Unidad de Procesamiento Tensor de Google
Tensor Processing Unit

TPU de Google

Yeison David López Camacho
Estudiante pregrado
Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
yeisonlo@hotmail.com

Yesid Alfonso Gutiérrez Guate
Estudiante pregrado
Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
yesidgutierrez.08@gmail.com

***Abstract*—This document analyzes Google's new technology implemented in its servers. This work describes and shows up the differences between similar arquitectures like CPU and GPU, further we compare them about their performance, energy cost and the use of those technologies. Beginning with the TPU's analysis, we found that it is an arquitecture that will bring big changes in the technology.**

***Keywords* - TPU; GPU; Tensor Processing Unit; Nvidia;**

Resumen – Este documento analiza la nueva tecnología implementada por Google (TPU) en sus servidores. Este trabajo describe y muestra las diferencias de varias arquitecturas similares como las CPU y GPU, además se comparan respecto a su rendimiento, costo energético y el uso de estas tecnologías. A partir del análisis de las TPU se encontró que es una arquitectura que traerá grandes cambios en la tecnología.

Palabras clave—Unidad de Procesamiento; procesamiento google; TPU; GPU;

I. INTRODUCCION

En Internet cada vez se genera mayor información, y por tanto más tráfico. Empresas como Google se llevan gran parte de él porque el buscador es la web más utilizada de todo el mundo, así como sus servicios son utilizamos por miles de millones de usuarios. Es por ello que en Google necesitan cientos de miles de servidores para procesar esta información, y cuanto más potentes y eficientes sean estos procesadores, más eficiente será el proceso.

Hace unos años Google se enfrentó a un problema relacionado con su asistente de voz. Si todos sus usuarios lo usaran durante tres minutos al día, habrían tenido que duplicar el número de servidores para gestionar todo el sistema de machine learning que se utiliza para transformar la voz en texto. En lugar de comprar nuevos servidores, la compañía decidió crear un hardware dedicado especialmente a estas tareas.

El resultado de esta decisión fue la llegada del Tensor Processing Unit (TPU). En este documento está detallado las ventajas de rendimiento frente a procesadores y tarjetas gráficas normales que se utilizan para ese tipo de funciones. En concreto, compara tanto el rendimiento bruto de ambas configuraciones como el rendimiento por vatio.

Al ser evaluado y comparada esta nueva tecnología, se logra ver que traerá grandes resultados para el procesamiento de datos y la computación a gran escala. Sin embargo, algunas empresas como Nvidia ve esta nueva tecnología como un nuevo reto de superar para nuevas tecnologías.

II. MARCO TEORICO

A. CPU-Unidad de procesamiento central

La CPU es un dispositivo de hardware que interpreta las instrucciones de cualquier programa

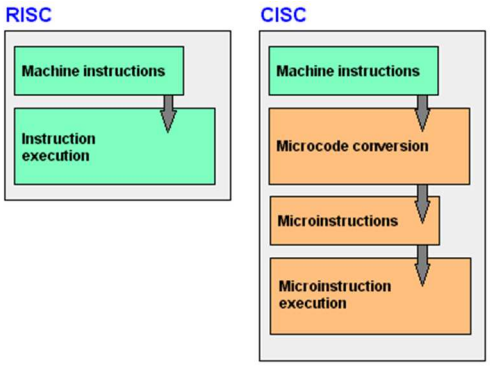
realizando operaciones aritméticas, lógicas y de entrada y salida, una CPU se compone principalmente de:

- ✓ La unidad aritmético-lógica (ALU): se encarga de realizar todas las operaciones aritméticas y lógicas.
- ✓ La unidad de control (CU): se encarga de extraer las instrucciones de memoria, decodificarlas y ejecutarlas realizando llamados a la ALU en caso de que lo requiera.
- ✓ Bus de datos: permite el paso de datos entre los periféricos, la memoria y el procesador.
- ✓ Registro de instrucción: Almacena la instrucción que se está ejecutando en el momento.
- ✓ Archivo de registros: Almacena datos temporalmente.
- ✓ Registro de direcciones de memoria: añade la siguiente dirección de memoria a la que se va a leer o escribir un dato.
- ✓ Contador de programa: contiene la dirección de memoria de la siguiente instrucción que va a ejecutarse de un programa.
- ✓ Reloj: Señal sincrónica a la que trabaja la CPU.

