Trabajo de investigación N: 2
TPU Tensor Processing Unit

Aguirre Prieto Ángel Ernesto, Castro Calvopiña Bryan Paúl y **Ramos Espinosa Christopher Lizardo**

Resumen –en el presente proyecto el cual lo hemos realizado por varias partes se ha realizado una investigación a fondo sobre las TPU (Unidad de Procesamiento Tensorial) y la plataforma presentada por google llamada "Google Colab" que durante el desarrollo de la investigación se demostró el uso de cada uno de los electos de análisis además se mostrara mediante la implementación de un ejemplo el uso de una TPU y de la Plataforma Google Colab.

Índice de Términos – Tensor Flow: as una biblioteca de código abierto para aprendizaje automático a través de un rango de tareas, y desarrollado por Google para necesidades de sistemas capaces de construir y entrenar redes neuronales para detectar y descifrar patrones y correlaciones, análogos al aprendizaje. **TPU:** Tensor Processing Unit. **IA:** Inteligancia artificial**. GPU**: graphics processing unit. **NVIDIA**: es una empresa multinacional especializada en el desarrollo de unidades de procesamiento gráfico y tecnologías de circuitos integrados

Abstract – In this project, which we have carried out by several parties, an in-depth investigation has been carried out on TPUs (Tensor Processing Unit) and the platform presented by Google called "Google Colab", which during the development of the investigation demonstrated the use of each of the elements of analysis, the use of a TPU and the Google Colab Platform will also be shown through the implementation of an example..

Keywords- Tensor Flow: is an open source library for machine learning across a range of tasks, and developed by Google for systems needs capable of building and training neural networks to detect and decipher patterns and correlations, analogous to learning. **TPU**: Tensor Processing Unit. **IA**: Artificial intelligence. **GPU**: graphics processing unit. **NVIDIA**: is a multinational company specialized in the development of graphics processing units and integrated circuit technologies

.

I. INTRODUCCIÓN

El Internet cada vez se genera un máximo de información, y por tanto más crucero. Empresas como Google se llevan gran noticiero de él porque el buscador es la web más utilizada de todo el dirigible, así como sus úrico son utilizamos por miles de millones de usuarios. Es por ello que en Google necesitan cientos de miles de servidores para procesar este informe, y cuanto más potentes y eficientes sean estos procesadores, más capaz será la causa. Hace unos primaveras Google se enfrentó a un dificultad relacionado con su adlátere de voz. Si todos sus usuarios lo usaran durante tres minutos al día, habrían tenido que duplicar el signo de servidores para diligenciar todo el sistema de machine laringe que se utiliza para transformar la voz en texto. En zona de agenciarse nuevos servidores, la monstruo decidió generar un hardware dedicado especialmente a estas tareas. El resultado de esta decisión fue la aparición del Tensor Procesan Unir (TPU). En oriente signo está minucioso las ventajas de provecho porte a procesadores y tarjetas gráficas normales que se utilizan para ese don nadie de funciones. En obvio, compara baza el rendimiento zote de ambas configuraciones como el fruto por vatio. Al ser evaluado y comparada esta novedad tecnología, se logra ver que traerá grandes resultados para el procesamiento de datos y la computación a gran escalera. Sin bloqueo, algunas empresas como Nidia ve esta nueva tecnología como una nueva oportunidad de lucha de cobrar para nuevas tecnologías.

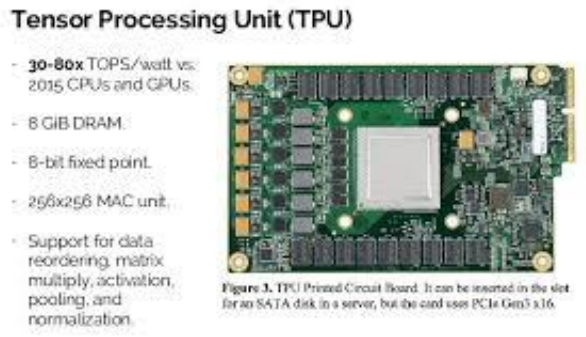

Figure 1. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

II. DEFINICIÓN DE OBJETIVOS NECESARIOS PARA LA INVESTIGACIÓN

A. Objetivo general

Para definir el objetivo general se tiene en cuenta el problema a estudiar y la búsqueda del producto solicitado después de realizar este proceso se logró definir el objetivo principal el cual es:

Conocer las cualidades y funcionalidades de las TPU y como verlas en funcionamiento mediante el uso de la plataforma Google Colab.

