپارچه (Integrated)

پردازنده‌های گرافیکی مجتمع یا مشترک (IGP) از بخشی از حافظه RAM کامپیوتر برای محاسبات خود استفاده می‌کند و دیگر مثل پردازنده‌های گرافیکی اختصاصی به یک حافظه RAM جداگانه مجهز نیست. اکثر این نوع پردازنده‌های گرافیکی به صورت مجتمع با برد اصلی ارائه می‌شوند. البته در APU ها از این هم فراتر می‌روند و در داخل خود CPU ساخته می‌شوند. امروزه ۹۰٪ کامپیوترها مجهز به پردازنده گرافیکی مجتمع هستند. این نوع پردازنده‌های گرافیکی نسبت به نوع اختصاصی باعث کاهش هزینه سیستم می‌شوند ولی از سویی قابلیت‌های آن‌ها نیز کمتر است. سابقاً این نوع GPU ها در اجرای بازی‌های سه بعدی و ویدئوهای HD با مشکل مواجه می‌شدند ولی با رشد تکنولوژی و معرفی نسل‌های جدیدتر، این پردازنده‌های گرافیکی نیز بهبود یافته و قدرتمندتر شدند. با این حال هنوز هم کاربران حرفه‌ای گرافیک و علاقه‌مندان به بازی‌های سنگین کامپیوتری باید یک پردازنده گرافیکی اختصاصی برای رایانه خود تهیه کنند.

از آنجا که GPU ها میزان ارجاع به حافظه بالایی دارند، ممکن است نوع مجتمع آن‌ها در مواقعی که فضای خالی حافظه رم کاهش می‌یابد، برای استفاده از آن با CPU وارد رقابت شود IGP. ها می‌توانند تا ۲۹.۸۵۶ گیگابایت بر ثانیه از پهنای باند حافظه را به خود اختصاص دهند. این در حالی است که کارت‌های گرافیکی اختصاصی می‌توانند تا ۲۶۴ گیگابایت از پهنای باند حافظه اختصاصی خود بهره گیرند. در مدل‌های قدیمی پردازنده‌های گرافیکی مجتمع امکان انتقال و نوردهی تصویر به صورت سخت‌افزاری وجود نداشت، اما مدل‌های جدیدتر این قابلیت‌ها را دارا می‌باشند. نکته مثبت این نوع پردازشگر، مصرف پایین انرژی و البته تولید گرمای کم‌تر است.

✓ GPU مجزا (Discrete)

پردازنده‌های گرافیکی بسیار قدرتمند معمولاً به‌طور جداگانه تهیه شده و بوسیله یک درگاه اختصاصی مثل PCI EXPRESS یا AGP یا HDMI به برد اصلی متصل می‌شوند. بدین ترتیب به راحتی می‌توان آن‌ها را تعویض نموده یا در صورتی که مادربرد از مدل‌های جدیدتر پشتیبانی کند آن‌ها را ارتقا داد. یک پردازنده گرافیکی اختصاصی لزوماً جاداشدنی نیست و حتی می‌تواند از طریق درگاه‌های دیگری به غیر از موارد استاندارد مذکور به برد اصلی متصل گردد. کلمه «اختصاصی» برای این استفاده می‌شود که در این نوع GPU ها یک حافظه RAM اختصاصی برای استفاده کارت گرافیک در نظر گرفته شده‌است. البته اکثر پردازنده‌های گرافیکی اختصاصی جاداشدنی و قابل تعویضند. در کامپیوترهای شخصی کوچک‌تر مثل لپ‌تاپ‌ها معمولاً برای اتصال GPU اختصاصی به برد اصلی از رابط‌های غیر استاندارد استفاده می‌شود تا حجم کمتری اشغال کنند، ولی اغلب آن‌ها با وجود تفاوتی که در شکل و اندازه دارند از همان منطق ارتباطی PCIe و AGP بهره می‌برند. مسلماً این نوع از پردازشگرها، توانایی بسیار بیشتری از GPU های یکپارچه دارند؛ البته که انرژی بیشتری مصرف هم می‌کنند و باید تدبیری برای خنک کردن آن‌ها اندیشید.

