Instituto Politécnico Nacional

“Escuela Superior de Cómputo”

Integrantes:

- Ignacio Cortés Atzin Maxela

Unidad de aprendizaje: Cómputo paralelo

Profesor: Luis Alberto Ibáñez Zamora

Grupo: 6BM2

“P3 Unidades de Procesamiento Neuronal (NPUs)”

Informe: Unidades de Procesamiento Neural (NPU)

1. Definición y Diferencias con CPU y GPU

Una Unidad de Procesamiento Neuronal (NPU, por sus siglas en inglés) es un tipo especializado de procesador diseñado específicamente para acelerar cargas de trabajo relacionadas con inteligencia artificial, en particular aquellas basadas en redes neuronales. A diferencia de las CPU, que son versátiles y pueden ejecutar una amplia gama de tareas secuenciales, y las GPU, que sobresalen en operaciones paralelas masivas como gráficos y aprendizaje profundo, las NPUs están optimizadas exclusivamente para la ejecución de inferencias de modelos de machine learning con alta eficiencia energética. Se utilizan en tareas donde se requiere velocidad y bajo consumo, como el reconocimiento de imágenes, el procesamiento de lenguaje natural y la visión artificial en dispositivos con recursos limitados. Las ventajas principales de las NPUs incluyen bajo consumo de energía, alta velocidad de inferencia y procesamiento dedicado. Sin embargo, entre sus desventajas se encuentra la limitada flexibilidad para tareas generales y la necesidad de adaptarse a marcos específicos o herramientas propietarias.

Características	CPU	GPU	NPU
Propósito	General	Paralelismo gráfico	Inferencia IA
Nº de núcleos	Bajo (1-16)	Medio/Alto (100s)	Alto (específicos)
Eficiencia IA	Baja	Alta	Muy alta
Consumo	Alto	Msedio/Alto	Bajo
Latencia	Alta	Media	Muy baja

2. Arquitecturas de NPUs

Existen diversas arquitecturas de NPUs desarrolladas por distintas compañías para aplicaciones embebidas, móviles o de borde. El **Google Coral / Edge TPU** es una NPU enfocada en el procesamiento local de modelos TensorFlow Lite, con una potencia aproximada de 4 TOPS y bajo consumo, disponible en formatos USB, PCIe y módulos SOM. **Intel Movidius / Myriad X**, utilizado en cámaras inteligentes y drones, ofrece alrededor de 1 TOPS y tiene una arquitectura con SHAVE cores para visión computacional, disponible como stick USB o integrable vía M.2. Por otro lado, el **Hailo-8** es una NPU de alto rendimiento con hasta 26 TOPS y eficiencia energética superior, diseñada para integrarse mediante interfaces como PCIe. Otras opciones incluyen el **BrainChip Akida**, que introduce un enfoque neuromórfico, el **Kneron KL520/KL730**, muy usados en dispositivos IoT por su tamaño y eficiencia, y el **Apple Neural Engine**, integrado directamente en chips A-series y M-series, con hasta 15.8 TOPS. Finalmente,

MediaTek APU, integrada en procesadores móviles, también proporciona un procesamiento de IA dedicado, compatible con diversas redes neuronales, se pueden observar mejor sus características a continuación.

Google Coral / Edge TPU

- Arquitectura: ASIC para TensorFlow Lite.
- Potencia: ~4 TOPS.
- Consumo: ~0.5 W.
- Integración: USB, M.2 A+E, PCIe.

Intel Movidius / Myriad X

- Arquitectura: 16 SHAVE cores + coprocesador.
- Potencia: ~1 TOPS.
- Consumo: ~1 W.
- Integración: USB, PCIe, M.2.

Hailo-8

- Arquitectura: Heterogénea con aceleradores dedicados.
- Potencia: 26 TOPS.
- Consumo: 2.5 W.
- Integración: M.2, PCIe, Mini PCIe.

BrainChip Akida

- Arquitectura: Redes neuronales espaciotemporales (SNN).
- Potencia: ~1 TOPS.
- Consumo: < 0.3 W.
- Integración: PCIe, M.2, standalone.

Kneron KL520 / KL730

- KL520: 0.3 TOPS, USB/SPI.
- KL730: 1.4 TOPS, video 4K, ISP incluido.

Apple Neural Engine (ANE)

- Arquitectura: Integrada en SoC.
- Potencia: 18 - 35 TOPS.
- Consumo: Bajo (optimizado para iOS).
- Integración: Interna.

MediaTek APU

- Arquitectura: Integrada en Helio y Dimensity.
- Potencia: 4.5-5 TOPS.
- Integración: Interna.

