REPÚBLICA BOLIVARIANA DE VENEZUELA
MINISTERIO DEL PODER POPULAR PARA LA DEFENSA
UNIVERSIDAD NACIONAL EXPERIMENTAL POLITECNICA
DE LA FUERZA ARMADA NACIONAL BOLIVARIANA
UNEFA NÚCLEO PUERTO CABELLO

Arquitectura del Computador 3do Corte.

ARQUITECTURAS DE LA ING. DE SISTEMAS.

Profesor: Ing. Javier Vilchez.

Grupo 1: El Fin del "Muro de la Memoria" e Interconexiones

Este grupo analiza por qué la memoria RAM tradicional es el actual "cuello de botella".

- **HBM3 vs. DDR5:** Diferencias físicas en el empaquetado y por qué la IA prefiere el ancho de banda sobre la capacidad.
- **Protocolo CXL (Compute Express Link):** Cómo permite que el procesador comparta su espacio de memoria física con aceleradores externos.
- **Tecnología NVLink:** El reemplazo del bus PCIe para la comunicación directa entre tarjetas gráficas.
- **Unified Memory Architecture:** Cómo Apple (Silicon) y NVIDIA están eliminando la separación entre memoria de sistema y memoria de video.
- Ejemplos de su aplicabilidad en la actualidad.

Grupo 2: Evolución de las ISA (Arquitecturas de Instrucciones)

Investigación sobre cómo el "idioma" que entiende el hardware se adapta a las matemáticas de la IA.

- **Intel AMX (Advanced Matrix Extensions):** Cómo los nuevos procesadores incluyen hardware físico dedicado a multiplicar matrices.
- **RISC-V y Personalización:** Por qué la arquitectura de código abierto es el futuro para crear chips de IA "a medida".
- **Instrucciones SIMD vs. Operaciones de Tensor:** Explicar la diferencia entre procesar vectores y procesar matrices multidimensionales.
- **Compiladores de Hardware (MLIR):** Cómo el software traduce modelos de IA a instrucciones físicas del procesador.
- Ejemplos de su aplicabilidad en la actualidad.

Grupo 3: Unidades Especializadas (NPU, TPU y Aceleradores)

Diferenciar los nuevos componentes físicos que se han sumado a la arquitectura tradicional.

- **Arquitectura de las NPU (Neural Processing Units):** Cómo funcionan los núcleos de IA en los nuevos procesadores de consumo.
- **Google TPU (Tensor Processing Unit):** El uso de "Arreglos Sistólicos" para procesar datos como un flujo continuo.
- **Eficiencia en TOPS (Tera Operations Per Second):** Cómo medir el rendimiento de un computador de IA más allá de los GHz.
- **Hardware de Precisión Reducida (FP8, INT8):** Por qué la IA no necesita precisión matemática exacta y cómo esto ahorra energía física.
- Ejemplos de su aplicabilidad en la actualidad.

Grupo 4: Arquitectura Distribuida y Centros de Datos

Cómo se escalan las bases de datos y la arquitectura lógica para la inteligencia masiva.

- **Clústeres de GPU:** El diseño físico de un "Supercomputador de IA".
- **Hardware para Bases de Datos Vectoriales:** Cómo se optimiza el almacenamiento para buscar información por similitud y no por tablas tradicionales.
- **InfiniBand vs. Ethernet:** Por qué las redes de servidores de IA necesitan cables y protocolos especiales para evitar latencia.
- **Sistemas de Enfriamiento Líquido:** El reto físico de mantener operativas las arquitecturas de alto rendimiento.
- Ejemplos de su aplicabilidad en la actualidad.

Grupo 5: Edge AI y Arquitecturas de Bajo Consumo

La IA llevada a la parte física de los dispositivos móviles y sensores.

- **Arquitectura ARM en el Escritorio:** Cómo los procesadores de bajo consumo están desplazando a la arquitectura x86.
- **In-Memory Computing:** El concepto de procesar datos directamente donde se almacenan (en la RAM/ROM) para evitar el calor.
- **Cuantización en Hardware:** Cómo "comprimir" un modelo de IA para que quepa en la memoria limitada de un microcontrolador.
- **Seguridad a Nivel de Chip:** Cómo la arquitectura protege los datos del usuario cuando la IA se ejecuta localmente.
- Ejemplos de su aplicabilidad en la actualidad.

NOTA:

La evaluación será tipo defensa grupal, mas diapositiva a los temas correspondientes por grupo.

Presencia.

Léxico.

Dominio del tema.

Uso del Material de apoyo.