✓ GPU ترکیبی یا هیبریدی

این نوع GPU مابین دو نوع قبلی قرار می‌گیرد. معمول‌ترین روش‌های پیاده‌سازی این نوع پردازنده‌های گرافیکی، روش Hyper Memory کمپانی ATI و Turbo Cache انویدیا هستند. کارت‌های گرافیک هیبریدی نسبت به IGP ها قیمت بالاتری دارند و از کارت‌های گرافیک اختصاصی ارزان‌تر هستند. این نوع GPU ها از حافظه مشترک با سیستم استفاده می‌کنند و در کنار آن یک حافظه کش اختصاصی نیز برای خود دارند، تا با استفاده از آن تاخیر بالای ناشی از دسترسی به

حافظه رم را جبران کنند. در تبلیغات این نوع محصولات گاهی گفته می‌شود که حجم حافظه رم آن‌ها ۷۶۸ مگابایت است، ولی در واقع این میزان حجمی از رم سیستم است که این نوع پردازنده گرافیکی می‌تواند از حافظه اصلی رایانه اشغال کند.

GPU ابری:

بدون شک ابرها، انقلاب بسیار بزرگی در دنیای اینترنت بودند و هستند. با استفاده از ابرها، امکان استفاده از بسیاری از خدمات، در بستر اینترنت فراهم می‌شود. وقتی صحبت از GPU ابری می‌شود یعنی بدون اینکه GPU تهیه کنید، بتوانید از توان پردازشی یک GPU بهره ببرید. GPU ابری مناسب کسانی می‌باشند که تنها کار می‌کنند و بودجه چندانی ندارند. به همین خاطر، تهیه کردن GPUهای توانا و قدرتمند برای آن‌ها مقدور نیست. این دسته از افراد می‌توانند از ارائه‌دهنده‌های مختلف، به صورت ابری GPU بگیرند و بر اساس استفاده‌ای که داشته‌اند، هزینه را پرداخت کنند.

GPU های ابری شامل 5 مورد زیر می باشند (تمامی مقادیر نمونه و متغیر هستند)

✓ گرافیک و رم مانند : 1 tesla v100(ram 16 gig)

✓ پردازنده : XENON E5-2690 V4 12 THREADS

✓ رم 256 GB و هارد 1200 GB SSD

✓ Deep Learning Performance (24.2)

✓ سیستم عامل ویندوز و لینوکس

مزایای GPU ابری :

✓ راحت‌تر شدن امور برای کسب‌وکار : داشتن GPU اختصاصی، به معنای صرف شدن مقدار زیادی نیرو و انرژی برای نگهداری از آن است. تهیه GPU ابری، این نیرو و انرژی را برای شرکت یا فرد ذخیره می‌کند.

✓ تخصصی عمل کردن کسب‌وکار : وقتی انرژی و زمان لازم برای نگهداری از GPU ذخیره شود، یعنی مجموعه یا فرد می‌تواند از آن‌ها برای پیشبرد بهتر امور اصلی استفاده کنند. به این ترتیب، امکان تمرکز بیشتر روی هدف اصلی و تخصصی فراهم می‌شود.

✓ کاهش هزینه‌ها : احتمالاً می‌دانید که قطعه‌ای مثل GPU، قیمت نسبتاً بالایی دارد. شاید تهیه کردن آن برای همه توجیه‌پذیر نباشد. با استفاده از GPU ابری، به اندازه نیازی که دارید از پردازشگر گرافیکی (GPU) استفاده می‌کنید و به همان میزان هم پول خرج می‌کنید.

✓ افزایش خلاقیت : مشخصاً استفاده از GPU ابری، کمک می‌کند تا دست هر تیمی برای بروز خلاقیت‌ها و ارائه نوآوری کاملاً باز باشد.

تاریخچه GPU

دهه ۱۹۷۰: تراشه‌های ANTIC و CTIA امکان کنترل سخت‌افزاری حالات گرافیکی و متنی و سایر تأثیرات را بر روی داده‌های ۸-بیتی را فراهم می‌کردند. تراشه ANTIC پردازشگری مخصوص برای نگاشت متن و داده گرافیکی به خروجی تصویری بود. طراح تراشه ANTIC جی مینر، طراحی تراشه گرافیکی را برای کمودور آمیگا نیز بر عهده داشت.

دهه ۱۹۸۰: شاید بتوان ایده اولیه شکل‌گیری مفهوم GPU را به دهه ۸۰ و کمودور آمیگا نسبت داد. این کامپیوتر اولین دستگاهی بود که با یک واحد پردازش گرافیکی استاندارد ارائه می‌شد و از ویژگی‌های گرافیکی قابل ملاحظه‌ای در زمان خود بهره می‌برد. در همین دهه و در سال ۱۹۸۷ کمپانی Texas Instrument اولین ریزپردازنده دارای قابلیت‌های گرافیکی را با نام TMS3401 ارائه کرد.