3. Integración CPU-GPU-NPU

La integración de una NPU con CPU y GPU implica una distribución de tareas donde cada componente aporta según sus fortalezas. La CPU gestiona el sistema operativo, orquestación de tareas y preprocesamiento de datos (como decodificación de imágenes o audio). La GPU puede encargarse de tareas gráficas o de postprocesamiento, como visualización o renderizado, mientras que la NPU se dedica a la inferencia del modelo de IA. Un ejemplo típico sería el flujo de una cámara inteligente: la CPU prepara la imagen, la NPU ejecuta la inferencia para detectar objetos, y la GPU renderiza los resultados. Esta integración puede realizarse mediante buses como USB, M.2, PCIe o SPI, dependiendo del tipo de dispositivo y de los requerimientos de ancho de banda. El balance de carga eficiente entre estos componentes permite optimizar el rendimiento y reducir el consumo energético, especialmente importante en dispositivos embebidos y móviles.

La CPU coordina las tareas y ejecuta el preprocesamiento (e.g., normalización de imagen), mientras que la NPU realiza la inferencia. La GPU se usa para postprocesamiento o visualización.

Ejemplo: - CPU: Captura y redimensiona imagen. - NPU: Inferencia (detección, clasificación). - GPU: Renderiza o aplica efectos visuales.

Interconexión: - USB 3.0: Universal, limitada en velocidad. - PCIe/M.2: Alta velocidad, baja latencia. - SPI/I2C: Bajo consumo, para IoT.

4. Aplicaciones de NPUs

Las NPUs se utilizan ampliamente en aplicaciones de inteligencia artificial en tiempo real.

a) Visión Artificial: permiten realizar tareas como clasificación de imágenes, detección de objetos y segmentación semántica, incluso en dispositivos de bajo consumo

- Detección de objetos, clasificación.
- Ejemplo: Nest Cam con Coral TPU.

b) Video en Tiempo Real: se usan para detección de movimiento, mejora de imagen y compresión inteligente.

- Seguimiento de movimiento, compresión inteligente.
- Ejemplo: Videocámaras industriales con Hailo-8.

c) Seguridad y Encriptación: permiten aplicar inferencias de detección de amenazas, control de acceso con reconocimiento facial y cifrado de datos con IA.

- Autenticación facial, detección de amenazas.
- Ejemplo: Accesos biométricos con Kneron o Akida.

Dispositivos: - Cámaras inteligentes, smartphones (iPhone, Xiaomi), robots (Boston Dynamics), Raspberry Pi + Coral TPU.

Ejemplos concretos de uso incluyen cámaras de seguridad inteligentes que reconocen personas o placas vehiculares sin conectarse a la nube; teléfonos móviles que usan la NPU para acelerar el reconocimiento facial y los asistentes inteligentes; y robots autónomos que procesan su entorno visual en tiempo real. Además, dispositivos IoT como Raspberry Pi pueden usar NPUs externas para procesar redes neuronales de forma local, sin requerir conectividad constante.

5. HATs y Módulos de Integración

Los HATs (Hardware Attached on Top) o carrier boards son placas de expansión que permiten conectar módulos como NPUs a plataformas embebidas como Raspberry Pi o Jetson Nano. Estas tarjetas proporcionan interfaces físicas (como PCIe o USB) y compatibilidad eléctrica para que los módulos de procesamiento se integren sin necesidad de rediseñar el hardware base. Por ejemplo, el **Waveshare PCIe-to-M.2 E-Key HAT+** permite conectar módulos M.2 con NPUs como el Coral o Hailo a una Raspberry Pi 5 vía PCIe. El **SparkFun MicroMod Machine Learning Carrier Board** integra un conector para módulos TinyML y sensores, ideal para desarrollos portátiles. El **Seeed Mini PCIe Adapter** permite utilizar tarjetas Coral o Movidius Myriad X con placas Jetson o CM4. También, el **AAEON AI HAT para CM4** ofrece integración con módulos AI Edge y soporte de drivers para cámaras e inferencia acelerada. Es fundamental verificar compatibilidades, necesidades de alimentación y drivers para asegurar un funcionamiento correcto.

Un **HAT (Hardware Attached on Top)** es una placa complementaria que se conecta a una SBC (Single Board Computer) como Raspberry Pi o Jetson, extendiendo sus capacidades.

Ejemplos:

- **Waveshare PCIe-to-M.2 HAT+**: Conecta Edge TPU o Hailo-8 a Raspberry Pi 5.
- **SparkFun MicroMod ML Carrier**: Soporta módulos con NPU para prototipos.
- **Seeed Mini PCIe Adapter**: Para integrar mini-PCIe NPUs.
- **AAEON AI HAT para CM4**: Diseñado para Coral o Myriad X, ideal para visión artificial.

