Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

Arquitecturas de GPU:

Las Unidades de Procesamiento Gráfico **(GPU)** son procesadores especializados en realizar cálculos en paralelo a gran escala, lo que los hace ideales para tareas tales como: el procesamiento de gráficos, la inteligencia artificial, entre otros.

A diferencia de las Unidades de Procesamiento Central **(CPU)**, que están diseñadas y optimizadas para ejecutar las tareas de forma secuencial con unidades de procesamiento de alta velocidad, las GPU están diseñadas para manejar varias tareas en paralelo con un numero considerable de núcleos simples y eficientes. Las GPU mas actuales se basan en procesamiento paralelo masivo, permitiendo el cálculo simultáneo de millones de operaciones en flujos de datos grandes.

Componentes Principales de una GPU

- **Los núcleos:** son unidades de cómputo especializadas en operaciones de coma flotante y enteros. Dependiendo del fabricante, se les denomina de diferentes maneras: NVIDIA: Núcleos CUDA**,** AMD: Stream Processors. Intel: Xe Cores.

- **Multiprocesadores de Flujo (SM):** Las GPU se organiza en grupos de núcleos que son responsables de dividir y coordinar la ejecución de tareas. Cada SM contiene: Varios núcleos de cómputo, Memoria compartida. Unidades de control de flujo.

- **Memoria:** cuenta con varios niveles de memoria para gestionar los datos:

  Memoria Global (VRAM): Es la memoria de video, es un tipo de memoria RAM (memoria volátil) haciendo su rendimiento es muchísimo más rápido.

  Memoria Compartida: La memoria compartida de GPU es la . Es un recurso temporal que se ajusta automáticamente según las necesidades del hardware

  Registros: Memoria más rápida utilizada por los hilos de ejecución.

La velocidad y ancho de banda de la memoria juegan un papel fundamental en el rendimiento de una GPU.

2.4. Unidades de Textura y ROPs

- **Unidades de Textura (TMUs)**: Se encargan de aplicar efectos gráficos como filtros y mapas de texturas.
- **Render Output Units (ROPs)**: Procesan la salida final de la imagen en la memoria de video.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

3. Tipos de Arquitecturas de GPU por Fabricante

Cada fabricante tiene su propia arquitectura de GPU, optimizada para diferentes tareas. Veamos las principales:

3.1. Arquitectura NVIDIA

Las GPU de NVIDIA están basadas en **CUDA (Compute Unified Device Architecture)**, que permite ejecutar aplicaciones altamente paralelas.

Algunas de sus arquitecturas más destacadas incluyen:

- **Fermi (2010)**: Primera en introducir la memoria caché L1 y L2.
- **Kepler (2012)**: Mayor eficiencia energética.
- **Pascal (2016)**: Introducción de la memoria HBM2.
- **Volta (2017)**: Incorporación de Tensor Cores para IA.
- **Turing (2018)**: Introducción de RT Cores para trazado de rayos en tiempo real.
- **Ampere (2020)**: Mejoras en IA y gráficos en tiempo real.
- **Hopper (2022)**: Enfocada en IA y computación de alto rendimiento.

3.2. Arquitectura AMD

AMD desarrolla sus GPU bajo la tecnología **RDNA (Radeon DNA) y CDNA** (Compute DNA para computación avanzada).

Algunas arquitecturas destacadas son:

- **GCN (Graphics Core Next, 2012-2019)**: Base de sus GPU durante años.
- **RDNA (2019 - Actualidad)**: Enfoque en gaming y eficiencia energética.
- **CDNA (2020 - Actualidad)**: Diseñada para inteligencia artificial y supercomputación.

3.3. Arquitectura Intel

Intel también ha desarrollado sus propias GPU bajo la línea **Intel Xe**, con distintas variantes:

- **Intel Xe-LP**: GPU integrada en procesadores modernos.
- **Intel Xe-HPG**: Diseñada para gaming.
- **Intel Xe-HPC**: Para inteligencia artificial y computación de alto rendimiento.

4. GPU para Diferentes Aplicaciones

Las GPU no solo se usan para gráficos, sino también en diversas áreas:

4.1. Gaming

Las GPU en juegos permiten renderizar gráficos de alta calidad, con tecnologías como:

- **Ray Tracing** (Trazado de Rayos) para iluminación realista.
- **DLSS (Deep Learning Super Sampling)** de NVIDIA para mejorar rendimiento con IA.