دهه ۱۹۹۰: در دهه ۹۰ روند رشد کارت‌های گرافیکی و پردازنده‌های دارای قابلیت‌های گرافیکی به شکل روزافزونی پیش رفت و برای اولین بار پردازنده‌ها و کارت‌های گرافیکی با قابلیت رندر تصاویر و ویدیوهای سه‌بعدی ارائه شدند. ولی نقطه آغاز آنچه که امروز آن را GPU می‌نامیم به سال ۱۹۹۹ و محصول متفاوت کمپانی انویدیا باز می‌گردد.

Nvidia GeForce 256 که به آن NV10 نیز می‌گویند (اولین کارت گرافیک تجاری بود که توانست قابلیت‌های سه‌بعدی و انتقال و نوردهی را در یک سخت‌افزار تجاری برای مصرف‌کننده ارائه دهد) و از آنجا که بعدها قابلیت سایه‌زنی پیکسلی و براداری به آن اضافه شد، بسیار انعطاف‌پذیرتر و قابل برنامه‌ریزی‌تر از کارت‌های گرافیکی سه‌بعدی هم‌دوره خود بود.

۲۰۰۶ تاکنون: با معرفی سری GeForce8، پردازنده‌های گرافیکی به دستگاه‌های محاسباتی عمومی‌تری تبدیل شدند. امروزه GPU های موازی در مقابل CPU ها رشد کرده‌اند و به همین دلیل یکی از زمینه‌های پژوهشی مهم کنونی، یعنی GPU های همه‌منظوره یا GPGPU رایانش و محاسبات عمومی بر روی GPU توانسته در بسیاری از حوزه‌ها مثل فراگیری ماشین (یادگیری ماشین)، هوش مصنوعی، اکتشاف نفت، پردازش تصویر، جبر خطی، آمار، بازسازی سه‌بعدی و حتی تعیین قیمت در بازار سهام، به یکی از پرکاربردترین زمینه‌های تحقیقاتی و صنعتی تبدیل گردد.

کمپانی‌های تولیدکننده GPU

کمپانی‌های بسیاری در زمینه تولید GPU فعالند. در سال ۲۰۰۸، اینتل، انویدیا و AMD/ATI به ترتیب با ۴۹.۵٪ و ۲۷.۸٪ و ۲۰.۶٪ از سهم بازار در صدر تولیدکنندگان GPU بودند. البته در این آمار، پردازنده‌های گرافیکی مجتمع با CPU اینتل نیز محاسبه شده‌اند. اگر آن‌ها را در نظر نگیریم، انویدیا و ATI تقریباً کل بازار را در اختیار دارند. دو شرکت کوچکتر S3 Graphics و Matrox نیز به تولید GPU می‌پردازند.

GPU و نمونه‌هایی از رایانش ابری

✓ Gpu و شبیه‌سازی میکرو مغناطیس: خدمات رایانش ابری مجموعه‌ای جامع از ابزارها برای انجام محاسبات بر روی منابع سخت‌افزاری هستند که از طریق اینترنت ارائه می‌شوند. ارائه‌دهندگان، دسترسی به رایانه‌های مجازی را می‌فروشند که به عنوان نمونه‌هایی شناخته می‌شوند که بر روی سخت‌افزار آنها اجرا می‌شوند و می‌توانند بر اساس تقاضا یا رزرو راه‌اندازی شوند. نمونه‌ها در انواع پیکربندی‌های سخت‌افزاری وجود دارند که در قیمت‌های ساعتی آن‌ها

منعکس می‌شوند. شبیه‌سازی‌های میکرومغناطیسی پیش‌بینی‌های کمی را برای فیزیک مغناطیسی پیچیده، از جمله تأثیرات مغناطیس زدایی، گشتاور انتقال اسپین و تعامل Dzyaloshinskii Moriya ارائه می‌کنند پیشرفت‌های اخیر در واحدهای پردازش گرافیکی (GPU) باعث ادغام چنین ظرفیت محاسباتی در بسته‌های میکرومغناطیسی شده است. ویژگی موازی انبوه پردازنده‌های گرافیکی مخصوصاً برای تسریع محاسبات اختلاف محدود بزرگ، مانند شبیه‌سازی دینامیک مغناطیسی در فیلم‌های گسترده و ساختارهای لایه کامل اتصالات تونل مغناطیسی مناسب است. با این حال، محاسبات مبتنی بر GPU به سخت افزار تخصصی نیاز دارد. علاوه بر این، شبیه‌سازی‌های مبتنی بر GPU فعلی بر اساس کتابخانه نرم افزار CUDA هستند، که محدود به سخت افزارهای ساخت NVIDIA است و دسترسی به آنها را بیشتر محدود می‌کند. هنگام استفاده از خدمات رایانش ابری برای انجام شبیه‌سازی میکرومغناطیسی مبتنی بر GPU، کاربر ابتدا یک نمونه GPU را بر روی سرورهای ارائه‌دهنده راه اندازی می‌کند. به جای نصب نرم‌افزار لازم پس از هر راه‌اندازی، می‌توان نمونه‌ها را بر اساس «تصاویر» قبلی ایجاد کرد که بسته‌های میکرومغناطیسی و پشتیبانی را از قبل نصب کرده‌اند.