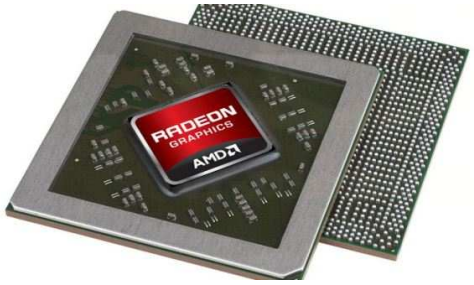
Así mismo la CPU puede clasificarse en dos grupos dependiendo del conjunto de instrucciones que admita:

- Procesadores CISC (Complex Instruction Set Computers). Su entramado de instrucciones es extenso y complejo, ya que éstas operan sobre los elementos internos de la computadora y son ejecutadas por un microprograma.

- Procesadores RISC (Reduced Instruction Set Computers). Poseen un juego de instrucciones reducido, pues cada una de ellas desarrolla una tarea sencilla. En caso de tener directrices más complejas, se llevan a cabo mediante una secuencia de las instrucciones sencillas disponibles.



B. GPU- Unidad de procesamiento gráfico



Es un coprocesador dedicado al procesamiento de gráficos u operaciones de coma flotante, para aligerar la carga de trabajo del procesador central en aplicaciones como los videojuegos o aplicaciones 3D interactivas. De esta forma, mientras gran parte de lo relacionado con los gráficos se procesa en la GPU, la unidad central de procesamiento (CPU) puede dedicarse a otro tipo de cálculos. La GPU implementa ciertas operaciones gráficas llamadas primitivas optimizadas para el procesamiento gráfico. Una de las primitivas más comunes para el procesamiento gráfico en 3D es el antialiasing, que suaviza los bordes de las figuras para darles un aspecto más realista. Adicionalmente existen primitivas para dibujar rectángulos, triángulos, círculos y arcos. Las GPU actualmente disponen de gran cantidad de primitivas, buscando mayor realismo en los efectos.

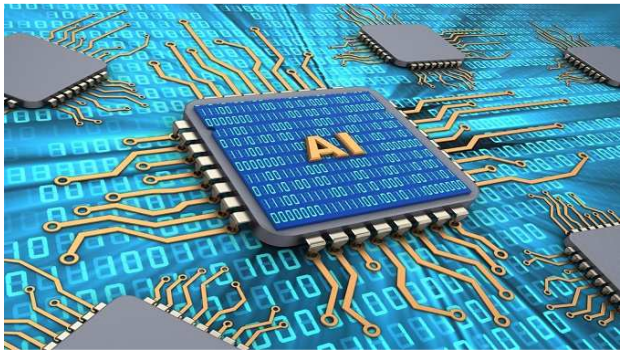
Las GPU están presentes en las tarjetas gráficas.

A diferencia de los procesadores centrales, diseñados con pocos núcleos, pero altas frecuencias de reloj, las GPU suelen tener grandes cantidades de núcleos de procesamiento a frecuencias de reloj relativamente bajas. En la actualidad, la mayoría de los núcleos de procesamiento están dirigidos a dos funciones: procesamiento de vértices y de píxeles.

El procesamiento de vértices es relativamente sencillo para las unidades de procesamiento gráfico modernas, siendo de los que menos recursos consumen. En términos sencillos se trata de obtener la información de los vértices, previamente calculada por el CPU, y procesar su ordenamiento espacial, rotación, y qué segmento del vértice será gráficamente visible, para así continuar con el pixelado. Luego, se procede a procesar los pixeles. Éste es el proceso más complejo y que requiera más carga de procesamiento, pues se aplicaran todas las capas y efectos necesarios para crear texturas complejas y obtener gráficos lo más realistas posibles.

