Implementación de las GPUs para un futuro inteligente

Fredy Alejandro Mendoza López
Universidad Industrial de Santander
Ingeniería de Sistemas
ifredomendoza@gmail.com

Andrés Felipe Uribe García
Universidad Industrial de Santander
Ingeniería de Sistemas
andresfelipeuribe11@gmail.com

Orlando Alberto Moncada Rodriguez
Universidad Industrial de Santander
Ingeniería de Sistemas
orlandomoncada610@gmail.com

David Santiago Morales Norato
Universidad Industrial de Santander
Ingeniería de Sistemas
d.s.norato@gmail.com

Laura Marcela Mantilla Romero
Universidad Industrial de Santander
Ingeniería de Sistemas
lauramantilla123@gmail.com

Resumen

A lo largo de los años, NVIDIA se ha caracterizado por ser uno de los líderes en el desarrollo de procesamiento gráfico en alta calidad. En la SC19 NVIDIA presenta un plan para superar sus avances de manera significativa, ya que se ha llegado a un límite físico dado en la cantidad de transistores posibles a usar, lo cual afecta la tasa de aumento en operaciones por segundo. La solución propuesta, es abordar los diferentes problemas dividiendo los procesos según la idoneidad de cada arquitectura. Todo esto nos lleva a trazar un camino del avance de la humanidad en diversos campos de la ciencia el cual está fuertemente enlazado con la inteligencia artificial, GPUs, HPC, entre otros procesos de optimización que permitirá dar al hombre ese gran paso a cosas que nunca imaginamos.

Abstract

Over the years, NVIDIA has been characterized as one of the leaders in the development of high-quality computing equipment in graphic processing. In SC19, NVIDIA presents us with a design to overcome its advances significantly, since it has reached a point where increasing the number of transistors per unit of space is limited by the physical space, which affects the rate of increase in operations per second. The proposed solution is handle the different problems by dividing the processes according to the suitability of each architecture. All mentioned previously is useful to trace the future of humanity in many the fields of science, this future is heavily relationed with Artificial inteligence, GPUs, HPC, and other important branches of science, this will introduce us to a era wich we had never imagined before.

Index Terms

Ray Tracing, HPC, IA, NVIDIA.

I. Introducción

La invención de las GPUs hizo posible el sombreado programable en tiempo real, dando a los artistas una paleta infinita de expresión. Este artículo ofrece una visión general del estado actual de las técnicas de GPU para la visualización interactiva de volúmenes a gran escala, para ello, es importante que el lector conozca acerca de la actualidad de NVIDIA y su importancia. NVIDIA es una compañía especializada en el diseño de unidades de procesamiento de gráficos (GPU) que se utilizan en computadoras [1], se caracteriza por ser líder en este mercado, especialmente para el mercado de juegos, así como de sistemas en chips o SOCs y mercados de computación móvil y automotriz. NVIDIA ha revolucionado el desarrollo de las GPU, convirtiendo estas en un cerebro de computadora en la intersección de la realidad virtual, la computación de alto rendimiento y la inteligencia artificial (IA).

Con el uso de la inteligencia artificial hoy en día las máquinas tienen la capacidad de .ªprender", de modo que gracias a esto se pueden resolver grandes desafíos que han estado más allá del alcance del ser humano, la IA es una entidad que impulsa el cambio en todas las industrias de todo el mundo, a medida que evoluciona también lo hace la humanidad, pero para que esto suceda, debe ser alimentada por un poder de cómputo masivo. Gracias a la implementación de las GPU de NVIDIA, se están logrando cosas como la revolución analítica, prevención de enfermedades y construcción de ciudades inteligentes, hoy en día, estas tecnologías están siendo una base importante para las organizaciones, con el fin de llevar sus proyectos a la realidad.

La ley de Moore nos expresa que aproximadamente cada dos años se duplica el número de transistores en un microprocesador, pero hoy en día la demanda de potencia informática es mucho mayor por lo que la computación de alto rendimiento (HPC) está definiendo una nueva ley sobrealimentada para reemplazar la ley de Moore. Comienza con un procesador paralelo de GPU altamente especializado y continúa a través del diseño del sistema, software, algoritmos y aplicaciones optimizadas.

Cada servidor acelerado por GPU reemplaza a docenas de servidores de CPU básicos, brindando un impulso dramático en el rendimiento de la aplicación y el ahorro de costos por lo cual esta computación nos ofrece un rendimiento informático en donde se podrán realizar miles de millones de cálculos por segundo, desde la predicción del clima y la ciencia de los materiales hasta la simulación del túnel de viento. La computación acelerada por GPU NVIDIA es el corazón de las áreas de descubrimiento más prometedoras de HPC.