- **FSR (FidelityFX Super Resolution)** de AMD para mejorar escalado de imágenes.

4.2. Inteligencia Artificial y Machine Learning

Las GPU son clave en **deep learning**, ya que permiten ejecutar redes neuronales de forma paralela. Tecnologías como **Tensor Cores** de NVIDIA aceleran estos cálculos.

4.3. Computación Científica

Usadas en simulaciones físicas, genómica y modelado de datos climáticos. Plataformas como **CUDA, OpenCL y ROCm** permiten aprovechar su potencia.

4.4. Renderizado y Edición de Video

Programas como **Blender, Adobe Premiere y DaVinci Resolve** aprovechan la aceleración por GPU para renderizar gráficos 3D y videos en alta resolución.

5. Comparación entre CPU y GPU

Característica	CPU	GPU
Núcleos	Pocos (4-64)	Miles
Velocidad	Alta en tareas secuenciales	Alta en tareas paralelas
Memoria Caché	Grande (L1, L2, L3)	Pequeña pero con alto ancho de banda
Uso	Sistemas operativos, aplicaciones generales	Gráficos, IA, simulaciones, minería de criptomonedas

Las GPU son más eficientes en tareas paralelas, mientras que las CPU son mejores en ejecución secuencial y tareas generales.

6. Conclusión

Las arquitecturas de GPU han evolucionado para abordar múltiples aplicaciones más allá de los gráficos. Desde el gaming hasta la inteligencia artificial y la computación científica, las GPU han cambiado la manera en que procesamos grandes volúmenes de datos. Con fabricantes como **NVIDIA, AMD e Intel** impulsando la innovación, el futuro de las GPU promete más potencia, eficiencia y versatilidad.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

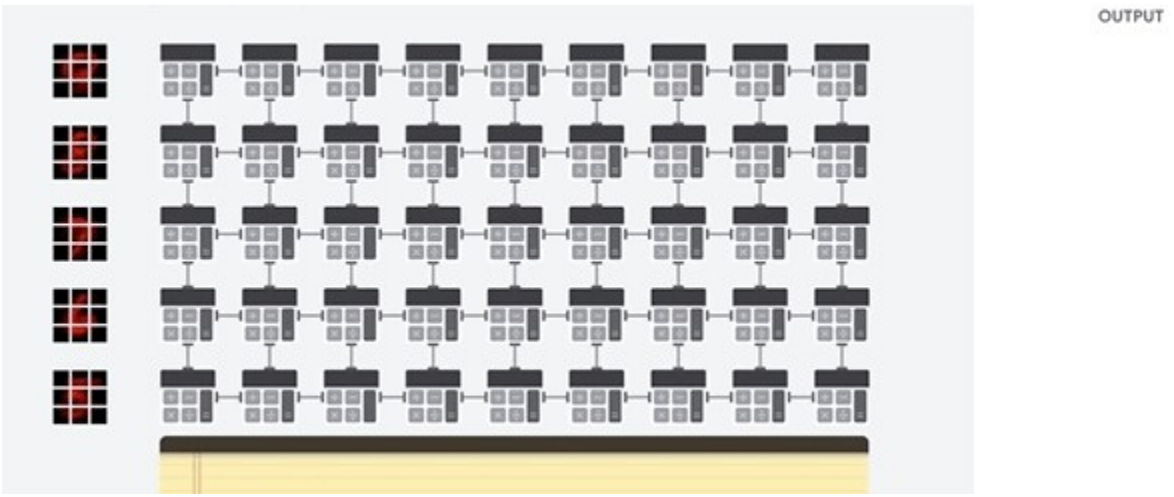
Arquitecturas TPU

Google diseñó Cloud TPU como un procesador matricial especializado para cargas de trabajo de redes neuronales. Las TPU no pueden ejecutar procesadores de texto, controlar motores de cohetes ni ejecutar transacciones bancarias, pero pueden manejar operaciones de matrices masivas que se usan en redes neuronales a velocidades rápidas.

La tarea principal de las TPU es el procesamiento matricial, que es una combinación de operaciones de multiplicación y acumulación. Las TPU contienen miles de acumuladores multiplicadores que están conectados directamente entre sí para formar una gran matriz física. Esto se conoce como arquitectura de arreglo sistólico. Cloud TPU v3 contiene dos arreglos sistólicos de 128 x 128 ALU en un solo procesador.