فایل‌های ورودی شبیه‌سازی به نمونه در حال اجرا منتقل می‌شوند و شبیه‌سازی بر روی سخت‌افزار راه دور تا تکمیل اجرا می‌شود. سپس داده‌ها به رایانه محلی کاربر منتقل می‌شوند. در این مرحله می‌توان نمونه را متوقف کرد تا از تحمیل هزینه‌های ساعتی بیشتر جلوگیری کرد، یا برای ادامه شبیه‌سازی‌های دیگر باز نگه داشت. واحدهای پردازش گرافیکی موازی (GPU) می‌توانند سرعت شبیه‌سازی‌های مغناطیسی را در مقایسه با محاسبات معمولی با استفاده از واحدهای پردازش مرکزی (CPU) به طور قابل توجهی بهبود بخشند. ما یک استراتژی برای انجام شبیه‌سازی‌های میکرومغناطیسی شتابدار GPU با استفاده از دسترسی مقرون‌به‌صرفه GPU ارائه‌شده توسط سرویس‌های رایانش ابری با یک برنامه منبع باز مبتنی بر پایتون برای اجرای کد میکرومغناطیسی MuMax3 از راه دور ارائه می‌کنیم. ما مقیاس بندی و مزایای هزینه استفاده از محاسبات ابری برای میکرومغناطیس را تجزیه و تحلیل می‌کنیم. پلتفرم‌های رایانش ابری برای سرویس‌های وب، می‌توانند سرعت قابل توجهی را هنگام اجرای دسته‌های بزرگ شبیه‌سازی ارائه دهند، بدون اینکه متحمل چیزی فراتر از هزینه‌های نمونه ساعتی استاندارد شوند. ما یک اسکریپت منبع باز Python به نام MuCloud ایجاد کرده‌ایم که شبیه‌سازی‌های MuMax3 را بر روی نمونه‌های GPU AWS بدون توجه به سیستم عامل محلی کاربر اجرا می‌کند.

برای اهداف امنیتی، پورته ایمن (SSH) و پروتکل انتقال فایل امن (SFTP) تضمین می‌کنند که تمام داده‌ها هنگام عبور بین رایانه‌های محلی و راه دور رمزگذاری شده‌اند. رابط وب MuMax3 در دسترس است به طوری که کاربر محلی می‌تواند شبیه‌سازی‌ها را در نمونه راه دور در زمان واقعی کنترل و نظارت کند.

خدمات رایانش ابری ابزاری را برای محققان فراهم می‌کند تا بدون سرمایه‌گذاری در سخت‌افزار تخصصی کامپیوتر، بهبود عملکرد شبیه‌سازی‌های میکرومغناطیسی مبتنی بر GPU را به دست آورند. این امکان‌های جدیدی مانند شبیه‌سازی همزمان در تعداد زیادی از نمونه‌های راه دور را باز می‌کند. ما یک برنامه منبع باز (MuCloud) ارائه می‌کنیم که به شبیه‌سازی MuMax3 اجازه می‌دهد تا بر روی نمونه‌های AWS اجرا شود، به طوری که محققان می‌توانند به راحتی به این راه جدید برای میکرومغناطیس دسترسی داشته باشند. با استفاده از این ابزارها، ما نشان می‌دهیم که هنگام شبیه‌سازی سیستم‌های بزرگ، می‌توان عملکرد تقریباً ده برابری را بر روی کدهای میکرومغناطیسی مبتنی بر CPU به دست آورد.