Hace décadas, el trazado de rayos ha usado una serie de algoritmos complejos para representar gráficos de computadora de manera realista. Al rastrear los rayos de luz individuales a medida que rebotan en múltiples superficies, puede recrear los reflejos, la dispersión por debajo de la superficie, la translucidez y otros matices que ayudan hacer una escena convincente. Se usa comúnmente en películas y publicidad, pero la gran cantidad de tiempo de procesamiento necesario lo ha mantenido alejado de las aplicaciones en tiempo real como los juegos. Con el GeForce R-TX de NVDIA, usando su mini-arquitectura de GPU Turing, el objetivo es implementar un soporte mejorado en tiempo real para RTX, su biblioteca de trazado de rayos de alto rendimiento que puede producir escenas detalladas a velocidades de fotogramas hiperrealistas hasta en los juegos más exigentes.

Con la implementación de estas tecnologías, la aceleración de GPU es el camino más eficiente y accesible en energía para las computadoras más poderosas del mundo, los caminos de la informática de alto rendimiento y la innovación de la inteligencia artificial están convergiendo, desde las supercomputadoras más grandes del mundo hasta los grandes centros de datos que alimentan la nube, este nuevo modelo de computación está ayudando a responder preguntas complejas, a descubrir nuevas ciencias y brindar capacidades sorprendentes a nuestros dispositivos móviles.

II. Estado del arte

En noviembre de 2019, NVIDIA mostró al público de la SC19 una pequeña simulación del aterrizaje de un viaje tripulado a Marte, en donde se muestra el campo de [2] alrededor de un módulo de aterrizaje. Este viaje puede durar alrededor de diez meses, no obstante, puede ser en vano si algo sale mal en los últimos seis minutos de la misión, ya que se debe reducir su velocidad de 12 mil millas por hora a casi cero para obtener un aterrizaje suave, por ende, es necesario que la simulación involucre la física necesaria. La simulación hecha en una GPU NVIDIA, cuenta con más de 6 mil millones de datos de células de tetraedros, prismas y pirámides, por lo que se convierte en la visualización de volumen interactivo más grande del mundo, la cual se analiza en el sistema Summit impulsado por GPU NVIDIA, la supercomputadora más rápida del mundo.

En cuanto a la inteligencia artificial se han mostrado sorprendentes trabajos en la literatura, un ejemplo es el descubrimiento de un nuevo antibiótico gracias al Deep Learning [3], esto representa uno de los trabajos más interesantes en el presente año ya que la medicina está empezando a hablar sobre la resistencia de las bacterias hacia los antibióticos, el uso de antibióticos muy fuertes es más frecuente y preocupa a la comunidad científica si en un futuro las bacterias generen total inmunidad. Por esto el descubrimiento de un nuevo antibiótico por este método representa que el Deep Learning tiene gran futuro en nuestra sociedad.

GeForce NOW es uno de los productos NVIDIA que en los últimos años ha atraído más atención de propios y extraños en el reciente nicho de mercado ocupado por el “cloud gaming service”, servicio de streaming de videojuegos en la nube que le permite al cliente jugar títulos propios en un ordenador remoto, soportado por un clúster GPU de alto rendimiento basado, resultado de la utilización de tecnología de tarjetas gráficas Nvidia RTX en las respectivas GPU’s que conforman cada nodo del Clúster que provee el servicio.

III. Marco teórico

III-A. Rasterización

El salto de una representación 2D a 3D marcó un antes y después en los videojuegos, ya que este último requería de nuevas tecnologías para poder ser implementado. Una técnica muy popular y que se usa actualmente es la rasterización, la cual consiste en llevar gráficos 3D a una pantalla bidimensional, esto, por medio de una malla de triángulos virtuales, los cuales crean modelos tridimensionales de objetos [4]. Los vértices asocian mucha información relevante, incluyendo la normal, la cual determina la forma del objeto que se verá reflejada en la pantalla digital. En sí, la implementación de esta técnica ya es una tarea intensiva en términos de computo, ya que se actualiza alrededor de 30 a 140 veces por segundo, no obstante, el procesamiento de pixeles se realiza sobre la imagen 2D, incluyendo los reflejos y la iluminación, esto indica que, los reflejos que encontramos en los videojuegos, son efectos predeterminados que el desarrollador implementa para hacer alusión a un entorno con luz presente, por lo tanto, la iluminación no es renderizada en tiempo real, ni la luz se comporta como el ser humano la percibe diariamente, lo cual genera un nuevo reto en el campo del procesamiento gráfico.