Tabla comparativa de NPUs (TOPS, consumo, interfaz, costo)

NPU	Potencia (TOPS)	Consumo Aproximado	Interfaz / Formato	Costo Aproximado (USD)
Google Edge TPU	4	~0.5 - 2 W	USB, M.2, PCIe	$60 - $120
Intel Movidius Myriad X	1	~1 W	USB (Neural Compute Stick), PCIe	$70 - $100
Hailo-8	26	~2.5 W	M.2, PCIe	$120 - $200
Apple Neural Engine	15.8 (M1), 35+ (M4)	Integrado (chip)	Interno (SoC en iPhone/Mac)	No disponible por separado
MediaTek APU 3.0	4	~1 - 2 W	Interno (en SoC Dimensity)	Incluido en dispositivos
BrainChip Akida	~1.2	<1 W	M.2, USB	$100 - $150
Kneron KL520	~0.6	<0.5 W	USB, SPI	$50 - $90
Kneron KL730	1.2 – 2	~1 W	USB, PCIe	$90 - $130

Tabla de compatibilidad de HATs/carriers

HAT / Carrier Board	Compatible con	Interfaz / Conector	Soporte para NPU	Drivers necesarios
Waveshare PCIe-to-M.2 E-Key HAT+	Raspberry Pi 5	PCIe (M.2 E-Key)	Google Edge TPU, Hailo-8, Kneron	Sí (según NPU, Linux)
SparkFun MicroMod Machine Learning Carrier	MicroMod CPU boards (ESP32, SAMD51)	Qwiic, GPIO	Syntiant NDP101, Himax WE-I Plus	Sí (firmware + librerías)
Seeed Mini PCIe Adapter	Jetson Nano, CM4 con PCIe baseboard	Mini PCIe	Coral, Movidius, Hailo-8	Sí (varía por NPU y sistema)
AAEON AI HAT para Raspberry Pi CM4	Raspberry Pi Compute Module 4	PCIe (via baseboard)	Myriad X, Hailo-8, Kneron	Sí (incluye soporte oficial)
Coral M.2 Accelerator with Dual Edge TPU	Raspberry Pi 5, Jetson Xavier NX, PC	M.2 B+M Key (PCIe)	Edge TPU	Sí (Google Coral drivers)
Luxonis OAK-1 HAT Adapter	Raspberry Pi 4 / CM4	USB 3.0 / GPIO	Movidius Myriad X (con cámara)	Sí (DepthAI SDK)

Conclusión

A través de esta investigación, comprendí de forma clara qué son las Unidades de Procesamiento Neuronal (NPUs) y cómo se diferencian de otros procesadores como la CPU y la GPU. Aprendí que las NPUs están diseñadas específicamente para acelerar tareas de inteligencia artificial, especialmente la inferencia de redes neuronales, y que su principal ventaja es su eficiencia energética y rapidez al ejecutar modelos entrenados. Esta característica las hace ideales para aplicaciones en dispositivos embebidos, móviles o que operan sin conexión a la nube.

También descubrí que existen diferentes tipos de NPUs desarrolladas por empresas como Google, Intel, Apple, Hailo y MediaTek, cada una con arquitecturas, interfaces y niveles de potencia distintos. Compararlas me ayudó a entender mejor el equilibrio entre rendimiento (TOPS), consumo energético, tipo de conexión (USB, PCIe, M.2) y precio. Me resultó especialmente interesante saber que algunas NPUs, como la Coral o la Myriad X, pueden integrarse fácilmente a dispositivos como la Raspberry Pi mediante adaptadores HAT o carrier boards.

Otro aprendizaje valioso fue conocer cómo se distribuye el procesamiento entre CPU, GPU y NPU en un sistema. Ahora entiendo que, por ejemplo, una imagen puede ser preprocesada por la CPU, procesada por la NPU para detectar objetos, y postprocesada por la GPU para visualizar resultados. Este reparto inteligente mejora la eficiencia del sistema completo, algo que no conocía en profundidad antes de este trabajo.

Finalmente, descubrí que las NPUs tienen aplicaciones prácticas en áreas como visión artificial, análisis de video y seguridad, y que ya forman parte de productos reales como cámaras inteligentes, teléfonos, robots autónomos y dispositivos IoT. Esta investigación me permitió no solo adquirir conocimientos técnicos, sino también dimensionar el impacto que estas tecnologías tienen en el desarrollo de soluciones más rápidas, seguras y eficientes para la vida diaria.

Fuentes

- Waveshare. (s.f.). *PCIe to M.2 E-Key HAT+ for Raspberry Pi 5*. https://www.waveshare.com
- Coral. (s.f.). *Coral by Google*. https://coral.ai
- Intel. (2020). *Intel® Movidius™ Myriad™ X VPU*. https://www.intel.com/content/www/us/en/products/details/processors/movidius/myriad-x.html
- Hailo. (s.f.). *Hailo-8 AI Processor*. https://hailo.ai/product/
- Apple Inc. (2023). *Apple Neural Engine*. https://developer.apple.com/machine-learning/neural-engine/