Por último, una vez procesada la información gráfica, esta es llevada a un monitor digital o analógico (en este último caso, previo paso por un convertidor), según las necesidades propias del ordenador.

C. TPU- Unidad de Procesamiento Tensor

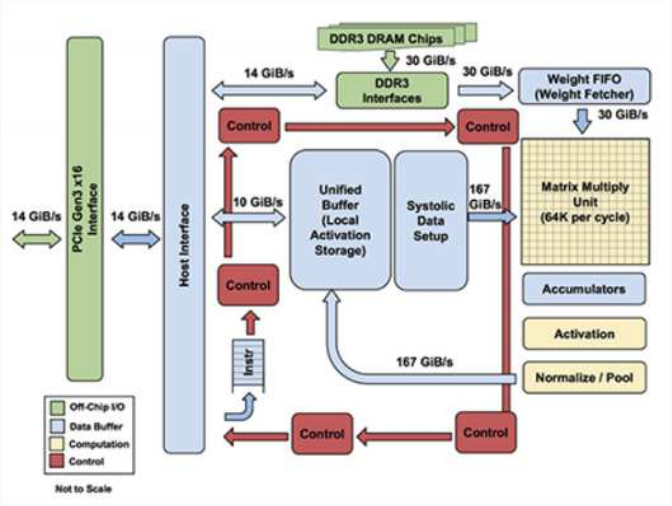


Las unidades de procesamiento de tensor son circuitos integrados desarrollados específicamente para el aprendizaje de máquinas. En comparación con las unidades de procesamiento gráfico (que a partir de 2016 se usan con frecuencia para las mismas tareas), estas unidades están diseñadas explícitamente para un mayor volumen de cálculo de precisión reducida y carecen de hardware para la rasterización/mapeo de textura. El término ha sido acuñado para un chip específico diseñado para el marco TensorFlow de Google. Otros diseños de aceleradores de IA están apareciendo también en otros proveedores y están dirigidos a mercados de robótica e incrustados.

PRIMERA GENERACIÓN

La primera generación del TPU de Google y se presentó en el Google I/O del 2016 diseñado específicamente para apoyar la aplicación de redes neuronales entrenadas. 7 Estas TPU tienen menos precisión en comparación con las CPU o GPU normales y una especialización alcanzado por operaciones matriciales.

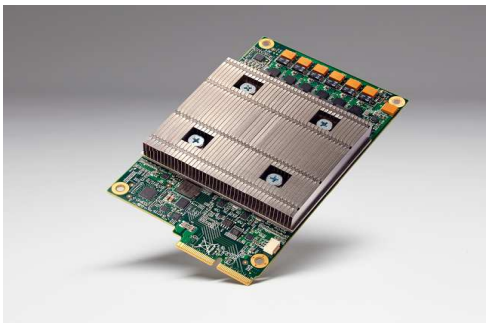
El TPU es una matriz de 8 bits multiplique el motor, impulsado con instrucciones CISC por el procesador host a través de un bus PCIe 3.0. Se fabrica en un proceso de 28 nm con un tamaño de troquel ≤ 331 mm2. La velocidad del reloj es de 700 MHz y tiene una potencia de diseño térmico de 28-40 W. Tiene 28 MiB 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 convoluciones, y aplican las funciones de activación 8



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.

III. 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. Esto es aproximadamente equivalente a la tecnología de avance rápido cerca de siete años en el futuro (tres generaciones de la Ley de Moore).

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.

El TPU ASIC se basa en un proceso de 28nm, funciona a 700 MHz y consume 40W cuando se ejecuta. Porque teníamos que desplegar el TPU a los servidores existentes de Google lo más rápido posible, elegimos para empaquetar el procesador como una tarjeta aceleradora externa que encaja en una ranura de disco duro SATA para la instalación drop-in. El TPU está conectado a su huésped a través de un bus x16 PCIe Gen3 que proporciona 12.5GB / s de ancho de banda efectivo.