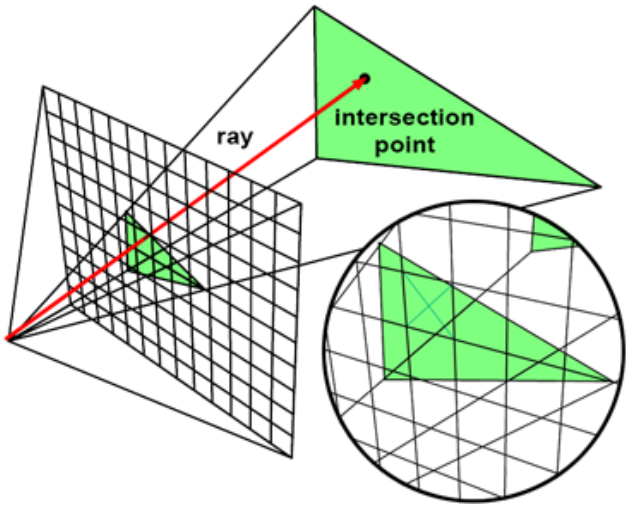

Figura 1. Como funciona la rasterización

III-B. Ray Casting y Ray Tracing

Los objetos del día a día, se pueden apreciar gracias a una fuente de luz; los fotones que perciben los ojos pueden ser sometidos a diferentes factores antes de llegar a ellos, como rebotar de un objeto a otro, el bloqueo de luz generando sombra, el reflejo de un objeto en otro, y los diferentes fenómenos que genera la luz. En 1968, Arthur Apple junto a IBM, lograron implementar un algoritmo que recogiera todos estos fenómenos y se pudieran ver reflejados gráficamente [5], la cual denominaron Ray Casting. En 1979, Turner Whitted, mostró como capturar reflejos, sombras y refracción, además, mostró como los choques de la luz entre objetos, contribuyen al color final del pixel, a esta técnica la denominó Ray Tracing [6]. El Ray Tracing cubre el problema del comportamiento de la luz en tiempo real, ya que, a diferencia de la rasterización común, la luz es aplicada en el objeto tridimensional, generando así los fenómenos deseados sobre la imagen. Esta técnica fue puesta en práctica hace varios años, un claro ejemplo de esto son las películas de animación, las cuales la usaron con el objetivo de obtener imágenes más realistas, sin embargo, la implementación del ray tracing a una GPU no era posible hasta hace algunos años.

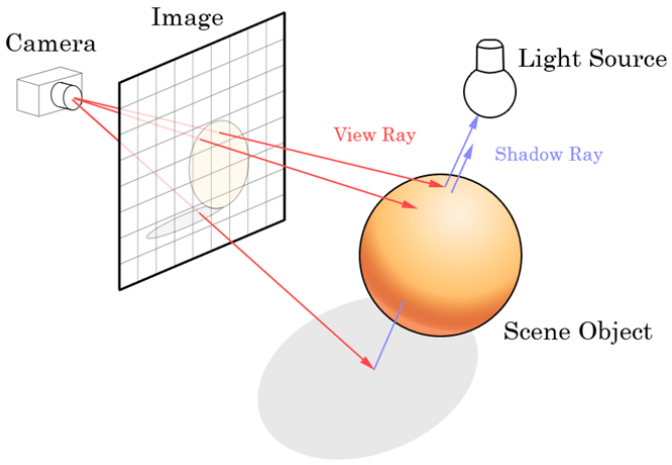

Figura 2. Como funciona Ray Tracing

III-C. Redes neuronales

Una neurona en el contexto de la inteligencia artificial es una función matemática que toma un vector de entrada que se multiplica por un vector de pesos, se suman los elementos del vector resultante y ese resultado es pasado a través de una función activación no lineal que es la encargada de agregar no linealidad característica de los problemas resueltos con deep

learning, luego creando conexiones entre las neuronas se crean las redes neuronales, se organizan en capas como se muestra en la Fig.3, estas redes neuronales pasan por un proceso llamado entrenamiento donde se minimiza una función de conste definida previamente, este proceso requiere conjuntos de datos de gran tamaño, esta necesidad más representa un problema que será expuesto más adelante.

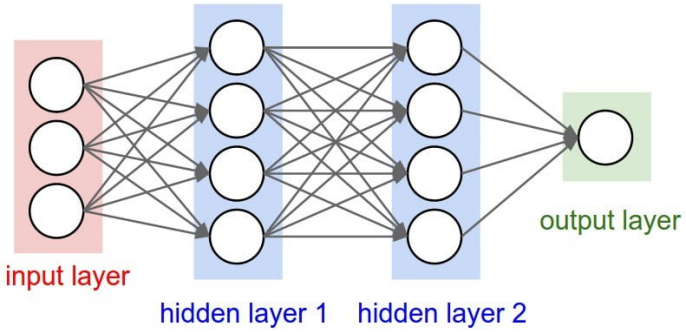

Figura 3. Representación gráfica usual de una red neuronal

III-D. Clúster

En la computación el termino clúster refiere al conglomerado de servidores conectados por medio de una red de alta velocidad y configurados para comportarse como un único servidor, resultado de la convergencia de varias tendencias tecnológicas actuales, entre las que destacan la disponibilidad de microprocesadores de alto rendimiento a precios económicos, que dan como resultado un ordenador con un balanceo de carga, rendimiento y escalabilidad superior al obtenido por computadores individuales de precio y disponibilidad comparable, siendo además más sencillos en su montaje debido a su flexibilidad, permitiendo configuraciones homogéneas en hardware y sistema operativo, o por el contrario arquitecturas y sistemas operativos que van desde lo similar como por ejemplo configuraciones de modelos de la misma GPU, por ejemplo, 8800GT mezclado con 8800GTX, a lo radicalmente distinto valiéndose de hardware de distintas IHV como NVIDIA y su competidora AMD.