B. Objetivos específicos

Cuando hablamos de definir los objetivos específicos tenemos que pensar que los mismos se descubren cuando se empiezan a hacer las respectivas investigaciones sobre el objetivo general mediante las cuales encontramos los siguientes objetivos específicos:

- Conocer las diferencias de funcionamiento de las TPU con las CPU
- Conocer las diferencias de funcionamiento de las TPU con las GPU.
- Implementar una n ejemplo funcional para el uso de las TPU.

II MARCO TEORICO

A. TPU- Undead de Procesamiento Tensor

Las unidades de procesamiento de tensor son circuitos integrados desarrollados específicamente para la educación de máquinas.

En representación con las unidades de procesamiento manifiesto (que a dividir de 2016 se usan con frecuencia para las mismas tareas), estas unidades están diseñadas explícitamente para un mayor volumen de litiasis de explicación corta y carecen de hardware para la rasterización/mapeo de textura. El definición ha sido acuñado para un microprocesador peculiar diseñado para el puerta Tensarlo de Google.

Otros diseños de aceleradores de IA están apareciendo asimismo en otros proveedores y están dirigidos a mercados de robótica e incrustados.

PRIMERA GENERACIÓN

La primer generación del TPU de Google y se presenta en el Google I/O del 2016 diseñado específicamente para recostar la empecinamiento de redes neuronales entrenadas. 7 Estés TPU tienen fuera de precisión en representación con las CPU o GPU normales y una específico alcanzado por operaciones matriciales.

El TPU es una matriz de 8 bits multiplique el motor, impulsado con instrucciones CISC por el procesador host a través de un bus PCIe 3.0. Se fabrica en un proceso de 28 nm con un tamaño de troquel ≤331 mm2. La velocidad del reloj es de 700 MHz y tiene una potencia de diseño térmico de 28-40 W. Tiene 28MiB de memoria en chip y 4 MiB de 32 bits
Acumuladores tomando los resultados de una matriz 256x256 de multiplicadores de 8 bits. Las instrucciones transfieren datos a o desde el huésped, realizan multiplicaciones de matriz o consolaciones, y aplican las funciones de activación

SEGUNDA GENERACIÓN

La segunda generación de TPU de Google fue presentado en Google I/O del 2017. Esto no sólo va a acelerar la aplicación de redes neuronales (inferencia), sino también la formación de estas redes. Estas TPU tienen una potencia de procesamiento de 180 TFLOPS y están interconectados a un "pod" con 11,5 petaflops. La Topología de la arquitectura del sistema de grupos tiene esferas en forma de red de 8 × 8 TPUs.

El TPU de segunda generación forman parte del Google Compute Engine, una oferta en la nube de Google, utilizable.
Los detalles técnicos de la segunda generación actualmente (mayo de 2017) no están disponibles. Sin embargo, se supone que utiliza GDDR5 SRAM.

B. NUEVA TECNOLOGIA TPU GOOGLE

El resultado se llama una Unidad de Procesamiento Tensor (TPU), un ASIC a medida que construimos específicamente para el aprendizaje de máquina - y adaptado para TensorFlow.

Hemos estado corriendo TPU dentro de nuestros centros de datos durante más de un año, y las hemos encontrado para entregar un orden de magnitud mejor rendimiento optimizado
Por vatio de aprendizaje automático.

TPU está adaptado a aplicaciones de aprendizaje automático, permitiendo que el chip sea más Tolerante a la reducida precisión de cálculo, lo que Significa que requiere un menor número de Transistores por operación. Debido a esto, se puede Exprimir más operaciones por segundo en el silicio,

Utilizar modelos de aprendizaje automático más Sofisticadas y potentes y aplicar estos modelos con Mayor rapidez, por lo que los usuarios obtienen Resultados más inteligentes con mayor rapidez. Un Tablero con un TPU encaja en una ranura de la unidad De disco duro en nuestros bastidores de centros de Datos

VENTAJAS DEL USO DE LAS TPU:

Las TPU minimizan el tiempo necesario para alcanzar la exactitud cuándo entrenas modelos de redes neuronales complejas y extensas. Un modelo que antes tardaba semanas en entrenarse con otras plataformas de hardware ahora puede converger con las TPU en cuestión de horas.

C. MODELOS DE PROGRAMACION DE LAS TPU Y SUS CARACTERISTICAS

Las TPU realizan cálculos de vector y matrices densos muy rápidamente. La transferencia de datos entre TPU y la memoria del host es lenta en comparación con la velocidad del procesamiento. Esto se debe a que la velocidad del bus PCIe es mucho menor que la de la interconexión de TPU y la de la memoria de gran ancho de banda (HBM) en el chip. Esto significa que una compilación parcial de un modelo, en el que la ejecución "va y viene" entre el host y el dispositivo, usa este último de forma muy poco eficiente, ya que permanecería inactivo la mayor parte del tiempo, a la espera de que lleguen los datos por medio del bus PCIe. A fin de solucionar este problema, el modelo de programación de TPU está diseñado para ejecutar gran parte del entrenamiento en la TPU, idealmente el ciclo de entrenamiento completo.

