Foro Empresas Innovadoras

amper

DOCUMENTO COMPLEMENTARIO AL INFORME "SITUACIÓN Y TENDENCIAS EN EL USO DE LA IA EN EL SECTOR DE LA DEFENSA"

Hardware/Chips para inteligencia artificial aplicada a la defensa y seguridad

José María Insenser
Miembro del Foro de Empresas Innovadoras (FEI)
AMETIC (IPCEI Chip)

Septiembre 2025

Tabla de contenido

Introducción

La microelectrónica es, desde su aparición en la década de los 50 del siglo pasado, con la invención del circuito integrado[1], una tecnología habilitadora clave que ha posibilitado el avance de otras tecnologías que han transformado la economía y la sociedad, como la industria 4.0, las telecomunicaciones, la automoción/movilidad, las ciudades inteligentes, la aeronáutica, el sector espacial, la energía, la instrumentación, la defensa y la seguridad, la industria agroalimentaria, el Control de tráfico, la Monitorización de infraestructuras, la ciberseguridad, etc. y , por supuesto, la Inteligencia Artificial.

La microelectrónica está en la base de todos los procesos de digitalización y automatización industriales clave, contribuyendo de forma notable al aumento de la productividad (véase *Figura 1*).

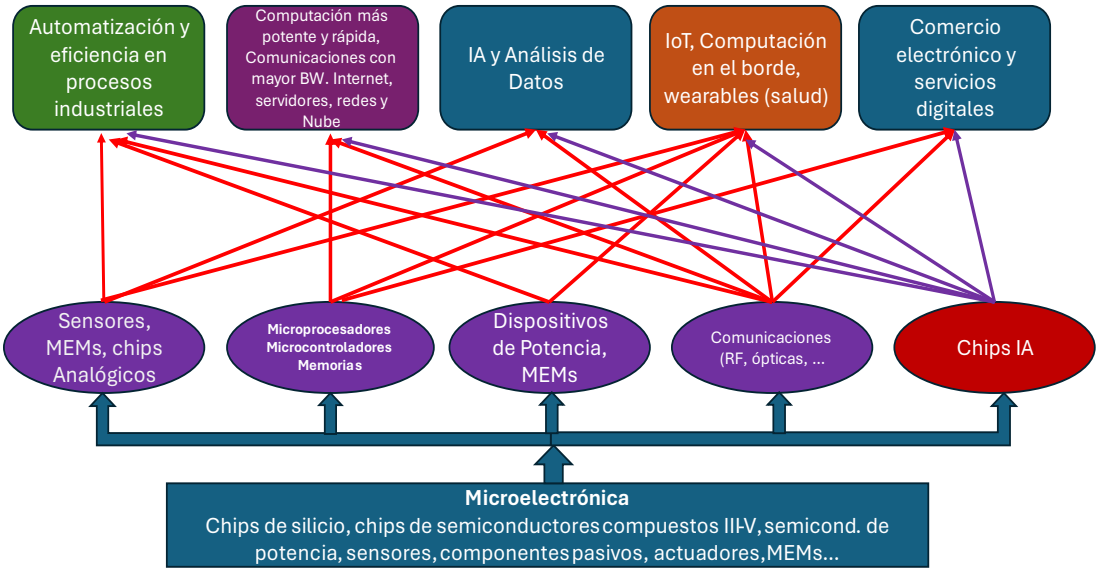


Figura 1. La microelectrónica y los semiconductores son la base para el aumento de la productividad en los distintos procesos de digitalización y automatización industriales. Fuente: elaboración propia

En efecto, las **TIC (Tecnologías de la Información y las Comunicaciones)** cuyo principal substrato hardware es la microelectrónica, han contribuido desde la invención de los microprocesadores por parte de *Federico Faggin* en Intel, a principio de la década de los 70 del siglo pasado, a mejorar la productividad de las empresas que las han usado. La contribución de las TIC y los semiconductores al incremento de productividad de las empresas ha conducido, en los países donde se han implantado suficientemente las TIC en el tejido productivo, a un incremento de la Productividad Total de los Factores (PTF)[2] y, en consecuencia, a un crecimiento de su PIB.

[1] El circuito integrado, entendido como la interconexión monolítica de dispositivos activos (transistores), fue desarrollado a finales de la década de 1950 por Bob Noyce en Fairchild y Jack Kilby en Texas Instrumens. Este último fue galardonado con el premio Nobel de Física en 2.000. Desafortunadamente Bob Noyce falleció antes.

[2] La productividad, ya sea a nivel macroeconómico o de las empresas, puede mejorarse a través de tres canales principales: a) Aumentando **el stock de capital** mediante la inversión en maquinaria y equipo, lo

En este informe nos centramos en una parte específica de los dispositivos semiconductores que son los **chips IA para la Defensa, la Seguridad y el Espacio.** Los chips IA, junto a los algoritmos de IA que ejecutan, de forma similar a como han hecho los equipos electrónicos dotados de microprocesadores con su software correspondiente, también contribuyen a un incremento de la Productividad total de los Factores (PTF).

En qué medida afectará en los próximos diez años la generalización del uso de la IA, tanto en el campo civil como en el de la Defensa y el Espacio, a la Productividad Total de los Factores y, en definitiva, al bienestar de la sociedad, todavía hay pocos datos para sacar conclusiones, pero algunos economistas han construido modelos predictivos. Según el modelo propuesto por (Acemoglou, May 2024) **los efectos de la IA sobre la productividad Total de los factores (PTF) en los próximos años no deberían ser más del 0,66 % del total, o aproximadamente un aumento del 0,064% en el crecimiento de la PTF anualmente.**

El informe se estructura en las siguientes secciones: La sección dos trata de los objetivos generales del capítulo. La sección tres se refiere al refuerzo mutuo entre la IA y las tecnologías de semiconductores, así como al carácter dual de los semiconductores, y específicamente, los chips de IA, también se analizan los impulsores actuales de la industria de semiconductores. La sección cuarta da una panorámica sucinta de la evolución de las tecnologías de semiconductores centrándose en el concepto de nodo tecnológico que ha seguido la conocida como *"Ley de Moore"*, que es una predicción tecnológica/económica, finalizando con la hoja de ruta IRDS (*International Roadmap Devices and Systems*).

La sección quinta introduce las características principales de los chips para la Defensa y el Espacio, haciendo hincapié en los diversos materiales semiconductores empleados, las características del diseño y el proceso tecnológico de este tipo de chips. La sección sexta se centra en los tipos de chips específicos para IA, se comentan las arquitecturas de las CPUs y GPUs, se introducen los conceptos de SoC (System on Chip) y SiP (System in Package) e integración heterogénea (Chiplets..), GPUs, FPGAs y ASICs (Chips dedicados) DNN (Redes Neuronales Profundas), chips neuromórficos, diversos chips comerciales utilizados, se aborda también como muy importante para la Defensa, los chips para computación en el borde, integración de sensores, los chips IA para su uso en Defensa (nube y borde): Proyectos de chips IA para defensa financiados por Agencias Gubernamentales (NASA, ESA, etc.) , por último, en esta sección, se analizan las tendencias futuras de chips IA: Chips de computación en memoria, chips de fotónica integrada, microelectrónica para computación cuántica.

La sección séptima se centra en el modelo de negocio de los semiconductores, basándose en su cadena de valor, proporcionándose algunos datos sobre las previsiones del mercado de semiconductores en los próximos años. La sección octava propone ejemplos de uso avanzado de chips IA para la defensa, como chips para aplicaciones en enjambre de drones, para guiado de misiles hipersónicos, para IoT militar y reconocimiento de imágenes. La última sección, la novena, se centra en geopolítica y semiconductores.

---

que permite a los trabajadores generar una mayor producción con los mismos insumos o con menos; b) Elevando la **calidad del capital humano** mediante mejoras específicas en las habilidades y competencias de los empleados; y c) **Aumentando la eficiencia de la interacción entre el capital y el trabajo**. Esta medida se conoce como **productividad total de los factores (PTF)**. Un aumento de la PTF refleja un mayor nivel de eficiencia en el proceso de producción, en el que se genera más producción sin un aumento proporcional de los insumos. Capta **los efectos del progreso tecnológico (innovaciones**) y la mejora de las prácticas de gestión.