IV. Historia de NVIDIA

NVIDIA fue fundada por JEN-HSUN "JENSEN" HUANG, CHRIS MALACHOWSKY y CURTIS PRIEM en 1993 y situaron su sede en California en abril 1993. NVIDIA nace con la creencia en que la PC algún día se convertiría en un dispositivo esencial a la hora de crear, diseñar y desarrollar juegos y multimedia. El año 1995 fue uno de los más importantes para esta empresa, ya que es lanzada su primera procesador gráfico denominado NVIDIA NV1 cuyo nombre comercial fue Diamond Edge 3D el cual permitía el diseño de juegos y la ejecución de estos en un mapeado 2D o 3D lo cual era una novedad para la época. Con NVIDIA focalizado en el procesamiento y renderizado del 3D, esta modalidad grafica fue aumentando su uso a nivel global rápidamente, después de todo era algo que nos acercaba al futuro planteado en Hollywood, por ende, NVIDIA siguió creando numeroso software y hardware que permitieran facilitar el trabajo con la tercera dimensión virtual. Una de sus más importantes creaciones en este sector fue RIVA 128, el primer procesador 3D que trabajaba a 128 bits. Este procesador vendió más de un millón de unidades en tan solo 4 meses. Con las acciones de NVIDIA subiendo y la demanda de sus productos aumentando, esta creo una alianza de gran vitalidad con Taiwán Semiconductor Manufacturing Company quienes empezaron a ayudar con el proceso de producción de los productos de NVIDIA. Además de ello en este mismo año NVIDIA aumenta su catálogo de procesadores con la ampliación de la familia RIVA, de esta forma nace RIVA128ZX y RIVA TNT, los cuales se caracterizaban por ser el de mayor capacidad de procesamiento 3D y ser el primer procesador en renderizar múltiples texturas en 3D respectivamente. Terminando el milenio NVIDIA crea la primera unidad de procesamiento grafico en el mundo llamada GeForce 256 el cual procesaba 10 millones de polígonos por segundo, otra GPU que salió en este mismo año fue la QUADRO la cual prontamente se volvería el estándar de los profesionales en el diseño, facilitando la creación de planos desde una simple pelota de tenis a las partes de un automóvil. En 2004 NVIDIA realizó el lanzamiento de su tecnología SLI la cual consiste en poder conectar varias GPU lo cual permitió conseguir un incremento drástico en el procesamiento grafico de una sola máquina. Juntamente con esto, llego una asociación de gran magnitud, el Rover de la NASA enviaba imágenes no tan detalladas del suelo marciano por lo cual NVIDIA permitió aumentar el nivel de calidad de las imágenes a una escala realista, esta alianza seguiría por varios años hasta modelara el viaje de personas a Marte. El evento más importante entre 2005 y 2006 fue el anuncio sobre el desarrollo y creación de los CUDA CORES, una arquitectura revolucionaria para el propósito general de las GPU. Esta arquitectura aprovecha las capacidades de procesamiento en paralelo permitiendo a las GPU afrontar retos

más complejos para la computación. Entre 2007 y 2009 NVIDIA realizó el lanzamiento de la arquitectura Tesla en las GPU la cual paso a ser parte de uno de los computadores enumerados en el top 500 de supercomputadores, donde más adelante en 2010 pasaría a ser parte del súper computador chino Tianhe-1A, también el lanzamiento de procesador móvil Tegra en el cual se desato Android. En 2012 nace la arquitectura Kepler en las nuevas tarjetas GTX de serie 600 además de ello Tegra pasa a su 3 versión donde es equipada en dispositivos móviles como tablets y celulares, el cual sería reemplazado en 2013 por el Tegra 4, año en el que NVIDIA anunciaría la GPU GeForce GTX TITAN, una de las más poderosas GPU usadas por video jugadores. Por el año 2015 se rescata el nacimiento de nuevas arquitecturas y ramas de NVIDIA, ya no solo había cuidado por diseñadores y video jugadores, sino las empresas de automóviles y el entretenimiento del hogar se hacía más presente con la NVIDIA DRIVE Y NVIDIA SHIELD. También cabe resaltar la aparición de la NVIDIA Jetson permitiendo una nueva generación inteligente y autónoma de las maquinas. Los años 2016 y 2017 representaron para NVIDIA un salto gigantesco en la producción de GPUs de gran rendimiento con la salida de la arquitectura PASCAL Y VOLTA respectivamente en cada año se centró en una tarea fundamental. En el primero se presentó una mejora para el Deep Learning en GPUs como la DGX-1 mientras que al siguiente año con esta tarea resuelta se pasó a incursionar en la inteligencia artificial en la tarjeta gráfica VOLTA 100. El avance en las distintas ramas de investigación donde trabajaba NVIDIA con el pasar de los días necesitaba más capacidad de procesamiento, por ello se creó una nueva arquitectura denominada TURING las cuales cuentan con Ray Tracing en tiempo real considerado el santo grial de los gráficos computacionales en la actualidad. Por ultimo en 2019 NVIDIA anuncio en la conferencia de supercomputación su mayor adquisición, Mellanox, una empresa de alto rendimiento en tecnología de interconectividad. [7]