El host de TPU transmite datos a una fila de infeed. La TPU carga datos de la fila de infeed y los almacena en la memoria HBM. Cuando se completa el procesamiento, la TPU carga los resultados en la cola de salida. Luego, el host de TPU lee los resultados de la cola de salida y los almacena en la memoria del host.

Para realizar las operaciones de matrices, la TPU carga los parámetros de la memoria HBM en la unidad de multiplicación de matrices (MXU).



Luego, la TPU carga datos de la memoria HBM. A medida que se ejecuta cada multiplicación, el resultado se pasa al siguiente acumulador de multiplicación. El resultado es la suma de todos los resultados de multiplicación entre los datos y los parámetros. No se requiere acceso a la memoria durante el proceso de multiplicación de matrices.

Como resultado, las TPU pueden lograr una alta capacidad de procesamiento computacional en los cálculos de redes neuronales.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

Chip de TPU

Un chip de TPU contiene uno o más TensorCores. La cantidad de TensorCores depende de la versión del chip TPU. Cada TensorCore consta de una o más unidades de multiplicación de matrices (MXUs), una unidad vectorial y una unidad escalar. Para obtener más información sobre TensorCores, consulta Un supercomputador específico de dominio para el entrenamiento de redes neuronales profundas.

Una MXU se compone de acumuladores multiplicadores de 256 x 256 (TPU v6e) o 128 x 128 (versiones de TPU anteriores a la v6e) en un array sistólico. Las MXU proporcionan la mayor parte del poder de procesamiento en un TensorCore. Cada MXU puede realizar 16,000 operaciones de multiplicación y acumulación por ciclo. Todas las multiplicaciones toman entradas de bfloat16, pero todas las acumulaciones se realizan en formato de número FP32.

La unidad vectorial se usa para el procesamiento general, como las activaciones y el softmax. La unidad escalar se usa para el flujo de control, el cálculo de direcciones de memoria y otras operaciones de mantenimiento.

pod de TPU

Un pod de TPU es un conjunto contiguo de TPU agrupadas en una red especializada. La cantidad de chips TPU en un pod de TPU depende de la versión de TPU.

Porción

Una porción es una colección de chips que se encuentran dentro del mismo pod de TPU y que están conectados por interconexiones entre chips (ICI) de alta velocidad. Las porciones se describen en términos de chips o TensorCores, según la versión de TPU.

Forma del chip y *topología del chip* también se refieren a las formas de las rebanadas.

Multislice frente a Single Slice

Multislice es un grupo de porciones que extiende la conectividad de las TPU más allá de las conexiones de interconexión entre chips (ICI) y aprovecha la red del centro de datos (DCN) para transmitir datos más allá de una porción. El ICI sigue transmitiendo los datos de cada fragmento. Con esta conectividad híbrida, Multislice habilita el paralelismo entre las porciones y te permite usar una mayor cantidad de núcleos de TPU para un solo trabajo de lo que puede admitir una sola porción.

Las TPU se pueden usar para ejecutar una tarea en una sola porción o en varias. Consulta la Introducción a la tomografía multislice para obtener más detalles.

Cubo de TPU

Una topología 4x4x4 de chips TPU interconectados. Esto solo se aplica a las topologías 3D (a partir de la TPU v4).

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

Sparse

SparseCores son procesadores de flujo de datos que aceleran los modelos que se basan en incorporaciones que se encuentran en los modelos de recomendación. v5p incluye cuatro SparseCores por chip, y v6e incluye dos SparseCores por chip.

Resiliencia de ICI de Cloud TPU

La resiliencia de la ICI ayuda a mejorar la tolerancia a fallas de los vínculos ópticos y los conmutadores de circuitos ópticos (OCS) que conectan las TPU entre cubos. (Las conexiones de ICI dentro de un cubo usan vínculos de cobre que no se ven afectados). La resiliencia de ICI permite que las conexiones de ICI se enruten alrededor de las fallas de OCS y de ICI ópticas. Como resultado, mejora la disponibilidad de programación de las porciones de TPU, con la desventaja de una degradación temporal en el rendimiento de ICI.

En Cloud TPU v4 y v5p, la resiliencia de ICI está habilitada de forma predeterminada para las porciones que son de un cubo o más, por ejemplo:

- v5p-128 cuando se especifica el tipo de acelerador
- 4 × 4 × 4 cuando se especifica la configuración del acelerador

Versiones de TPU

La arquitectura exacta de un chip de TPU depende de la versión de TPU que uses. Cada versión de TPU también admite diferentes tamaños y configuraciones de rebanadas.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

ARQUITECTURA NEUROMÓRFICAS

la computación neuromórfica, también conocida como ingeniería neuromórfica, es un enfoque de la computación que imita la forma en que funciona el cerebro humano. Implica diseñar hardware y software que simulen las estructuras y funciones neuronales y sinápticas del cerebro para procesar información.