Mientras que otras compañías están discutiendo acerca de si las GPU, FPGAs, o VPU son más adecuados para el aprendizaje automático, Google salió con la noticia de que se ha estado utilizando su propia hecha a la medida Unidad de Procesamiento Tensor (TPU) durante más de un año, alcanzando un afirmaban aumento de 10 veces en la eficiencia. La comparación se hizo probable en relación con las GPU, que son actualmente los chips estándar de la industria para el aprendizaje automático.

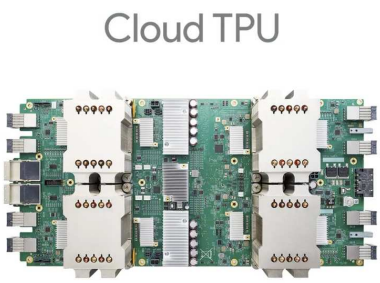
Análisis tensorial es una extensión del cálculo vectorial, que está en la base de Google (recientemente liberado como código abierto) Tensorflow marco para el aprendizaje de la máquina.

Las nuevas unidades de procesamiento Tensor, como era de esperar, están diseñadas específicamente para hacer cálculos sólo tensores, lo que significa que la empresa puede colocar más transistores en el chip que hacer bien una sola cosa - el logro de una mayor eficiencia que otros tipos de chips.

Esta clase de fichas se llama ASIC (circuitos integrados de aplicación específica), y que han sido utilizados, por ejemplo, en módems inalámbricos, tales como módem Icera de Nvidia, y en Bitcoin equipos de minería para órdenes de magnitud mayor eficiencia en comparación con las GPU y FPGAs .

Movidius dijo que la filosofía de TPU de Google es más en línea con lo que ha estado tratando de lograr con su 2 Myriad unidad de procesamiento visual (VPU) exprimiendo más ops / W de lo que es posible con las GPU.

Google afirmó en un blog que su TPU podría lograr un orden de magnitud mayor eficiencia para el aprendizaje de la máquina, lo que sería equivalente a cerca de siete años de progreso después de la Ley de Moore. Google dijo que una junta de TPU podría caber dentro de una ranura de la unidad de disco duro en sus centros de datos.
Google también dio a conocer por primera vez que los TPU se utiliza no sólo para impulsar su producto Street View, sino también AlphaGo en su Ir coincide en contra de Lee Sedol. Cuando la empresa anteriormente habló de AlphaGo, sólo se menciona el uso de CPU y GPU, aunque eso fue meses antes de las coincidencias Go sucedieron. Google dijo que el TPU permitió AlphaGo a “pensar” mucho más rápido y permite que se vea más adelante entre los movimientos.



Google ahora puede utilizar TPU no sólo para mejorar sus propios productos, pero también puede ofrecer ese incremento en el rendimiento a los clientes aprendizaje automático. La compañía ahora puede parecer mejor la competencia en este tipo de mercado, ya que otros pueden todavía sólo ofrecer servicios de aprendizaje automático quizá basados en FPGA o basada en la GPU.

ASIC, en comparación con FPGAs, son codificados de forma rígida y no pueden ser reprogramadas en el campo. Esta falta de flexibilidad restringe muchas de emplear los procesadores especializados en despliegues a gran escala. Sin embargo, Google indicó que saltó de la primera prueba de silicio a un entorno de producción en tan sólo 22 días. Este increíble rampa indica que Google tiene la capacidad de desarrollar e implementar otros ASIC optimizados en una línea de tiempo acelerado en el futuro.

TPU de Google ahora podría cambiar el panorama para el aprendizaje de máquina, a medida que más empresas pueden estar interesadas en seguir el mismo camino para lograr el mismo tipo de ganancias de rendimiento y eficiencia. Google

también ha estado coqueteando con los ordenadores cuánticos , así como el OpenPower y RISC-V arquitecturas de chip.