Objetivos

El Objetivo general del informe es el de analizar las tecnologías actuales y las proyecciones futuras en el diseño y fabricación de chips para la IA aplicada a la Defensa (nube y computación en el borde). También se da una panorámica general de las tecnologías de semiconductores, centrándonos en su cadena de valor y sus segmentos principales: Diseño de los chips, Fabricación de estos en el proceso tecnológico para el que se ha diseñado y encapsulado y prueba. De suma importancia es considerar que los chips IA son para aplicaciones de Defensa y del Espacio, lo que implica una serie de consideraciones en el diseño, fabricación, encapsulado y prueba de los chips, relacionados con el cumplimiento de estándares específicos para Defensa y Espacio, es decir, tienen que estar certificados para dichas aplicaciones. En cuanto a objetivos específicos, se consideran, principalmente, los siguientes:

- Analizar las características específicas de los semiconductores adaptados a las necesidades de la Defensa y del Espacio.
- Describir y analizar los requerimientos de los chips para la IA aplicada a la Defensa y al Espacio.
- Analizar la influencia de la geopolítica en la cadena de suministros para la IA aplicada a la Defensa y al Espacio.
- Extraer conclusiones que conduzcan a la elaboración de una serie de recomendaciones concretas, en particular para aumentar, en lo posible, la autonomía estratégica y alineamientos con las estrategias europeas de semiconductores e IA.

IA/Tecnología semiconductores

Círculo virtuoso IA/Tecnología de semiconductores.

Los diferentes algoritmos en los que se basa la IA necesitan una potencia de cálculo importante, lo que implica la utilización de arquitecturas en las que se puedan realizar muchas operaciones aritméticas, como multiplicaciones y acumulaciones de forma simultánea. Esta necesidad impone arquitecturas paralelas, que implementadas en tecnologías de semiconductores con nodos tecnológicos superiores a los 100 nm tendrían superficies prohibitivas en coste y en consumo de energía, por lo que su implementación ha sido posible a medida que los nodos tecnológicos, siguiendo la ley de Moore, han ido descendiendo de tamaño por debajo de 40 nm y la frecuencia de operación ha ido subiendo hasta el orden de 1 GHz., lo que también ha provocado un **consumo considerable de energía**, que es uno de los desafíos de las arquitecturas actuales de los chips de IA.

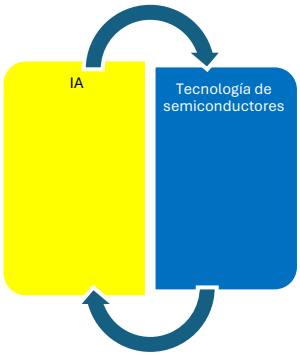


Figura 2. El círculo virtuoso entre la Inteligencia artificial y las tecnologías de semiconductores. Fuente: elaboración propia

Por otro lado, la aplicación de la IA para obtener chips de IA que puedan implementar algoritmos de IA más eficientemente conduce a un círculo virtuoso entre la IA y la tecnología de semiconductores. La IA mejora las **herramientas EDA[3] para el diseño de chips** con mejor rendimiento, potencia, área, costo y "time to market" (PPACt). El ciclo virtuoso de IA que impulsa la propia IA ha arraigado en la industria de los semiconductores, impulsando nuevas arquitecturas y, en algunos casos, reduciendo drásticamente los ciclos de desarrollo a prácticamente la mitad, (por ejemplo de dos años a uno) aumentando considerablemente la productividad del diseño, uno de los factores de coste más elevados para obtener un chip IA.

La IA también se aplica para mejorar el rendimiento de los procesos de fabricación de los chips. A su vez, la **evolución constante de los nodos tecnológicos** permite reducir el tamaño de los chips y aumentar la velocidad de procesado. Con el concepto "More Moore" se consigue aumentar la velocidad de procesado, lo que mejora considerablemente las aplicaciones de la IA.

La industria de semiconductores es al mismo tiempo **impulsora** y **consumidora** de sistemas de IA. Sin lugar a duda, la IA será una característica destacada de la industria de semiconductores durante las próximas décadas.

Uso dual de los semiconductores

El **uso dual de los semiconductores** se refiere a su capacidad de ser utilizados tanto en aplicaciones **civiles** como **militares**. Esta característica los convierte en una **tecnología estratégica** y, a menudo, objeto de regulaciones gubernamentales y restricciones comerciales.

Ejemplos de uso dual:

- **Sensores y comunicaciones**: Chips utilizados en **antenas 5G** o **radares de automóviles** también pueden integrarse en **sistemas de radar militar** o en **equipos de guerra electrónica**.
- **Procesadores y Computación de Alto Rendimiento:** Un procesador de alto rendimiento como los de AMD o Intel puede ser usado en **centros de datos** para el entrenamiento de IA y también en **simulaciones militares** o **criptografía avanzada**.
- **Electrónica de Consumo vs. Militar:** Un chip de inteligencia artificial usado en un **smartphone** puede ser el mismo que se emplea en sistemas de **reconocimiento de imágenes** en drones militares.

En la Fig. 3 se representa un diagrama de bloques de un SoC, concretamente el SAF85xx de NXP Semiconductor que es claramente de uso dual. En efecto, este SoC es un radar para automoción altamente integrado que opera en la banda de 76 a 81 GHz.

Su utilización en automoción se extiende cada vez más, no solo por la utilización de Sistemas Avanzados de Asistencia al Conductor (ADAS) en los vehículos actuales, sino con el advenimiento

[3] Herramientas EDA (Electronic Design Automation) son, principalmente, el conjunto de herramientas software que los ingenieros de diseño microelectrónico utilizan para el diseño, simulación y verificación de los chips. También se consideran herramientas EDA el software para la simulación de procesos tecnológicos y caracterización de dispositivos electrónicos. Igualmente, en la emulación se utilizan equipos hardware programables que interaccionan con las herramientas EDA clásicas como los simuladores y sintetizadores. Actualmente, las herramientas EDA más avanzadas están impulsadas por la IA con objeto de aumentar la productividad del diseño de los chips y su eficiencia.

de los vehículos autónomos donde este tipo de circuitos, y la aplicación de la IA en la computación en el borde, serán ampliamente utilizados.

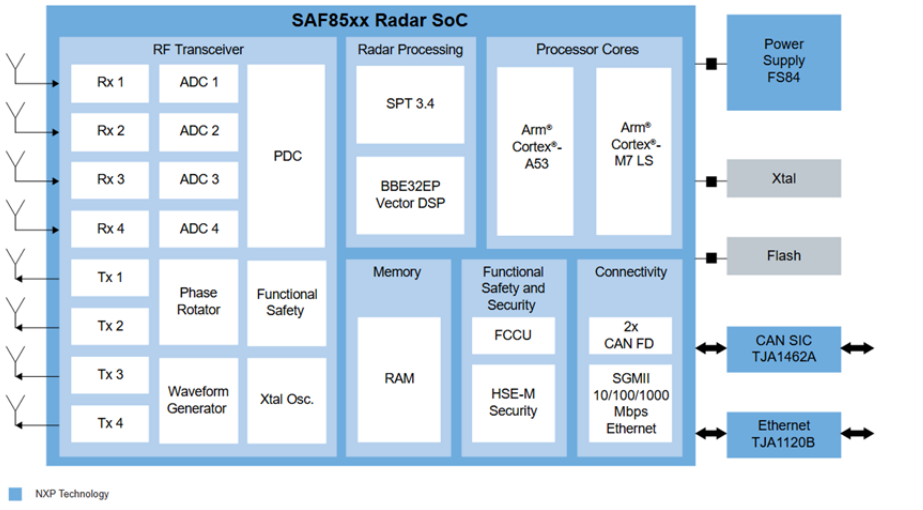


Figura 3. Diagrama de bloques del SoC (System on Chip) SAF85xx Radar para automoción de NSP Semiconductor. Fuente: SAF85xx, Radar SoC | NXP Semiconductors

En la Figura 4, puede verse el aumento progresivo de sensores de radar en los vehículos. Buenos ejemplos son el frenado de emergencia autónomo y el control de puntos ciegos, que utilizan sensores de radar. Este razonamiento sobre la utilización de sensores de radar para ADAS y vehículos autónomos (sin conductor), puede aplicarse también para vehículos militares. La tecnología es la misma. Únicamente los chips además de cumplir la norma de Calidad en Automoción AEC-Q100 , tendrían que adaptarse para cumplir la norma MIL-STD.