- Se guardan todos los parámetros del modelo en la memoria de gran ancho de banda en el chip.
- La ejecución de muchos pasos de entrenamiento en un mismo ciclo permite amortizar el costo de iniciar cálculos en TPU.
- Se transmiten los datos de entrenamiento de entrada a una cola de entrada en TPU. Un programa que se ejecuta en TPU obtiene los lotes desde esas colas durante cada paso de entrenamiento.
- El servidor de TensorFlow que se ejecuta en la máquina host (la CPU adjunta al dispositivo de TPU) obtiene los datos y realiza procesamientos previos antes de ingresarlos en el hardware de TPU.
- Paralelismo de datos: Los núcleos de TPU ejecutan un programa idéntico que reside en sus respectivas HBM de manera síncrona. Se realiza una operación de reducción al final de cada paso de red neuronal en todos los núcleos.

D. RESULTADOS TPU

Esta primera generación de TPU dirigido inferencia (el uso de un modelo ya formado, en contraposición a la fase de formación de un modelo, que tiene características algo diferentes), y aquí están algunos de los resultados que hemos visto: En nuestras cargas de trabajo de producción de IA que utilizan la inferencia de redes neuronales, el TPU es 15x a 30x más rápido que las GPU y CPU contemporáneos.

El TPU también logra mucho mejor eficiencia energética que los chips convencionales, el logro de 30x a 80x mejora en TOPS / Watt medida (tera-operaciones [billón o 10 12 operaciones] de cálculo por vatio de energía consumida). Las redes neuronales que impulsan estas aplicaciones requieren una cantidad sorprendentemente pequeña de código: sólo 100 a 1500 líneas. El código se basa en TensorFlow , nuestro marco de código abierto de aprendizaje automático popular.

III. PROCESO DE INVESTIGACION SOBRE LOS ELEMENTOS QUE COMPONEN LAS TARJETAS DE DESARROLLO

A. PLANIFICACION Y CRONOGRAMA DE TRABAJO

En este paso se dividió la investigación en partes para cada uno de los de los integrantes del equipo de realización del trabajo, después de haber sido realizadas las investigaciones por separado se hizo uso de herramientas virtuales para explicar entre los miembros las respectivas partes investigadas por cada integrante además de hacer un cronograma para el resto de los pasos a realizar para cumplir con los objetivos del proyecto.

Cronograma

ID	TAREA	16/7/2020	17/7/2020	18/7/2020	19/7/2020	20/7/2020	21/7/2020	22/7/2020	23/7/2020
1	DIVISION TRABAJO								
2	INFORME								
3	ARTICULO Y DIAPOSITIVAS								
4	REVISION DE ERRORES								
5	REALIZACION DEL VIDEO								

B. Articulo

En este paso se encuentra el juntar la información que antes se dividió con sus respectivas investigaciones para así conocer cada una de las características de cada una de la tarjetas de desarrollo además de conocer de manera teórica la función de cada uno de los pines y la razón por la cual se encuentran ubicados en las tarjetas de desarrollo.

En este paso se puede realizar varios de los primeros pasos sobre el artículo y el proyecto además de depurar los conocimientos y aumentarlos debido a la investigación a fondo realizada para cada uno de los elementos de las tarjetas.

C. Planteamiento de problema y objetivos

El planteamiento de los objetivos es esencial debido a que gracias a estos se puede llevar acabo la realización del proyecto por tal motivo se definió los siguientes objetivos generales:

- Conocer las cualidades y funcionalidades de las TPU y como verlas en funcionamiento mediante el uso de la plataforma Google Colab.

Después de definir los objetivos generales se empezó a realizar las investigaciones respectivas y a raíz de so surgieron los objetivos específicos que son los siguientes

- Conocer las diferencias de funcionamiento de las TPU con las CPU
- Conocer las diferencias de funcionamiento de las TPU con las GPU.
- Implementar una n ejemplo funcional para el uso de las TPU.

D. Planteamiento del estado del arte y el marco teórico

En este paso se realizó una investigación a fondo y en el marco teórico se colocó parte de los archivos más actuales de las cuales nos proporcionaron una mejor visión del panorama general además de proporcionarnos varias acotaciones a nuestro conocimiento.