Los poderes de la familia Intel Xeon 99 por ciento de los centros de datos del mundo, y hay rumores de larga duración que Google está desarrollando su propia CPU para romper el dominio de Intel. El TPU Google puede ser un precursor de aún más en el futuro. trabajo exploratorio bien publicitada de Google con otras plataformas de computación estimuló Intel Xeon para comenzar a ofrecer a medida para casos de uso específicos. Será interesante ver los esfuerzos futuros de la empresa en el diseño de sus propios chips, y el impacto en el mercado en general.

IV. RESULTADOS TPU

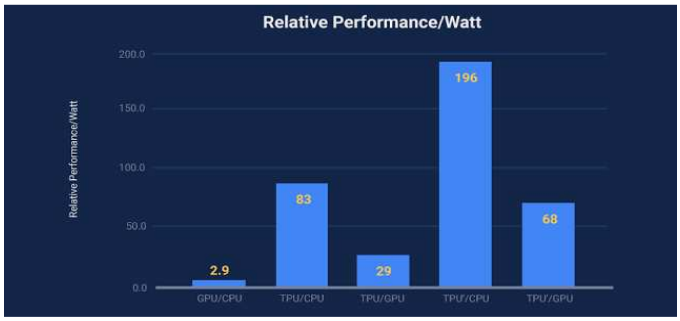
Esta primera generación de TPU dirigido inferencia (el uso de un modelo ya formado, en contraposición a la fase de formación de un modelo, que tiene características algo diferentes), y aquí están algunos de los resultados que hemos visto:

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.

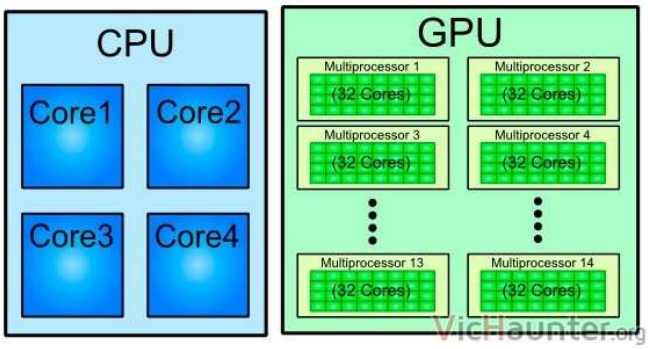
Más de 70 autores contribuyeron a este informe. Realmente se necesita una aldea para diseñar, verificar, implementar y desplegar el hardware y software de un sistema como este.



TPU nos permite hacer predicciones muy rápidamente, y lograr que los productos que responden en fracciones de segundo. TPU están detrás de cada consulta de búsqueda; que potencia exacta modelos de visión que subyacen en productos como Google Image Search, Google Fotos y la API de Google Cloud Vision; sustentan las mejoras de calidad innovadoras que Google Traducir puso en marcha el año pasado; y que jugaron un papel decisivo en la victoria de Google DeepMind sobre Lee Sedol , la primera instancia de un ordenador de derrotar a un campeón del mundo en el antiguo juego de Go.

V. DIFERENCIAS ENTRE CPU Y GPU

Estas son algunas diferencias básicas, que ayudaran a ver el avance tecnológico que se ha generado para la computación de gran escala y el procesamiento de datos.



- La CPU es procesador genérico y la GPU está especializada en representaciones gráficas.

- La velocidad de las GPU superan a las velocidades de la CPU.

- La GPU trabaja íntegramente en paralelo (se basa en el Modelo Circulante).

- La CPU puede remplazar una simple GPU (como los Intel i7) pero las GPU no pueden sustituir a las CPU.
- La ubicación: la CPU se sitúa en la placa base y la GPU va soldada en la circuitería de la representación gráfica.