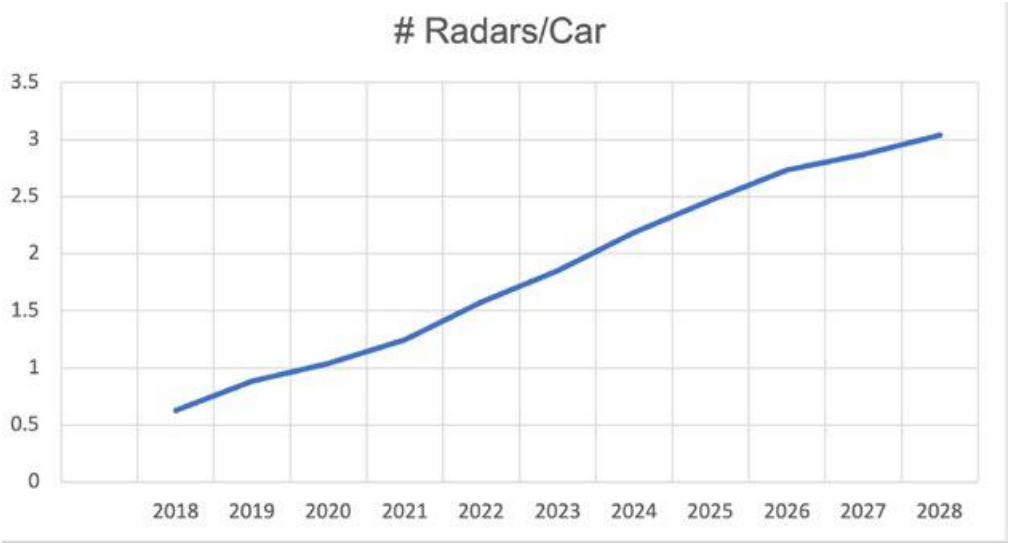


Figura 4. Número medio de sensores de radar por vehículo a lo largo del tiempo. Fuente ADAS Architectures and radar processing

Impulsores actuales de la industria de semiconductores

En cuanto al volumen de producción que permite mantener la capacidad de las fábricas de semiconductores el **motor principal es el sector civil**. En efecto, los impulsores clave en el sector civil, son actualmente:

- **Inteligencia Artificial y Cloud Computing**: Empresas como NVIDIA, Google y Microsoft están liderando la demanda de semiconductores avanzados para IA.
- **Automoción y Electrónica de Consumo**: La industria de los vehículos eléctricos (Tesla, BYD) y los smartphones (Apple, Samsung) requieren chips más potentes.
- **Computación Cuántica y 5G/6G**: Avances en telecomunicaciones impulsan el desarrollo de semiconductores especializados.

En Defensa muchos chips tienen un volumen de producción reducido, comparado con el sector civil, además de los requerimientos específicos de una mayor fiabilidad y un funcionamiento en un rango de temperaturas más extenso que el rango de temperaturas comercial, lo que conduce a un precio unitario mayor por chip, en igualdad de funcionalidad con el sector civil pues los costes fijos deben distribuirse en un menor número de unidades. Sin embargo, el sector de la Defensa es, claramente, un impulsor estratégico:

- **Autonomía en chips nacionales**: Países como EE.UU. y China están desarrollando fábricas locales para reducir la dependencia extranjera.
- **Supercomputadoras y Ciberseguridad**: Aplicaciones militares de alto secreto dependen de chips avanzados para simulaciones y criptografía.
- **Armas Autónomas y Guerra Electrónica**: Sistemas como los drones de combate o los misiles guiados requieren semiconductores especializados

El **sector civil** lidera la demanda en volumen, pero el **sector defensa** es un impulsor estratégico clave. Gobiernos como el de EE.UU. están invirtiendo miles de millones en **fábricas locales (TSMC, Intel, Samsung en Arizona)** para garantizar la producción de chips sin depender de Asia, lo que demuestra el papel clave del sector militar en la planificación a largo plazo.

En definitiva, aunque los consumidores y la IA lideran el crecimiento de la industria, la **geopolítica y la defensa están definiendo las estrategias de producción y seguridad de suministro**

Evolución de la Tecnologías de Semiconductores

Desde la invención del transistor en los Bell Labs en 1947 hasta la actualidad, han transcurrido 78 años, en los que la microelectrónica, la rama de la electrónica que se ocupa de los circuitos y componentes electrónicos miniaturizados a escala micrométrica (µm) y nanométrica (nm), ha sido la base física sobre la que se ha edificado la digitalización que ha propiciado las telecomunicaciones modernas, tanto cableadas como inalámbricas, la informática, Internet, la robótica y multitud de aplicaciones en la mayoría de actividades industriales, de salud, de ocio, financieras y económicas, que han conducido a un importante aumento de la productividad.

Estos avances se han producido por la constante inversión en I+D e innovación del diseño y procesos tecnológicos de los semiconductores, que han permitido aumentar el número de transistores por unidad de superficie de forma casi continua desde la invención del circuito integrado monolítico en 1959 por Kelby y Noyce.

Ley de Moore

En 1965, el cofundador de Intel, Gordon Moore, predijo que el número de transistores en un chip se duplicaría aproximadamente cada dos años, con un aumento mínimo en el costo. Esta predicción se conoce como "Ley de Moore" (véase la figura 5).

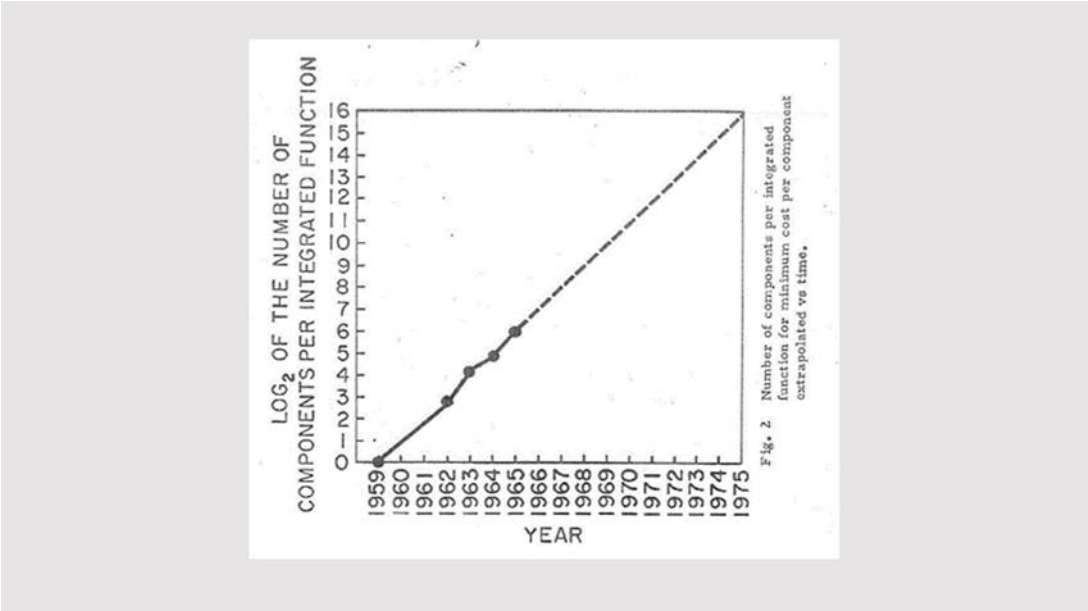


Figura 5. Gráfico original de Gordon Moore en el que en ordenadas se expresa el logaritmo en base 2 del "número de componentes por función integrada" y en abscisas el año. Fuente: (Moore, April 1965)

Como comenta (Kelleher, 2022): "*Durante más de 40 años, los ingenieros de Intel han innovado continuamente para integrar cada vez más transistores en chips cada vez más pequeños y mantener el ritmo de la Ley de Moore. A mediados y finales de la década de 2010, como ya lo ha hecho en varias ocasiones anteriores, la industria predijo que «la Ley de Moore ha muerto». Parafraseando un dicho famoso, creo que los informes sobre la muerte de la Ley de Moore son muy exagerados. La innovación no ha muerto, y mantendremos la Ley de Moore como siempre lo hemos hecho, mediante la innovación: innovación en procesos, en el empaquetado y en la arquitectura. Será un desafío, como siempre, e Intel está a la altura*".