La computación neuromórfica puede parecer un campo nuevo, pero sus orígenes se remontan a la década de 1980. Fue la década en la que Misha Mahowald y Carver Mead desarrollaron la primera retina y cóclea de silicio y las primeras neuronas y sinapsis de silicio que fueron pioneras en el paradigma de la computación neuromórfica.

Dado que la computación neuromórfica se inspira en el cerebro humano, toma muchos elementos de la biología y la neurociencia.

Las neuronas "son las unidades fundamentales del cerebro y el sistema nervioso".5 Como mensajeros, estas células nerviosas transmiten información entre diferentes áreas del cerebro y a otras partes del cuerpo. Cuando una neurona se activa o se activa, desencadena la liberación de señales químicas y eléctricas que viajan a través de una red de puntos de conexión llamados sinapsis, lo que permite que las neuronas se comuniquen entre sí.6

Estos mecanismos neurológicos y biológicos se modelan en sistemas informáticos neuromórficos a través de redes neuronales de impulsos (SNN). Una red de red neuronal de impulsos es un tipo de red neuronal compuesta por neuronas de impulsos y sinapsis.

Las neuronas de pico almacenan y procesan datos de manera similar a las neuronas biológicas, y cada neurona tiene sus propios valores de carga, retraso y umbral. Las sinapsis crean vías entre las neuronas y también tienen valores de retraso y peso asociados a ellas. Estos valores (cargas neuronales, retrasos neuronales y sinápticos, umbrales neuronales y pesos sinápticos) se pueden programar dentro de los sistemas informáticos neuromórficos.

En la arquitectura neuromórfica, las sinapsis se representan como dispositivos sinápticos basados en transistores, que emplean circuitos para transmitir señales eléctricas. Las sinapsis suelen incluir un componente de aprendizaje, que altera sus valores de peso a lo largo del tiempo de acuerdo con la actividad dentro de la red neuronal de picos.

A diferencia de las neural networks convencionales, las SNN tienen en cuenta el tiempo en su operación. El valor de carga de una neurona se acumula con el tiempo; y cuando esa carga alcanza el valor umbral asociado de la neurona, se dispara, propagando información a lo largo de su red sináptica. Pero si el valor de la carga no supera el umbral, se disipa y finalmente se “filtra”. Además, los SNN están impulsados por eventos, con valores de retraso neuronal y sináptico que permiten la difusión asincrónica de información.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

ARQUITECTURA CUÁNTICA

La computación cuántica esta basada en las propiedades de la interacción cuántica entre las partículas subatómicas, como la superposición simultanea de dos estados en una sola partícula subatómica. La superposición cuántica, propiedad fundamental de la interacción cuántica, es ampliamente aprovechada para el desarrollo teórico de los algoritmos cuánticos, logrando una capacidad de procesamiento exponencial.

La superposición cuántica permite mantener simultáneamente múltiples estados en un bit cuántico, es decir “0” y “1” a la vez; a diferencia del bit – elemento fundamental en la computación actual – que únicamente es capaz de mantener un estado discreto, alternativo, a la vez, el “0” ó “1” lógico. La computación cuántica, aprovecha la superposición cuántica, para lograr el paralelismo cuántico y el paralelismo cuántico masivo.

Cualquier interacción con el mundo subatómico, producirá un cambio en este, es decir, cualquier medición o lectura traerá indefectiblemente un cambio. Este fenómeno cuántico es aprovechado en la tele transportación cuántica para la transmisión de qubits, y así mismo es utilizada como mecanismo de seguridad en la criptografía cuántica.

El bit cuántico “qubit”

El elemento básico de la computación cuántica es el bit cuántico o qubit1 (quantum bit por sus siglas en inglés), un qubit representa ambos estados simultáneamente, un “0” y un “1” lógico, dos estados ortogonales de una sub partícula atómica, como es representada en la figura 1. El estado de un qubit se puede escribir como {0, 1⟩}, describiendo su múltiple estado simultáneo.

Un vector de dos qubits, representa simultáneamente, los estados 00, 01, 10 y 11; un vector de tres qubits, representa simultáneamente, los estados 000, 001, 010, 011, 100, 101, 110, y 111; y así sucesivamente. Es decir un vector de n qubits, representa a la vez 2 n estados.