V. Inteligencia Artificial

V-A. ¿Qué es la inteligencia artificial y el Deep learning?

Inteligencia artificial(IA) es una rama de la computación que ha tenido un rol significativo en el desarrollo de soluciones óptimas en los últimos años como alternativa a las soluciones analíticas o soluciones basadas en el modelo de un problema de optimización, IA es aplicable a muchos problemas en diversos campos como: visión por computador, procesamiento de lenguaje natural, segmentación de imágenes y videos, logrando así un gran avance en el estado del arte, hoy en día es imposible hablar del desarrollo de la humanidad sin tener en cuenta las posibles aplicaciones de la IA en nuestra sociedad. Este campo de estudio busca desarrollar algoritmos para resolver tareas que usualmente se necesitaría de una mente humana para resolver (tareas que requieren inteligencia), en este orden de ideas se puede decir que la inteligencia artificial es el desarrollo de algoritmos que simulen la inteligencia humana [8]. La IA es un conjunto de campos de estudio en el cual resalta el machine learning(ML), el ML se basa en hacer predicciones por modelos que se optimizan en un proceso llamado entrenamiento, el entrenamiento usa los datos para .ªprender"de estos y realizar unas predicciones acerca de nuevos datos nunca vistos anteriormente [9]. Dentro del machine learning se destaca el deep learning, esta técnica basada en redes neuronales, usa la neurona como unidad básica(usualmente llamado perceptrón) [10], las neuronas son una función matemática que toma un vector de entrada que se multiplican por un vector de pesos, se suman estos elemento y el resultado es pasado a través de una función no lineal de activación que es la encargada de agregar no linealidad característica de los problemas resueltos con deep learning, estas redes neuronales pasan por un proceso llamado entrenamiento donde se minimiza una función de conste definida previamente, este proceso requiere conjuntos de datos de gran tamaño, esta necesidad más representa un problema que será expuesto más adelante.

V-B. Historia de la inteligencia artificial

Los orígenes de la inteligencia artificial se remontan a los inicios de las computadoras, en los 60s los investigadores enfatizaban en la necesidad de desarrollar algoritmos para resolver problemas con un enfoque diferente al matemático clásico [11], los científicos de la computación trabajaban en el desarrollo de máquinas que aprendieran como el robot WABOT-1 [12], en los siguientes años se pasó por un periodo llamado el invierno de la inteligencia artificial debido a que las aplicaciones necesitaban una gran cantidad de datos y tiempo para poder lograr una solución aceptable, los computadores de aquella época no estaban suficientemente desarrollados para lograr tratar esa cantidad de datos en un tiempo relativamente corto, este periodo siguió hasta finales de la década de los 90s donde el creciente desarrollo tecnológico impulsó nuevamente la investigación y el desarrollo en IA. En 1997 la empresa IBM con su computadora Deep Blue [13] logró vencer por primera vez al campeón de ajedrez del mundo. En los últimos años gracias al alto desarrollo tecnológico al que ha llegado la humanidad y a la gran cantidad de datos que nuestra sociedad genera cada día, ha sido posible el desarrollo actual en la inteligencia artificial y especialmente en el deep learning ha mostrado en los últimos años, especialmente en procesamiento de lenguaje natural que c, reconocimiento del genoma, detección de objetos, reconstrucción de imágenes, procesamiento de videos.