En efecto, la Ley de Moore se ha ido cumpliendo gracias al esfuerzo en innovación que ha efectuado, no solo Intel, sino todos los fabricantes de semiconductores, aunque en Intel ha sido su caballo de batalla durante las últimas cuatro décadas.

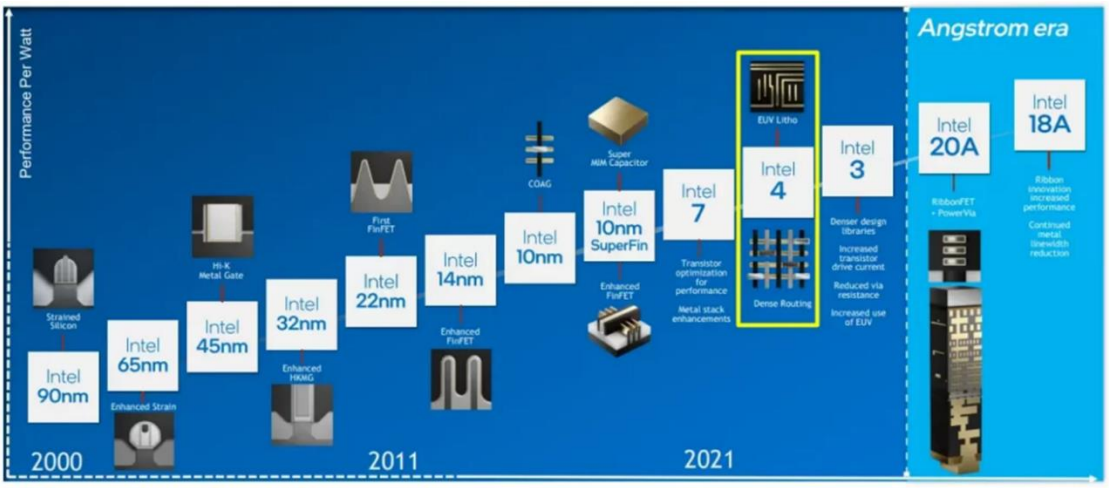


Figura 6. Muestra las innovaciones en proceso realizadas por Intel a lo largo del tiempo. En el eje de ordenadas se refleja las "Prestaciones por watio". Fuente: (Kelleher, 2022)

Las innovaciones de proceso a lo largo del tiempo pueden observarse en la figura 6. A lo largo de los años han ido introduciéndose verdaderas innovaciones que han permitido el cumplimiento de la Ley de Moore. Con inventos como la tecnología de puerta metálica de alta k[4], los transistores 3D tri-gate y el silicio tensado[5], Intel ha desarrollado constantemente tecnologías innovadoras para mantenerse al día con la Ley de Moore. A finales de la década de 2000, a medida que las dimensiones físicas seguían reduciéndose, la industria se dio cuenta de que se necesitaban áreas adicionales de innovación, como la ciencia de los materiales, la nueva arquitectura de procesos y la co-optimización de la tecnología de diseño (DTCO), para seguir el ritmo.

[4] A medida que se han ido escalando los dispositivos hacia dimensiones más reducidas, el dieléctrico de puerta que separa la puerta del canal MOS ha ido disminuyendo su grosor para mantener el rendimiento del dispositivo. Sin embargo, el dieléctrico usado en los procesos ha sido siempre el SiO_2 y este material tiene un límite si el espesor de la capa de SiO_2 es inferior a ~1 nm, la corriente de fuga debida al efecto túnel cuántico empieza a predominar, lo que causa graves problemas en el consumo de energía y el rendimiento del dispositivo. Este obstáculo técnico se ha superado reemplazando SiO_2 con aislantes que poseen constantes dieléctricas altas (alto-κ). Con dieléctricos de alto-κ, el espesor dieléctrico se puede aumentar con la misma capacitancia, suprimiendo así la corriente de fuga. Actualmente, los **dieléctricos de alto-κ** preferidos son HfO_2 , el ZrO_2 y el Al_2O_3

[5] La **mejora de la movilidad** obtenida mediante la aplicación de la deformación adecuada (El silicio tensado es una capa de silicio en la que los átomos de silicio se estiran más allá de su distancia interatómica normal.) proporciona una mayor velocidad de los portadores de carga en los canales MOS y una mayor corriente de excitación, respectivamente, con la misma tensión de alimentación y el mismo espesor de óxido de puerta. Esto implica que se deben utilizar óxidos de puerta más gruesos o una tensión de alimentación más baja para una corriente de excitación fija, lo que mitiga la relación de compensación entre la corriente de excitación, el consumo de energía y los efectos de canal corto.

También el encapsulado ha seguido evolucionando de forma paralela a la evolución del proceso tecnológico.

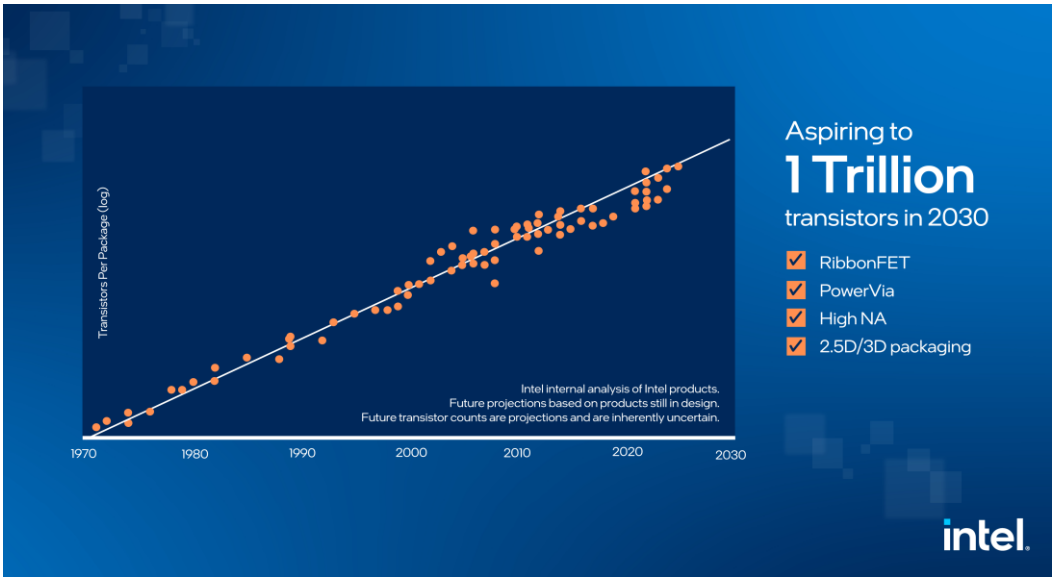


Figura 7. Ley de Moore, número de transistores por dispositivo: Pasado, presente y futuro. Fuente: : (Kelleher, 2022)

En la figura 7, puede observarse el número de transistores por dispositivo a lo largo de estos años. Este número, que indica que se dobla cada par de años, de acuerdo con lo que predice la Ley de Moore, solo se ha podido lograr con una apuesta total por la innovación. Durante los primeros 40 años, las ganancias provinieron principalmente de las innovaciones en el proceso tecnológico. En el futuro, las ganancias provendrán de las innovaciones tanto en el proceso como en el encapsulado. Intel y otros fabricantes consideran que los procesos tecnológicos seguirán ofreciendo mejoras históricas en densidad, mientras que las tecnologías de apilamiento 2D y 3D brindan a los arquitectos y diseñadores más herramientas para aumentar la cantidad de transistores por dispositivo.