VI. DIFERENCIAS ENTRE GPU Y CPU

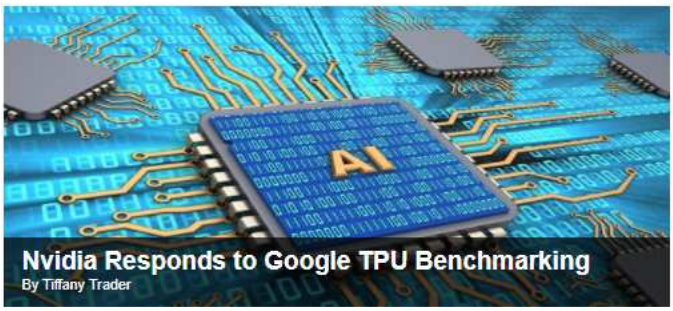
En septiembre de 2016, Google lanzó la GPU P40, basada en la arquitectura Pascal, para acelerar la inferencia de las cargas de trabajo de las aplicaciones modernas de IA, como la traducción de voz y el análisis de video. Google comparó el TPU con la antigua GPU K80 a finales del 2014, basados en la arquitectura Kepler. Nvidia creó el siguiente gráfico para "cuantificar el salto de rendimiento de K80 a P40 y mostrar cómo TPU se compara con la tecnología actual de NVIDIA. "

	K80 2012	TPU 2015	P40 2016
Inferences/Sec <10ms latency	1/13 TH	1X	2X
Training TOPS	6 FP32	NA	12 FP32
Inference TOPS	6 FP32	90 INT8	48 INT8
On-chip Memory	16MB	24 MB	11 MB
Power	300W	75W	250W
Bandwidth	320 GB/S	34 GB/S	350 GB/S

Basado en las especificaciones de TDP, el TPU es más eficiente que el P40 en una base de operaciones por vatio con un margen de 6,2X (para cargas de trabajo de inferencia de 8 bits).

Hay que tener en cuenta que el TPU sólo puede satisfacer las cargas de trabajo de inferencia. La fase de formación del aprendizaje profundo es mucho más complicada y las GPU tienen el liderazgo actualmente.

VII. RESPUESTA DE NVIDIA



Nvidia está eligiendo enmarcar los resultados recientes de TPU no como una potencial amenaza competitiva, sino como un claro signo de la ascendencia de la computación acelerada. "Sin la computación acelerada, la escala de salida de la IA simplemente no es práctico", es la conclusión que Nvidia plantea.

Dado que Google ha utilizado parte de la tecnología de Nvidia, puede evidenciarse que no existe un ámbito de competencia comercial entre ellos, sino un deseo de innovación propio de cada una de estas compañías basándose en la tecnología del otro, lo cual sería muy productivo para la computación misma, ya que la interacción de diversas tecnologías podría ayudar a entender, emprender e incluso desatar toda una serie nueva de ideas y propuestas a favor de la comunidad.
Entre estas compañías, que llevan enfoques diferentes, han concordado en algunos puntos acerca de los TPU que crean esa competencia de la que Nvidia habla:

- La AI requiere una computación acelerada. Los aceleradores proporcionan las demandas de procesamiento de datos significativas de aprendizaje profundo en una época en que la ley de Moore se está desacelerando.
- El procesamiento de los tensores es el núcleo de la entrega de rendimiento para la formación de aprendizaje profundo y la inferencia.
- El procesamiento de tensores es una nueva carga de trabajo importante que las empresas

deben tener en cuenta al construir modernos centros de datos.

- Acelerar el procesamiento de los tensores puede reducir drásticamente el costo de la construcción de centros de datos modernos.

VIII. CONCLUSION

El objetivo de esta nueva tecnología es ofrecer al mundo una potencia de cálculo acelerada a pedido, sin gastos de capital iniciales, ayudando así a que diferentes empresas puedan construir los mejores sistemas de aprendizaje de máquina usando el poder de los Cloud TPU de Google.

La respuesta de Nvidia es muy acertada porque la competencia entre dos empresas de alta gama de tecnología permite que los conocimientos surjan rápidamente y nuevas tecnologías aparezcan en el mercado. Al querer cada empresa estar en a la vanguardia en estas estructuras, cada uno se esforzará y hará aportes grandes a la computación acelerada, y al aprendizaje de las maquinas.

REFERENCES

[1] https://www.adslzone.net/2017/04/06/tpu-el-chip-de-google-hasta-30-veces-mas-potente-que-una-cpu-y-gpu-normales/J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] I.S. Jacobs 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/