V-C. *Retos y capacidades del DL*

No es posible imaginarse un futuro sin que la IA esté presente, incluso hoy en día hay innumerables webs y aplicaciones que hacen uso del deep learning para entregar cada vez mejores productos al mercado, un ejemplo de esto son los vehículos autónomos [14] en el cual cada vez recaen más esperanzas en él, ya que los avances que traería a la humanidad una industria de transporte que se conduzca sola, plantea grandes retos pero también grandes posibilidades para nuestra sociedad, se salvarían muchas vidas causadas por error humano incluso cabe la posibilidad de que muchos precios de productos se reduzcan considerablemente. Otro gran ámbito don de la humanidad se ve y se verá en un futuro cada vez más afectada es la medicina [15], con la creciente cantidad de nuevos dispositivos que brindan ayuda en la medicina viene el aumento en el número de datos que podemos tratar, así por ejemplo el DL ha presentado soluciones a la clasificación de cáncer [16], o a la identificación de células cancerígenas [17]. la utilización de deep learning en la medicina promete un gran aumento en la calidad de vida de los pacientes, reduciendo los tiempos de los procedimientos o incluso el descubrimiento de un nuevo antibiótico [3].

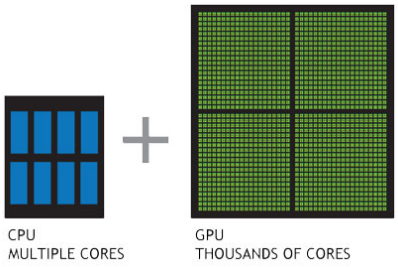

Figura 4. Número de núcleos de una CPU VS número núcleos de una GPU

VI. NVIDIA RTX Y SIMULACIÓN

A lo largo de los años, la industria de los videojuegos ha ido creciendo a pasos agigantados en el desarrollo gráfico, implementando cada vez más técnicas y trucos para lograr el hiperrealismo. Los primeros videojuegos eran bastante simples gráficamente, sencillamente eran pixeles en conjunto representados en una pantalla. A partir de esto, esta área fue evolucionando al tal punto que la conocemos ahora, sin embargo, las limitaciones que nos da la rasterización al aplicar luminosidad a un objeto bidimensional, no permite aprovechar al máximo los resultados gráficos que se pueden generar, por ende, se busca que a partir del Ray Tracing, se puedan producir los resultados anhelados.

VI-A. *NVIDIA RTX*

A partir del trabajo hecho por Whitted, en el 2018, NVIDIA logró implementar el Ray Tracing en una GPU, brindando así la renderización de luz en tiempo real en una unidad de procesamiento gráfico, denominando esta tecnología como RTX, obteniendo así la más actual e hiperrealista GPU, reemplazando así la GTX que tenía en el mercado. Los resultados obtenidos son notorios, los gráficos y la experiencia a la hora de jugar son mucho más realistas. La figura 5 nos muestra las diferencias que nos brinda la RTX ante una tarjeta gráfica convencional, se pueden apreciar fenómenos de la luz como la reflexión, refracción, el reflejo de las esferas en las otras, entre otros.

La serie RTX de NVIDIA, se basa en la novedosa micro-arquitectura de Turing, la cual implementa el trazado de rayos en tiempo real. El trazado de rayos se acelera mediante el uso de núcleos RT, los cuales utilizan diferentes técnicas de colisiones y procesamiento de cuadrantes [18].

VI-B. *El futuro del Ray Tracing*

Sí bien, la implementación del trazado de rayos en tiempo real en una GPU es una gran novedad, esta tecnología sigue en proceso, y aún numerosas compañías no han podido adoptarla ni optimizarla del todo. Es cierto que el Ray Tracing es necesario para llegar al hiperrealismo en los videojuegos, ya que la luz y todo lo que se puede generar a partir de esta, es muy importante en el ambiente gráfico. El apoyo de las empresas va a ser fundamental para el desarrollo y progreso de esta tecnología, la cual ya cuenta con el soporte de importantes APIs de videojuegos, como lo son Unreal Engine, Unity, Vulkan y DirectX [19]. Esta última ha sido uno de sus más grandes pasos, ya que Microsoft llegó a un acuerdo con NVIDIA para brindar soporte total al Ray Tracing con DirectX12, la cual denominó DXR. A partir de esta implementación se probó el rendimiento y la calidad gráfica de las sombras en varios videojuegos [20]. Sin embargo, al ser una tecnología reciente, tiene sus percances, su parte más débil se encuentra en la caída evidente de la tasa de fotogramas, la cual es muy importante para la experiencia del jugador, por esta razón, existe un gran porcentaje de personas que prefieren seguir usando la tecnología anterior, que brinda una tasa de fotogramas estable.

Figura 5. Comparación, sin RTX y con RTX

VI-C. Simulación

La simulación ha sido un proceso muy importante en el desarrollo de proyectos científicos, gracias a ella, se pueden obtener predicciones que aportan mayor probabilidad de éxito a la hora de llevar a cabo un proyecto. El constante desarrollo y mejora de la computación, ha hecho cada vez mayor la posibilidad de la simulación de algunos casos complejos, e inclusive, dar respuesta a algunas preguntas que por falta de tecnología no se podía brindar una solución.