En resumen, según (Kelleher, 2022): "cuando consideramos todas las innovaciones en procesos y empaquetado avanzado, hay numerosas opciones disponibles para seguir duplicando la cantidad de transistores por dispositivo al ritmo que demandan nuestros clientes. La Ley de Moore solo se detiene cuando se detiene la innovación, y la innovación continúa sin cesar en Intel en procesos, empaquetado y arquitectura. Seguimos firmes en nuestra aspiración de entregar aproximadamente 1 billón de transistores en un solo dispositivo para 2030".Intel-Moores-Law-Investor-Meeting-Paper-final.pdf

Concepto de Nodo Tecnológico y su evolución

El término **"nodo" en la tecnología de semiconductores** se originó a partir de la medida de la mitad del paso de la celda de memoria de un chip DRAM[6].

[6] DRAM: Memoria de Acceso Aleatorio Dinámica. Es una memoria de lectura y de escritura de acceso aleatorio. El tiempo de acceso es rápido y tiene la mayor densidad, ya que un bit se almacena en un único transistor. Para mantener la información, la memoria necesita "refrescarse" (un reloj debe actuar constantemente) lo cual implica que el consumo de energía sea considerable. Es una memoria volátil, es decir, que, al apagar la alimentación, la memoria no retiene la información.

Dado que las mayores capacidades se lograron a través de una mayor densidad, fue la DRAM la que se convirtió en el impulsor de la escalabilidad tecnológica. El nombre del nodo generalmente se alinea con la mitad del paso, conocido como “medio paso” (half-pitch), del área activa en la matriz de celdas de memoria. El medio paso de la celda de memoria en un chip DRAM es una dimensión crucial que influye en la capacidad total del chip. Ver Figuras 8 y 9.

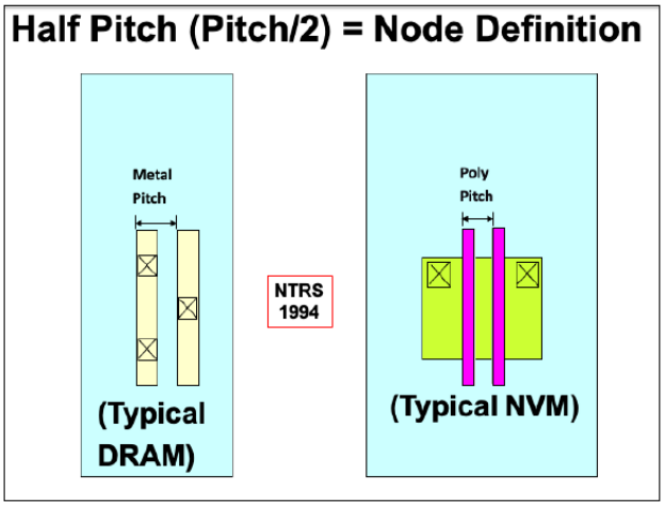


Figura 8. Definición de nodo tecnológico: Mitad del paso (Half pitch) del área activa en una celda de memoria DRAM. Fuente: 2022IRDS_ES.pdf

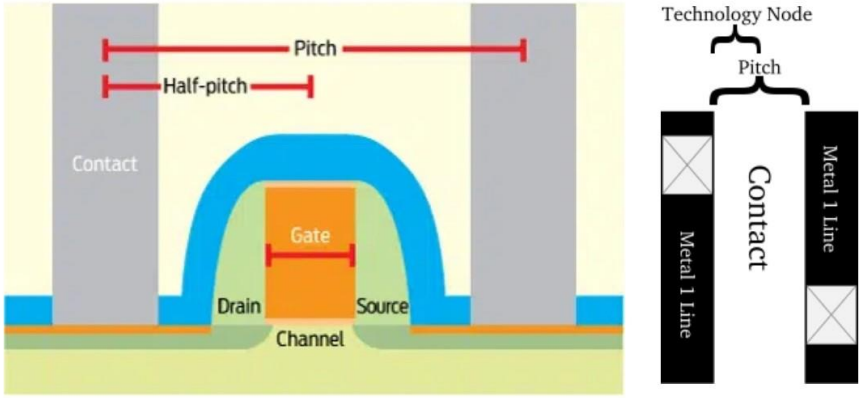


Figura 9. Definición de nodo tecnológico: Mitad del paso (Half Pitch) visto en sección. Fuente: All about Technology Node | LinkedIn

La definición de nodo tecnológico es muy importante porque determina propiedades esenciales del chip, como frecuencia de operación, densidad de este por mm^2, rendimiento de la oblea y otras.

A medida que los fabricantes de semiconductores siguieron aplicando la Ley de Moore (la observación de que la cantidad de transistores en un microchip se duplica aproximadamente cada dos años), se encontraron con obstáculos importantes a finales del siglo XX.

Los MOSFET planares tradicionales (véase la figura 10 izquierda) que se construían en capas superpuestas sobre el sustrato de silicio, se estaban acercando a sus límites físicos. Estos límites se debían principalmente a problemas relacionados con la corriente de fuga y la disipación de potencia, que se volvieron cada vez más problemáticos a medida que los transistores seguían reduciéndose de tamaño. Para superar estos desafíos, los ingenieros y científicos comenzaron a

explorar diseños alternativos de transistores. Uno de los avances se produjo en forma de ***FinFET*** (*transistores de efecto de campo Fin*). Véase figura 10 (derecha).

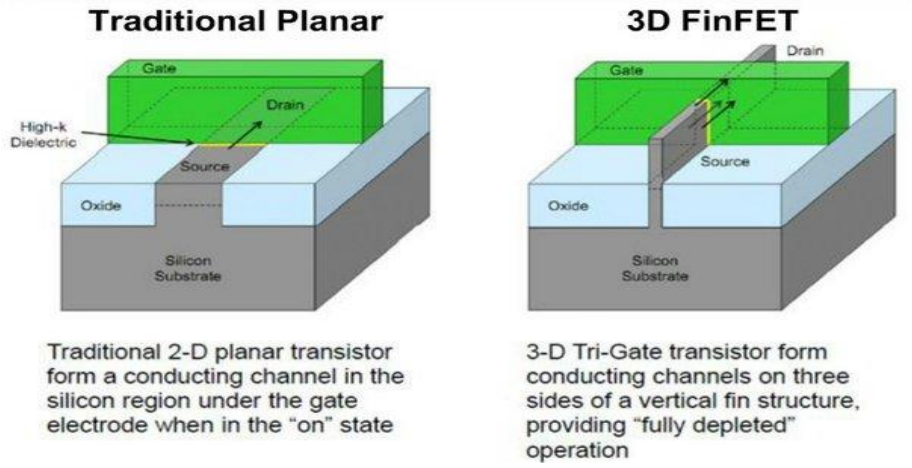


Figura 10. Transistor MOSFET tradicional Planar (izda.) y Transistor FinFET (3D). Fuente: Semiconductor Engineering