E. MAPA DE VARAIBLES Y PREREQUISITOS

En esta parte de la investigación encontramos que las tarjetas de desarrollo tienen variaciones en específico la RASPBERRY tiene sus variaciones las cuales tienen cambios en sus pines y en la capacidad de procesamiento que tiene, otra variación que se encuentra es que algunas de ellas hacen uso e lenguajes de programación con algunas variaciones uno de estos casos es el de ARDUINO ya que hace uso del C++ modificado (simplificado).

F. DESARROLLO DE LOS EJEMPLOS DER CADA UNA DE LAS TARJETAS DE DESARROLLO

Ejemplo

1- Acceder a la página dependiente de google (google colab) y crear un nuevo block de notas o atreves del drive ya que es un complemento del mismo.
https://colab.research.google.com/

2- Escribir el código en lenguaje de programación adecuado (Python) además de elegir sen que acelerador queremos que se ejecute en este caso se lo realizó en varios de los aceleradores como son la CPU,GPU y TPU

3- Ejecución del programa escrito en código, además tomar el tiempo que se demora en ejecutarlo de manera correcta y realizar la respectiva comparación de los tiempos de retardo que tienen cada uno de los aceleradores

IV. CONCLUSIONES

Realizada la investigación sobre las tarjetas de desarrollo se llegó a las conclusiones sobre los objetivos específicos las cuales fueron:

- Las TPU son dispositivos que realizan con una mayor velocidad ya que la misma trabaja con matrices y con bloques de código que en comparación de las CPU es un 30% a 40 % más veloz y exacta en los cálculos

- Las TPU son dispositivos que realizan con una mayor velocidad ya que la misma trabaja con matrices y con bloques de código que en comparación de las GPU es un 15% a 20 % más veloz y exacta en los cálculos.

- EL ejemplo funcional se especificó en pasos anteriores además de mostrar la diferencia de tiempos en la implementación del código en el CPU, GPU, TPU.

Las conclusiones planteadas son de cada objetivo específico respectivamente los cuales ya fueron mostrados en un paso anterior.

Después de llegar a la conclusión de cada uno de los objetivos específicos se tiene que llegar a la evolución final del objetivo principal, que después de la evaluación se llegó a su respectiva conclusión, la cual es:

- Durante el desarrollo de los objetivos específicos se llegó a la conclusión que el objetivo general se cumplió a cabalidad ya w que gracias a la implementación del ejemplo mostrado anteriormente se mostró la diferencia en la velocidad de procesamiento que sus contrapartes como son las CPU y las GPU.

V. RECOMENDACIONES

- Al momento de realizar la investigación nos mostró que se deben estudiar más afondo de la tecnología TPU ya que la misma al ser un concepto reciente no existe una información muy amplia.
.
- Conocer los distintos tipos de plataformas elegibles para realizar simulaciones sobre los diferentes elementos estudiados en este artículo.

- Tener en cuenta al momento de realizar cualquier simulación en las TPU, GPU y las CPU se deben tomar en cuenta los lenguajes de programación elegidos para el componente en el que se necesita simularlo...

VI. REFERENCIAS

[1] https://www.adslzone.net/2017/04/06/tpu-el-Chip-de-google-hasta-30-veces-más-potente-Que-una-CPU-y-ju-normales/J. Clerk Maxwell, A Tratéis en Electricista and Magnetismo, 3rd ed., Vol. 2. Oxford: Clareando, 1892, pp.68-73.

[2] I.S. Jacob and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] K. Elissa, “Title of paper if known,” unpublished.

[4] https://es.wikipedia.org/wiki/Unidad_central_de_procesamiento

[5] https://www.definicionabc.com/tecnologia/cpu.php

[6] http://www.valortop.com/blog/que-es-la-cpu-o-procesador-de-un-ordenador

[7] https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

[8]https://es.wikipedia.org/wiki/Unidad_de_procesamiento_de_tensor

[9] https://www.adslzone.net/2017/04/06/tpu-el-chip-de-google-hasta-30-veces-mas-potente-que-una-cpu-y-gpu-normales/

[10] https://www.blog.google/topics/google-cloud/google-cloud-offer-tpus-machine-learning/

[11] https://www.nextplatform.com/2017/05/17/first-depth-look-googles-new-second-generation-tpu/

[12] http://www.informatica-hoy.com.ar/aprender-informatica/Diferencias-CPU-GPU-APU.php

[13] https://www.hpcwire.com/2017/04/10/nvidia-responds-google-tpu-benchmarking/

[14] http://www.forosdeelectronica.com/f37/procesadores-dedicados-redes-neurologicas-procesamiento-tensores-152755/