El avance de la supercomputación, el desarrollo de las GPUs y la inteligencia artificial, mantienen una estrecha relación con la simulación, ya que hoy en día se puede apreciar en múltiples casos el gran tamaño de datos que se deben procesar para llevarla a cabo, un claro ejemplo de esto es la simulación del aterrizaje a Marte mencionado anteriormente. Todo este tipo de proyectos generan retos en el campo de la simulación, ya que cada vez se vuelven más exigentes, no obstante, esto al parecer no ha sido un problema, ya que a medida que la simulación se vuelve más minuciosa, le computación avanza de manera equilibrada.

VII. High Performance Computing

VII-A. ¿Que es la HPC?

En el mundo de la computación todo tiende a ser más pequeño a la par de más rápido, esto se hace aún más evidente si hablamos a nivel de HPC (High Performance Computing) y las supercomputadoras, pues a pesar de seguir siendo maquinas ruidosas y voluminosas, su capacidad de procesamiento se ha multiplicado exponencialmente a lo largo de los años, cualesquiera de las máquinas más potentes de hace una década hoy en día no podrían siquiera acercarse al actual top 500. Estos aparatos normalmente asociados a la investigación y el mercado de datos, nichos especializados en la tecnología que parecen alejarse totalmente de la realidad del consumidor doméstico, sin embargo estas supercomputadoras ya se encuentran siendo parte de nuestras vidas seamos o no conscientes de ello, en servicios de streaming, o en uno de los usos más comunes de estas; el modelado de sistemas climáticos, que simula datos pasados, calcula sobre el tiempo actual y predice el clima futuro, ayudando a planificar los vuelos y eludiendo las tormentas más peligrosas antes de que se ciernan sobre nosotros [21].

VII-B. GPGPU como solución de HPC

Si bien las GPU (graphic processing unit) fueron concebidas para dedicarlas al procesamiento gráfico, en nuestros tiempos son muchas las aplicaciones que se han adaptado para que estas arquitecturas alcancen una aceleración significativa en el ámbito de HPC. La velocidad de procesamiento que se puede obtener con estas nuevas plataformas es indiscutible debido a su diseño paralelo ofrecido por sus múltiples núcleos, que permiten el lanzamiento de un alto número de hilos en simultaneidad [22]. Las GPUs, a diferencia de las CPUs, tienen una cantidad mayor de transistores dedicados al procesamiento de datos, en cambio las CPUs dedican muchos de sus transistores para el flujo de control y grandes cachés. Si bien hasta la actualidad, la cantidad de problemas de propósito general resolubles en placas gráficas es menor al número que puede resolverse computacionalmente en las CPUs, estas aplicaciones se han visto altamente beneficiadas disminuyendo el tiempo de procesamiento utilizando sólo una GPU. Hay que aclarar que al hablar de clústeres de GPU se hace referencia a una arquitectura heterogénea, pues es indispensable la existencia de un subsistema conformado por varios cores CPUs y su sistema de memoria entrada/salida, y por otro lado un subsistema GPU con sus memorias on/off chip, dichas entidades están conectadas por medio de un bus PCI-E motivo el cual un inconveniente de su configuración es el cuello de botella que se da al transferir datos entre nodos de distintas arquitecturas.

Pero para poder dividir estas cargas de trabajo de manera eficiente según sea el tipo de operación y peso se debe tener soporte en software capacitado en la realización de esta tarea, para este fin NVIDIA lanzó en 2007 CUDA una herramienta para desarrollo de computación en paralelo que además incluye un compilador y un conjunto de librerías de desarrollo creadas

por la empresa para permitir a los programadores usar variaciones en el lenguaje de programación C para codificar algoritmos en GPU.

VII-C. NVIDIA en el mercado

Actualmente NVIDIA ha comenzado a producir aceleradores de GPU Tesla V100 diseñados específicamente para clústeres de GPU. Estas GPU pensadas para HPC, disponibles como cajas de montaje en rack de 1U contienen cuatro dispositivos GPU, para la conexión a nodos HPC con alimentación y refrigeración adecuadas para su instalación interna ayudando a resolver uno de lo mayores problemas de la HPC, la disipación de calor.

Tras su lanzamiento reciente el 4 de Febrero del presente año GeForce Now ya parece posicionarse en el top de plataformas que prestan servicio de streaming de videojuegos en la nube, a pesar de competir con servicios como Google Stadia, PlayStation Now y Microsoft xCloud, el motivo parece ser su calidad gráfica y baja latencia, pues al estar conformado por una red de servidores basados en centros de datos que alojan y sirven la biblioteca de juegos, la plataforma puede transmitir juegos a una resolución de hasta 1080p a 60 cuadros por segundo, resultado del aprovechamiento del gran paralelismo, y el alto ancho de banda de la memoria en las GPU para trabajos de alto costo aritmético en contra de realizar numerosos accesos a memoria principal, lo cual podría causar cuello de botella.