Entanglement

La capacidad de procesamiento paralelo de la computación cuántica, es enormemente 0incrementada por el procesamiento masivamente en paralelo, debido a una interacción que ocurre durante algunas millonésimas de segundo. Este fenómeno de la mecánica cuántica es llamado “entanglement”.

Debido al “entanglement”, dos partículas subatómicas, permanecen indefectiblemente relacionadas entre si, si han sido generadas en un mismo proceso. Por ejemplo la desintegración en un positrón y un electrón. Estas partículas forman subsistemas que no pueden describirse separadamente. Cuando una de las dos partículas sufre un cambio de estado, repercute en la otra. Esta característica se desencadena cuando se realiza una medición sobre una de las partículas.

Arquitectura de una computadora cuántica

La arquitectura de una computadora cuántica es similar a la de las computadoras tradicionales, con ciertos elementos propios de la computación cuántica. Oskin et al propone una arquitectura de una computadora quántica que esta conformada por una ALU cuántica, memoria cuántica, y un

planificador dinámico. La corrección de errores es un aspecto que debe ser tomado muy en cuenta en el diseño de una arquitectura cuántica.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

CLOUD COMPUTING

La computación en la nube es la disponibilidad a pedido de recursos de procesamiento (como infraestructura y almacenamiento), como servicios a través de Internet. Elimina la necesidad de que las personas y las empresas administren ellos mismos los recursos físicos y solo paguen por lo que usan.

os modelos de servicio de computación en la nube se basan en el concepto de compartir información, software y recursos de procesamiento bajo demanda en Internet. Las empresas o las personas pagan para acceder a un grupo virtual de recursos compartidos, incluidos los servicios de procesamiento, almacenamiento y herramientas de redes, que se encuentran en servidores remotos que son propiedad de los proveedores de servicios y son administrados por ellos.

Una de las muchas ventajas de la computación en la nube es que solo pagas por lo que usas. Esto permite que las organizaciones escalen con mayor rapidez y eficiencia sin la carga de comprar y mantener sus propios centros de datos físicos y servidores.

En términos más simples, la computación en la nube usa una red (por lo general, Internet) para conectar a los usuarios con una plataforma en la nube en la que solicitan y acceden a servicios de computación alquilados. Un servidor central controla toda la comunicación entre los dispositivos y los servidores del cliente para facilitar el intercambio de datos. Las funciones de seguridad y privacidad son componentes habituales para proteger esta información.

Cuando se adopta la arquitectura de computación en la nube, no hay una solución universal. Lo que funciona para otra empresa puede no satisfacer tus necesidades ni las de tu empresa. De hecho, esta flexibilidad y versatilidad es uno de los sellos distintivos de la nube, lo que permite a las empresas adaptarse con rapidez a mercados o métricas cambiantes.

Existen tres modelos diferentes de implementación de computación en la nube: nube pública, nube privada y nube híbrida.

Tipos de modelos de implementación de computación en la nube

Nube pública

Las nubes públicas se ejecutan mediante proveedores de servicios en la nube externos. Ofrecen recursos de procesamiento, almacenamiento y red a través de Internet, lo que permite que las empresas accedan a recursos compartidos bajo demanda en función de sus requisitos y objetivos comerciales únicos.

Nube privada

A las nubes privadas las compila, administra y posee una sola organización y se alojan de forma privada en sus propios centros de datos, lo que comúnmente se conoce como “local” o “en las instalaciones”. Proporcionan mayor control, seguridad y administración de datos, a la vez que

permiten que los usuarios internos se beneficien de un conjunto compartido de recursos de procesamiento, almacenamiento y red.

Nube híbrida

Las nubes híbridas combinan modelos de nube pública y privada, lo que permite a las empresas aprovechar los servicios de nube pública y mantener las capacidades de seguridad y cumplimiento que suelen encontrarse en las arquitecturas de nube privada.

Existen tres tipos principales de modelos de servicios de computación en la nube que puedes seleccionar según el nivel de control, flexibilidad y administración que necesite tu empresa:

Infraestructura como servicio (IaaS)

Infraestructura como servicio (IaaS) ofrece acceso bajo demanda a los servicios de infraestructura de TI, incluidos el procesamiento, el almacenamiento, las herramientas de redes y la virtualización. Proporciona el nivel más alto de control sobre tus recursos de TI y se asemeja más a los recursos de TI locales tradicionales.