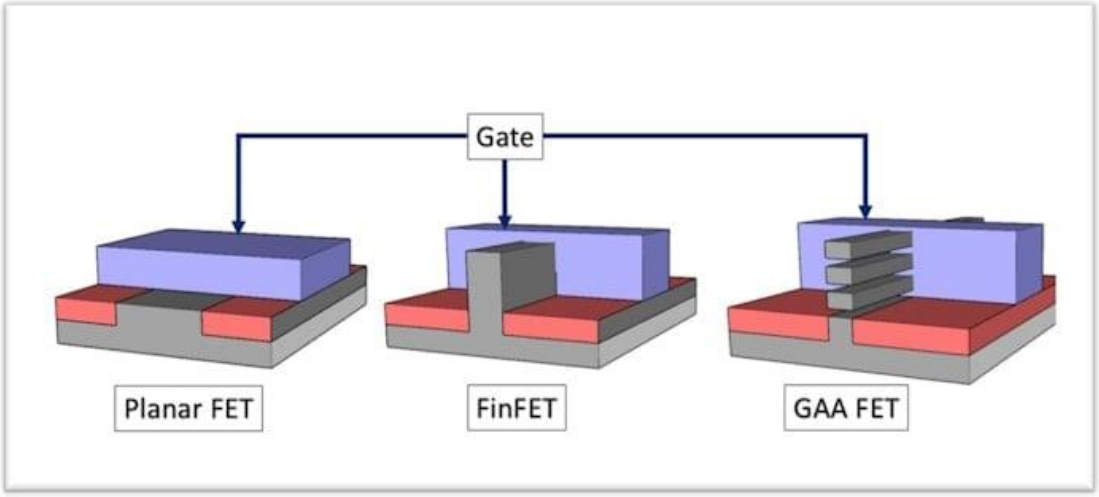


Figura 11. Evolución de la estructura de los MOSFET: Planar, FinFET y GAA FET. Fuente: Semiconductor Engineering

Mejoras sobre la tecnología planar:

- Menor corriente de fugas
- Mejora de las prestaciones: Velocidad de conmutación más alta.
- Mejora de le eficiencia energética.
- Continuación del escalado -> **¡Se sigue salvando la Ley de Moore!**

Así pues, la reducción de las dimensiones del nodo tecnológico no ha sido simplemente una reducción de dimensiones clásica, sino que ha habido que innovar en materiales, procesos y dispositivos. Véase la Fig. 11 en la que se ve que a partir de 22nm, se abandona la tecnología planar tradicional y se introduce el FinFET, para evolucionar a partir de 7nm a los dispositivos FET GAA (Gate-All-Around). FinFET y GAA son dispositivos 3D, en contraposición con la tecnología Planar.

La reducción de dimensiones de los dispositivos ha posibilitado integrar muchos más dispositivos (circuitos más complejos, verdaderos sistemas) en la misma área por lo que a media que disminuía el tamaño del nodo el coste del diseño del chip aumentaba considerablemente pues el diseño suele ser un sistema más complejo. Al mismo tiempo, para una misma oblea de 300 mm. a medida que se escalaba hacia abajo (para la misma complejidad de circuito) se obtenían precios unitarios más bajos (salían más chips por oblea). Sin embargo, los costes de fabricación de los prototipos (NRE) son más elevados pues, la amortización de equipos, así como el coste de las máscaras (procesos más complejos y dimensiones más pequeñas) son mayores.

En la figura 12 puede apreciarse el aumento del coste del diseño del chip (Ingeniería de diseño + NREs (fabricación, encapsulado y prueba de prototipos) en función del nodo tecnológico utilizado.

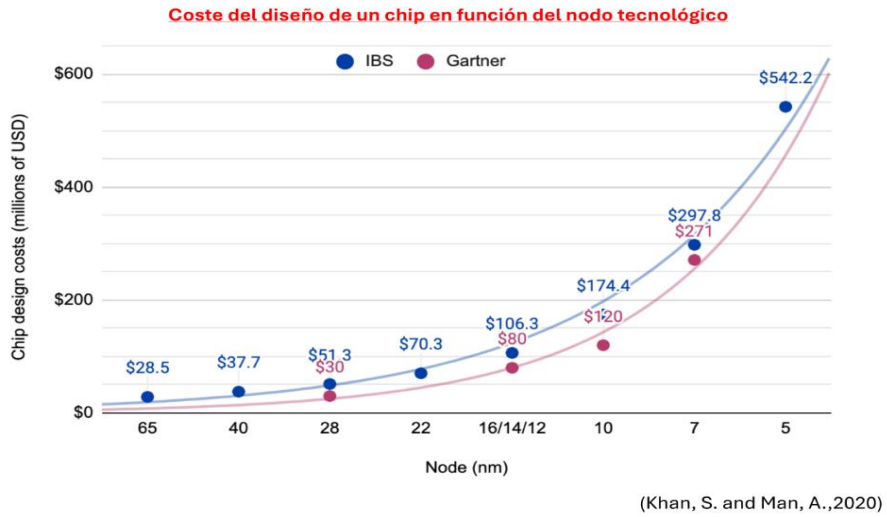


Figura 12. Coste del diseño de un chip en función del nodo tecnológico. Fuente: (Khan & Mann, April 2020)

Las mejoras en prestaciones de los nodos más avanzados con respecto a la tecnología de 90 nm pueden observarse en la figura 13. En el caso de la eficiencia del transistor, los del nodo de 5nm son 50 veces las del nodo de 90 nm.

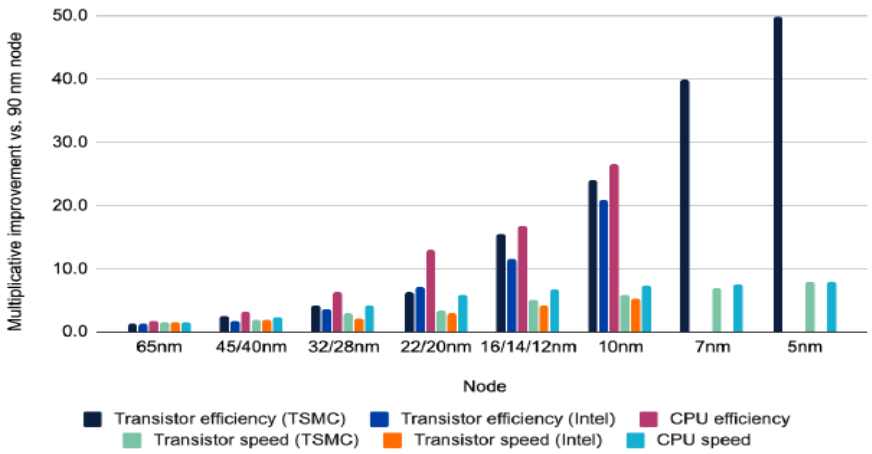


Figura 13. Mejoras en la eficiencia y la velocidad medidas en comparación con el nodo de 90 nm. Fuente: (Khan & Mann, April 2020)

Hoja de ruta IRDS

La **Hoja de Ruta Internacional para Dispositivos y Sistemas (IRDS)** define las tendencias clave en tecnología de semiconductores (véase la figura 14), incluyendo los tres conceptos siguientes:

1**. "More Moore" (MM)**

- Se refiere al escalado continuo de la tecnología CMOS según la Ley de Moore, que predice que el número de transistores en un circuito integrado se duplica aproximadamente cada dos años.
- Enfoque: Mejorar el rendimiento, reducir el consumo de energía y aumentar la densidad de transistores.
  Ejemplos de tecnologías:
- Transistores FinFET y Gate-All-Around (GAA)
- Litografía ultravioleta extrema (EUV)
- Arquitecturas avanzadas de transistores 3D

2**." More than Moore"**

- Se extiende más allá del simple escalado de transistores al integrar funcionalidades no digitales (p. ej., analógicas, RF, MEMS, fotónica) en dispositivos semiconductores.
- Enfoque: Diversificar las funcionalidades para permitir mejoras a nivel de sistema en lugar de simplemente aumentar la densidad de transistores. Ejemplos de tecnologías:
- Sistema en chip (SoC) y sistema en paquete (SiP)
- MEMS (sistemas microelectromecánicos)
- Circuitos de radiofrecuencia y ondas milimétricas para 5G/6G
- Fotónica integrada para comunicaciones ópticas