✓ GPU و بسته پردازش NFV: پارادایم مجازی سازی تابع شبکه NFV از جایگزینی سخت افزار با هدف خاص پشتیبانی از پردازش بسته با سخت افزارهای همه منظوره حمایت میکند و هزینه‌ها را کاهش میدهد و انعطاف پذیری و چابکی بیشتری را برای عملیات شبکه به ارمغان می‌آورد. با این حال، این تغییر می‌تواند عملکرد شبکه را به دلیل عدم بهینه بودن کاهش دهد.

قابلیت های پردازش بسته سخت افزار همه منظوره نیز چنین است. در همین حال، GPU در بسیاری از مراکز داده Data Center به دلیل استفاده گسترده از آنها در یادگیری ماشین Machine Learning مستقر شده اند. این پردازنده های گرافیکی میتوانند برای سرعت بخشیدن به قابلیت پردازش بسته NFV مورد استفاده قرار گیرند. اما تاخیر معرفی شده می تواند برای برخی از برنامه ها مشکل ساز باشد.

اخیراً پردازنده های گرافیکی در بسیاری از حوزه های غیر از رندر ویدئویی که در ابتدا برای آنها در نظر گرفته شده بود به دلیل بسیار زیاد قابلیت پردازش موازی، این قابلیت که به عنوان واحد پردازش گرافیکی با هدف عمومی GPGPU شناخته می شود، بسیاری از پیشرفت های اخیر را در یادگیری ماشین ممکن کرده است. کاربرد موفقیت آمیز پردازنده های گرافیکی در مناطق مختلف، ارائه دهندگان رایانش ابری را به استقرار آنها در مراکز داده سوق داده است. در شبکه، پردازنده های گرافیکی برای انتقال دهنده IPv4/v6 دروازه IPsec، بازرسی بسته عمیق و الگوریتم رمز اعمال شده اند. با این حال، برخی از جنبه ها، مانند تاخیر معرفی شده توسط GPU ها به پردازش بسته، تا کنون مورد مطالعه قرار نگرفته است.

GPU های واحد پردازش گرافیکی با هدف عمومی معمولاً در عملیات ضرب ماتریس استفاده می شوند. مفهوم GPGPU در سال 2001 با پشتیبانی از عملیات ممیز شناور در GPU ها به عنوان راهی برای محاسبه هر چیزی غیر از عملیات گرافیکی معرفی شد. در سال 2006، انویدیا معماری دستگاه یکپارچه محاسباتی CUDA را منتشر کرد که اجرای کد را بر روی GPU ها بدون نیاز به تبدیل کامل و صریح داده ها به/از یک فرم گرافیکی امکان پذیر می کرد. این معماری عامل اصلی پیشرفتهای اخیر در چندین زمینه، مانند آموزش مدل های ML در مقیاس بزرگ است. هر SM دارای چندین پردازنده واحد است. تمرکز معماری GPU بر تعداد واحدهای منطقی حسابی ALUs برای افزایش توان عملیاتی است. در مقابل، واحدهای پردازش مرکزی CPU روی بخش بزرگی از تراشه به حافظه نهان تمرکز میکنند و تأخیر دسترسی به حافظه را کاهش میدهند.

جریان CUDA مدیریت وظایفی که باید توسط یک برنامه CUDA انجام شود، مانند اجرای GPU و انتقال بین حافظه CPU-GPU، توسط صفی از عملیات به نام استریم انجام می شود که پس از اولین ورود، اولین خروج FIFO انجام می شود. اگر جریان CUDA که در آن اجرای برنامه باید برنامه ریزی شود مشخص نشده باشد، به یک جریان پیش فرض اختصاص داده می شود. در این حالت، هر عملیات به صورت موازی با بسیاری از هسته های CUDA اجرا میشود، اما به صورت متوالی انجام میشود، بنابراین اگر برنامه از عملیات ذاتا مستقل تشکیل شده باشد، منجر به کاهش عملکرد می شود. یک رویکرد جریان چندگانه می تواند برای جلوگیری از این مشکل با انجام عملیات متقابل دستگاه یا اجرای همزمان GPU بر روی یک دستگاه استفاده شود. تفاوت بین این رویکردها را با نشان دادن مثالی از یک برنامه کاربردی CUDA در حال اجرا در رویکردهای تک جریانی و چند جریانی برای دو وظیفه یعنی بافرهای بسته در زمینه این کار نشان می دهد. برای سادگی، هر اجرا توسط دو عملیات متوالی اولیه یعنی انتقال داده و پردازش بسته تشکیل می شود.

منابع

1. A GPU-Assisted NTV framework for intrusion detection system
2. GPU-accelerated micromagnetic simulation using cloud computing
3. GPU computing with NVIDIA CUDA