Plataforma como servicio (PaaS)

Plataforma como servicio (PaaS): ofrece todos los recursos de hardware y software necesarios para el desarrollo de aplicaciones en la nube. Con PaaS, las empresas pueden enfocarse por completo en el desarrollo de aplicaciones sin la carga de administrar y mantener la infraestructura subyacente.

Software como servicio (SaaS)

Software como servicio (SaaS): proporciona una pila de aplicaciones completa como servicio, desde la infraestructura subyacente hasta el mantenimiento y las actualizaciones del software de la app. Una solución de SaaS suele ser una aplicación de usuario final, en la que el proveedor de servicios en la nube administra y mantiene la infraestructura y el servicio.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

COMPUTACIÓN HETEROGÉNEA

La computación heterogénea es un paradigma en la arquitectura informática que integra múltiples tipos de procesadores y unidades informáticas dentro de un único sistema para lograr un rendimiento y una eficiencia optimizados. En tal entorno, varios procesadores, como CPU y GPU, matrices de puertas programables en campo (FPGA) y otros aceleradores especializados colaboran para ejecutar diversas tareas computacionales.

La esencia de la informática heterogénea radica en su capacidad de distribuir cargas de trabajo según las fortalezas de cada tipo de procesador. Cada tipo de procesador se destaca en el manejo de tipos específicos de operaciones: las CPU son adecuadas para tareas secuenciales, las GPU para procesamiento paralelo y las FPGA para tareas personalizables y de alto rendimiento. Esta distribución permite mejorar el rendimiento, ya que las tareas son procesadas de manera más rápida y eficiente por las personas más apropiadas, hardware. Además, mejora la eficiencia energética al reducir la carga computacional en procesadores menos adecuados, reduciendo así el consumo de energía.

Arquitectura de sistema heterogéneo

La arquitectura de sistemas heterogéneos (HSA) tiene como objetivo proporcionar una plataforma unificada donde diversas unidades de procesamiento puedan comunicarse y cooperar de manera eficiente, mejorando así el rendimiento general del sistema, la eficiencia energética y la programabilidad.

HSA aborda varios desafíos clave en los sistemas heterogéneos tradicionales, como la coherencia de la memoria, la complejidad de la programación y el intercambio eficiente de datos. Uno de los conceptos centrales de HSA es el uso de un modelo de memoria compartida, que permite que diferentes procesadores accedan al mismo espacio de memoria sin necesidad de copiar datos explícitos. Este modelo de memoria compartida simplifica la programación y mejora el rendimiento al reducir la sobrecarga asociada con la transferencia de datos entre procesadores.

En HSA, todos los procesadores se tratan como elementos informáticos de primera clase, cada uno de ellos capaz de acceder directamente a la memoria del sistema y comunicarse con otros procesadores a través de una interconexión de alta velocidad. Este enfoque elimina el cuello de botella tradicional de tener que enrutar todos los datos a través de la CPU, lo que permite un procesamiento paralelo más eficiente y una ejecución más rápida de las tareas, que se descargan a procesadores especializados como GPU o FPGA.

HSA también introduce un conjunto estandarizado de API y herramientas de programación que abstraen las complejidades de la informática heterogénea. Esta estandarización permite a los desarrolladores escribir aplicaciones que aprovechan al máximo las diversas capacidades de procesamiento del hardware compatible con HSA sin un conocimiento profundo de los detalles del hardware subyacente.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

Al proporcionar un marco común para la computación heterogénea, HSA tiene como objetivo acelerar el desarrollo de aplicaciones de alto rendimiento y eficiencia energética en varios dominios, incluido el procesamiento de gráficos, la computación científica, el aprendizaje automático y más.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

ARQUITECTURAS DISTRIBUIDAS EN LA NUBE

La arquitectura de nube es un elemento clave de la compilación en la nube. Se refiere al diseño y conecta todos los componentes y tecnologías necesarios para la computación en la nube.

La migración a la nube puede ofrecer muchos beneficios empresariales en comparación con los entornos locales, desde una mejor agilidad y escalabilidad hasta una mayor rentabilidad. Si bien muchas organizaciones pueden comenzar con un enfoque “lift-and-shift” en el que las aplicaciones locales se trasladan con modificaciones mínimas, en última instancia, será necesario crear e implementar aplicaciones de acuerdo con las necesidades y los requisitos de los entornos de nube.