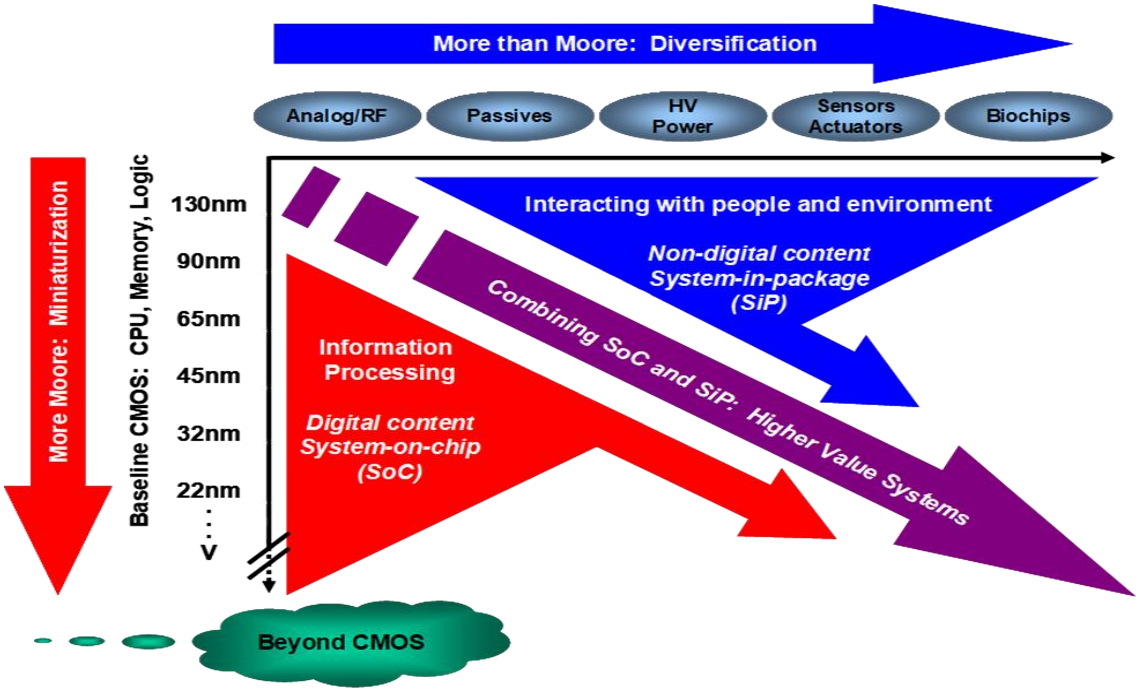


Figura 14. Hoja de ruta de los sistemas integrados (IRDS): Ley de Moore (More Moore) y More than Moore (MtM) (Diversificación). Combinando en un solo encapsulado SoC y MtM en SiP para obtener sistemas de más valor. Fuente: (IRDS, 2023)

3.” Beyond CMOS”

- Se refiere a las tecnologías post-CMOS que exploran paradigmas informáticos completamente nuevos, más allá de los transistores tradicionales basados en silicio.
- Enfoque: Superar las limitaciones físicas y de potencia del escalado de CMOS mediante el aprovechamiento de materiales y arquitecturas informáticas alternativas.
  Ejemplos de tecnologías:
- Computación cuántica (cúbits)
- Computación neuromórfica (arquitecturas inspiradas en el cerebro)
- Espintrónica (utilizando el espín del electrón para el procesamiento de datos)
- Materiales 2D (p. ej., grafeno, dicalcogenuros de metales de transición[7])

Estos tres enfoques se complementan: «More Moore» impulsa el escalado convencional, «More than Moore» integra nuevas funcionalidades y «Beyond CMOS» investiga futuros sustitutos de las tecnologías informáticas tradicionales.

“More Moore”

Figura 15. Big Data y Datos Instantáneos. Fuente: (IRDS MM, 2023)

La Ley de Moore y los conceptos de "More Moore" para el escalado de transistores permiten tecnologías como los sistemas en chip (SoC), con requisitos de rendimiento que implican un consumo excesivo de energía y un ancho de banda de interconexión limitado. La interacción entre big data y datos instantáneos presenta desafíos y limitaciones que el IRDS aborda en la hoja de ruta de “More Moore”. El escalado de sistemas, gracias al escalado de Moore, se ve cada vez más dificultado por la escasez de recursos como la energía y el ancho de banda de datos. Esto se ha vuelto aún más complejo debido a los requisitos de una interacción fluida entre el big data y los datos instantáneos (Ver Fig. 15). La generación instantánea de datos requiere dispositivos de consumo ultra bajo con función de "siempre activo" (“*always-on*”), así como dispositivos de alto rendimiento que puedan generar los datos al instante. El big data requiere abundantes recursos de computación, ancho de banda de comunicación y memoria para generar

[7] Los dicalcogenuros de metales de transición (TMD) son materiales bidimensionales con notables propiedades semiconductoras y elevados coeficientes de absorción óptica.

el servicio y la información que necesitan los clientes. La necesidad de datos para computación se aceleró con el **aumento de las cargas de trabajo de IA**.

El Equipo de *More Moore International Focus (IFT)* de la Hoja de Ruta Internacional de Dispositivos y Sistemas (IRDS) proporciona los requisitos físicos, eléctricos y de confiabilidad para las tecnologías de lógica y memoria con el fin de mantener el escalado de potencia, rendimiento, área y costo (PPAC) para aplicaciones en el borde y en la nube. Esto se realiza en un horizonte temporal de 15 años para la fabricación convencional/de alto volumen (HVM).

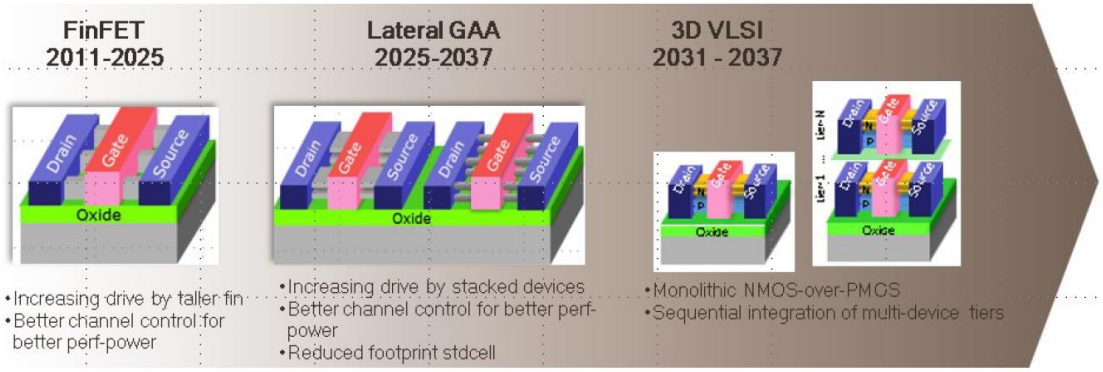


Figura 16. Evolución de las arquitecturas de dispositivos en la hoja de ruta de "More Moore" del IRDS. Fuente: (IRDS MM, 2023)

Es probable que el finFET se mantenga hasta 2025. En 2022, ya se inició la transición a **dispositivos GAA laterales** debido a las limitaciones del rendimiento limitado de una sola aleta (finFET), como resultado de la despoblación de aletas y el escalado del ancho de aleta (saturando el escalado de la Lgate para mantener el control electrostático). La mitigación de la capacitancia parásita, el ancho de accionamiento efectivo (Weff) y la integración de la puerta metálica de reemplazo (RMG) siguen siendo el desafío para mantener la adopción de GAA en múltiples nodos. La evolución proyectada de las arquitecturas de dispositivos se muestra en la Figura 16. Se proyecta que el FET complementario (CFET) será la evolución posterior de los GAA laterales en formato 3D, donde los dispositivos N se apilarán sobre los dispositivos P.

"More than Moore"

"More than Moore" se refiere a la incorporación en los dispositivos de funcionalidades que no necesariamente se escalan según la Ley de Moore, pero que **aportan valor añadido de diferentes maneras**. El enfoque **"More than Moore"** permite que las funcionalidades no digitales (por ejemplo, comunicación por radiofrecuencia, control de potencia, componentes pasivos, sensores, actuadores) migren del nivel de la placa del sistema al encapsulado (SiP) o al chip (SoC).