VIII. Conclusiones

NVIDIA sigue siendo líder en el mercado del GPU computing, ahora con su novedosa serie RTX, sigue mostrando cada vez más su posición dominante. El futuro del Ray Tracing depende de las compañías y de los usuarios en torno a la decisión de implementarla, no obstante, todo apunta a que será una tecnología bastante imponente en el futuro, ya que es un paso importante para llegar al hiperrealismo. Por otro lado, la simulación se encuentra en constante desarrollo y mejoramiento, cada vez hay computadores más rápidos, GPUs más potentes, y máquinas más eficientes para poder procesar la gran cantidad de datos necesarios para generarla, por lo tanto, se espera que en el futuro la simulación sea cada vez más precisas, y cada vez más rápida a la hora de desarrollarla. De esta misma manera evoluciona la inteligencia artificial y el deep learning, gracias a la entrada en el mercado de hardware cada vez más potente, sin embargo, ya se está viviendo un gran incremento de las aplicaciones que hacen uso de DL, se prevé que estas crezcan mucho más y superen a sus predecesoras con la salida de nuevas GPUs enfocadas para esta rama de la inteligencia artificial. Este amplio abanico de posibilidades que la empresa ha desarrollado a lo largo de los años al no haberse limitado, a diferencia de algunos de sus competidores a producir hardware, le ha valido a NVIDIA una parte importante en varios medios de servicio masivo, los cuales soporta en equipos de la más alta capacidad, implementar servicios de cloud usando sus propias tecnologías les ha valido la obtención de una configuración y optimización superior a la de sus pares, debido a la ventaja de tener conocimiento claro sobre la pantente de sus dispositivos.

Referencias

[1] "What is nvidia?" March 2020. [Online]. Available: https://www.cbronline.com/what-is/what-is-nvidia-4956361/

[2] "How nasa is using gpus to visualize the mars landing — nvidia blog. (2019)." [Online]. Available: https://blogs.nvidia.com/blog/2019/11/18/nasa-mars-landing-simulation-gpus/

[3] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackerman *et al.*, "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688–702, 2020.

[4] L. Seiler, "A hardware assisted design rule check architecture," in *19th Design Automation Conference*. IEEE, 1982, pp. 232–238.

[5] A. Appel, "Some techniques for shading machine renderings of solids," in *Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference*, ser. AFIPS '68 (Spring). New York, NY, USA: Association for Computing Machinery, 1968, p. 37–45. [Online]. Available: https://doi.org/10.1145/1468075.1468082

[6] T. Whitted, "An improved illumination model for shaded display," *Commun. ACM*, vol. 23, no. 6, p. 343–349, Jun. 1980. [Online]. Available: https://doi.org/10.1145/358876.358882

[7] "Nvidia's history." [Online]. Available: https://www.nvidia.com/en-us/about-nvidia/corporate-timeline/

[8] J. McCarthy, "What is ai?" Jan 1970. [Online]. Available: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

[9] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[10] "Deep learning: Overview of neurons and activation functions," https://medium.com/@srnghn/deep-learning-overview-of-neurons-and-activation-functions-1d98286cf1e4, (Accessed on 02/29/2020).

[11] S. Ray, "History of ai," Dec 2019. [Online]. Available: https://towardsdatascience.com/history-of-ai-484a86fc16ef

[12] "Wabot." [Online]. Available: http://www.humanoid.waseda.ac.jp/booklet/kato$_2$.$html$

[13] IBM, "Ibm100 - deep blue," https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/, (Accessed on 02/29/2020).

[14] "Autopilot — tesla," https://www.tesla.com/autopilot?redirect=no, (Accessed on 03/01/2020).

[15] T. D. D. Consulting, "The potential for artificial intelligence in healthcare," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/, (Accessed on 03/01/2020).

[16] Z. Alyafeai and L. Ghouti, "A fully-automated deep learning pipeline for cervical cancer classification," *Expert Systems with Applications*, vol. 141, p. 112951, 2020.

[17] Y. Song, L. Zhang, S. Chen, D. Ni, B. Li, Y. Zhou, B. Lei, and T. Wang, "A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 2903–2906.

[18] "Nvidia announces the geforce rtx 20 series: Rtx 2080 ti 2080 on sept. 20th, rtx 2070 in october."

[19] J. Burgess, "Rtx on–the nvidia turing gpu," in *2019 IEEE Hot Chips 31 Symposium (HCS)*. IEEE, 2019, pp. 1–27.

[20] "Ray tracing, respuestas a tus preguntas: Tipos de ray tracing."

[21] G. Hwang and S. Chang, "Speed up weather prediction on qct developer cloud: A case study on knights landing platform," in *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, June 2017, pp. 6–9.

[22] Y. Lin, C. Lin, C. Lee, and Y. Chung, "qcuda: Gpgpu virtualization for high bandwidth efficiency," in *2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, Dec 2019, pp. 95–102.