La arquitectura de nube determina cómo se integran los componentes, de modo que puedas agrupar, compartir y escalar recursos a través de una red. Considéralo como un plano de compilación para implementar y ejecutar aplicaciones en entornos de nube.

La arquitectura de nube se refiere a cómo varios componentes de la tecnología de nube, como el hardware, los recursos virtuales, las capacidades del software y los sistemas de red virtual, interactúan y se conectan para crear entornos de computación en la nube. Actúa como un plano que define la mejor manera de combinar de manera estratégica los recursos a fin de crear un entorno de nube para una necesidad empresarial específica.

Componentes de la arquitectura de nube

Los componentes de la arquitectura de nube incluyen lo siguiente:

- Una plataforma de frontend
- Una plataforma de backend
- Un modelo de entrega basado en la nube
- Una red (internet, intranet o intercloud)

En la computación en la nube, las plataformas de frontend contienen la infraestructura del cliente: interfaces de usuario, aplicaciones del cliente y el dispositivo o la red del cliente que permite a los usuarios interactuar y acceder a los servicios de computación en la nube. Por ejemplo, puedes abrir el navegador web en el teléfono celular y editar un Documento de Google. Estos tres elementos describen los componentes de la arquitectura de nube de frontend.

Por otro lado, el backend se refiere a los componentes de la arquitectura en la nube que conforman la nube en sí, incluidos los recursos de procesamiento, el almacenamiento, los mecanismos de seguridad, la administración y mucho más.

A continuación, se muestra una lista de los componentes de backend principales:

Aplicación: El software o la aplicación de backend al que accede el cliente desde el frontend para coordinar o cumplir con los requisitos y solicitudes de clientes.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

Servicio: El servicio es el núcleo de la arquitectura de nube y se encarga de todas las tareas que se ejecutan en un sistema de computación en la nube. Administra los recursos a los que puedes acceder, incluido el almacenamiento, los entornos de desarrollo de aplicaciones y las aplicaciones web.

Nube del entorno de ejecución: La nube del entorno de ejecución proporciona el entorno en el que se ejecutan los servicios y actúa como un sistema operativo que controla la ejecución de las tareas y la administración del servicio. Los entornos de ejecución usan tecnología de virtualización para crear hipervisores que representen todos tus servicios, como las apps, los servidores, el almacenamiento y las herramientas de redes.

Almacenamiento: El componente de almacenamiento en el backend es donde se almacenan los datos para operar aplicaciones. Si bien las opciones de almacenamiento en la nube varían según el proveedor, la mayoría de los proveedores ofrecen servicios flexibles de almacenamiento escalable que están diseñados para almacenar y administrar grandes cantidades de datos en la nube. El almacenamiento puede incluir discos duros, unidades de estado sólido o discos persistentes en los compartimentos del servidor.

Infraestructura: es probable que la infraestructura sea el componente más conocido de la arquitectura en la nube. De hecho, es posible que hayas pensado en que la infraestructura de nube *es* arquitectura de nube. Sin embargo, la infraestructura de nube abarca todos los componentes de hardware principales que impulsan los servicios en la nube, como la CPU, la unidad de procesamiento gráfico (GPU), los dispositivos de red y otros componentes de hardware necesarios para que los sistemas se ejecuten sin problemas. La infraestructura también se refiere a todo el software necesario para ejecutar y administrar todo.

Por otro lado, la arquitectura de nube es el plan que dicta cómo se organizan los recursos y la infraestructura de la nube.

Administración: Los modelos de servicios en la nube requieren que los recursos se administren en tiempo real de acuerdo con los requisitos del usuario. Es esencial usar software de administración, también conocido como middleware, para coordinar la comunicación entre los componentes de la arquitectura de backend y del frontend en la nube, y asignar recursos a tareas específicas. Además de middleware, el software de administración también incluirá capacidades para la supervisión de uso, la integración de datos, la implementación de aplicaciones y la recuperación ante desastres.

Seguridad: A medida que más organizaciones siguen adoptando la computación en la nube, es fundamental implementar herramientas y funciones de seguridad en la nube para proteger los datos, las aplicaciones y las plataformas. Es esencial planificar y diseñar la seguridad de los datos y de redes para proporcionar visibilidad, evitar la pérdida de datos y el tiempo de inactividad, y garantizar la redundancia. Esto puede incluir copias de seguridad regulares, depuración y firewalls virtuales.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

¿Cómo funciona la arquitectura de nube?

En la arquitectura de nube, cada uno de los componentes trabaja en conjunto para crear una plataforma de computación en la nube que proporciona a los usuarios acceso a recursos y servicios a pedido.

El backend contiene todos los recursos de computación en la nube, servicios, almacenamiento de datos y aplicaciones que ofrece un proveedor de servicios en la nube. Se usa una red para conectar los componentes de la arquitectura de nube de frontend y backend, lo que permite enviar y recibir datos entre ellos. Cuando los usuarios interactúan con el frontend (o la interfaz del cliente), envía consultas al backend mediante middleware, en el que el modelo de servicio realiza la tarea o solicitud específica.

Los tipos de servicios disponibles para usar dependen del modelo de entrega basado en la nube o del modelo de servicio que hayas elegido. Existen tres modelos principales de servicios de computación en la nube:

- **Infraestructura como servicio (IaaS):** Este modelo proporciona acceso a pedido a la infraestructura de nube, como servidores, el almacenamiento y las herramientas de redes. Esto elimina la necesidad de adquirir, administrar y mantener la infraestructura local.
- **Plataforma como servicio (PaaS):** Este modelo ofrece una plataforma de procesamiento con todas las herramientas de infraestructura y software subyacentes necesarias para desarrollar, ejecutar y administrar aplicaciones.
- **Software como servicio (SaaS):** Este modelo ofrece aplicaciones basadas en la nube que proporcionan y mantienen el proveedor de servicios, lo que elimina la necesidad de que los usuarios finales implementen software localmente.

Capas de arquitectura de nube

Una forma más simple de comprender cómo funciona la arquitectura de la nube es pensar en todos estos componentes como varias capas ubicadas una encima de la otra para crear una plataforma en la nube.

Estas son las capas básicas de la arquitectura de nube:

1. **Hardware:** servidores, almacenamiento, dispositivos de red y otros elementos de hardware que impulsan la nube.
2. **Virtualización:** una capa de abstracción que crea una representación virtual de los recursos físicos de almacenamiento y procesamiento. Esto permite que varias aplicaciones usen los mismos recursos.
3. **Aplicación y servicio:** esta capa coordina y admite solicitudes de la interfaz de usuario de frontend, y ofrece diferentes servicios según el modelo de servicio en la nube, desde la asignación de recursos hasta las herramientas de desarrollo de aplicaciones y las aplicaciones basadas en la Web.

Tarea 2: Arquitecturas de GPU, TPU, Neuromórficas, Cuántica, Cloud Computing, Computación Heterogénea y arquitecturas distribuidas en la nube.

Paula Sandoval – Digitales III

Tipos de arquitectura de nube

La adopción de la nube no es universal. Deberás considerar qué tipo de nube quieres compilar según tus inversiones en tecnología existentes, tus requisitos comerciales específicos y los objetivos generales que esperas lograr.

Hay tres tipos principales de arquitectura en la nube que puedes elegir: **pública, privada o híbrida**.

La **arquitectura de la nube pública** usa recursos de computación en la nube y una infraestructura física cuya propiedad y administración pertenecen a un proveedor externo de servicios de nube. Las nubes públicas te permiten escalar recursos con facilidad y sin la necesidad de invertir en tu propio hardware o software, pero puedes usar arquitecturas de multiusuario para otros clientes al mismo tiempo.

La **arquitectura de la nube privada** se refiere a una nube dedicada cuya propiedad y administración corresponde a tu organización. Se aloja de forma privada en las instalaciones locales de tu propio centro de datos, lo que proporciona más control sobre los recursos y más seguridad sobre la infraestructura y los datos. Sin embargo, esta arquitectura es considerablemente más costosa y su mantenimiento requiere más experiencia de TI.

La **arquitectura de nube híbrida** usa arquitectura de nube pública y privada para entregar una combinación flexible de servicios en la nube. Una nube híbrida te permite migrar cargas de trabajo entre entornos, lo que te permite usar los servicios que mejor se adapten a las demandas de tu empresa y a la carga de trabajo. Las arquitecturas de la nube híbrida suelen ser la solución preferida para las empresas que necesitan controlar sus datos, pero también quieren aprovechar las ofertas de nube pública.

En los últimos años, también surgió la **arquitectura de múltiples nubes** , ya que cada vez más organizaciones usan los servicios en la nube de varios proveedores de servicios en la nube. Los entornos de múltiples nubes están ganando popularidad por su flexibilidad y capacidad para hacer coincidir mejor los casos de uso con las ofertas específicas, independientemente del proveedor.