Además, la integración cada vez más íntima de software integrado complejo en los SoC y SiP significa que el software también podría tener que convertirse en un tejido en consideración que afecte directamente al escalado del rendimiento. (Co-diseño Hw/Sw)

El objetivo de **"More than Moore"** es ampliar el uso de la tecnología basada en silicio desarrollada en la industria de la microelectrónica para **proporcionar nuevas funcionalidades no digitales**. A menudo aprovecha las capacidades de escalado derivadas de los desarrollos de **"More Moore" para** incorporar funcionalidad digital y no digital en sistemas compactos y, eventualmente, en sistemas de sistemas.

“More than Moore” enfatiza las innovaciones en otros aspectos del diseño y la fabricación de chips para satisfacer las crecientes demandas de diversas aplicaciones. (IRDS MtM, 2023)

“Beyond CMOS”

A medida que el escalado dimensional de los CMOS se acerca a **límites fundamentales**, se están explorando **nuevos dispositivos de procesamiento de información y microarquitecturas para funciones existentes y nuevas**.

Esto está impulsando el interés en nuevos dispositivos para el procesamiento de información y la memoria, nuevas tecnologías para la **integración heterogénea** de múltiples funciones y nuevos paradigmas para la arquitectura de sistemas.

Los nuevos paradigmas informáticos y las nuevas aplicaciones (por ejemplo, big data, **Internet de las cosas (IoT), inteligencia artificial, sistemas autónomos**, computación a exaescala) introducen mayores requisitos de rendimiento y eficiencia, que son cada vez más difíciles de cumplir para las saturadas tecnologías de **More Moore**.

Las tecnologías **Beyond-CMOS** pueden proporcionar los dispositivos, procesos y arquitecturas necesarios para la nueva era de la informática. En Beyond-CMOS se contemplan tecnologías como las “Memorias emergentes”, los nanocables, la computación cuántica, etc. (ver figura 17). Trata de superar las limitaciones de la Ley de Moore, en eficiencia y prestaciones, mediante la combinación de materiales, dispositivos y arquitecturas emergentes.

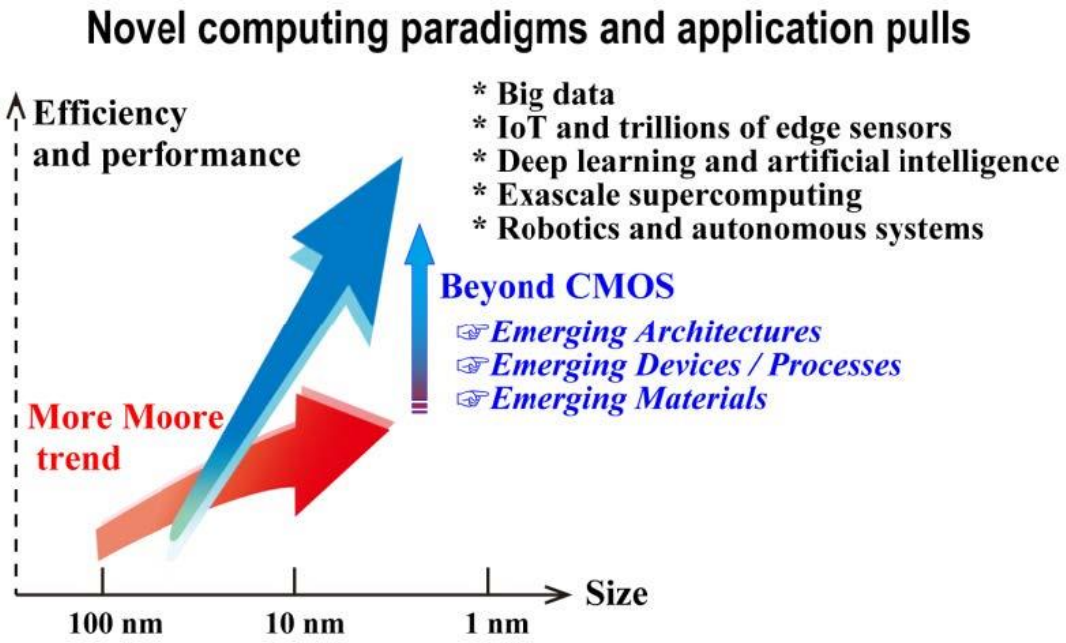


Figura 17. Relación entre More Moore, Beyond CMOS y nuevos paradigmas computacionales y sus aplicaciones. Fuente: (IRDS BC, 2023)

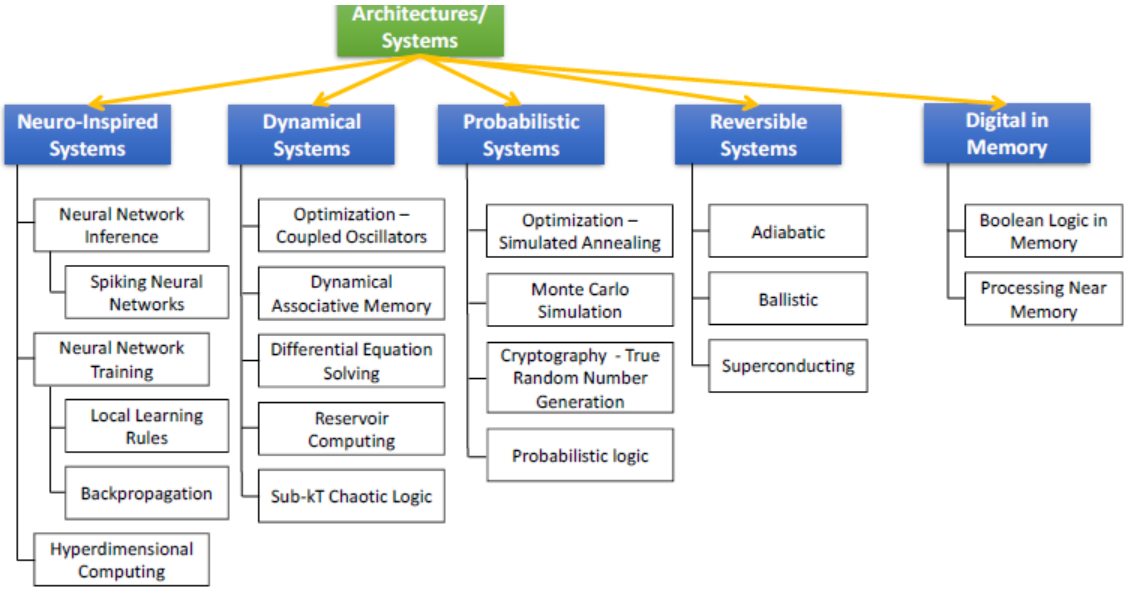


Figura 18. Arquitecturas/Sistemas para nuevos paradigmas computacionales enumerados por IRDS en la tendencia "Beyond CMOS". Requieren co-diseño con dispositivos emergentes. Fuente: (IRDS BC, 2023)

Las nuevas arquitecturas y sistemas que impulsan las nuevas aplicaciones como la robótica y los sistemas autónomos, la IA y el Aprendizaje profundo, el IoT y la IA en el borde con millones de sensores, la supercomputación, etc. exigen nuevos dispositivos que superen las limitaciones del final de la Ley de Moore, los dispositivos emergentes (dispositivos sinápticos, dispositivos estocásticos y dispositivos osciladores) *(IRDS BC, 2023).*

SoC y SiP

Los sistemas electrónicos actuales pueden integrarse en un único chip. En este caso se habla de un **SoC** (System on Chip) (Ver figura 19). Sus principales características son:

- Integra múltiples componentes (CPU, GPU, memoria, E/S, etc.) en una única pieza de silicio (Circuito integrado monolítico).
- Diseñado para aplicaciones compactas, de alto rendimiento y bajo consumo.
- Optimizado para tareas específicas (p. ej., smartphones, dispositivos IoT, sistemas embebidos). Su fabricación se basa generalmente en procesos avanzados de fabricación CMOS.
- Ofrece comunicación de alta velocidad entre componentes gracias a interconexiones cortas.
- Escalabilidad y flexibilidad limitadas en comparación con las soluciones multichip.

Ejemplos de SoC

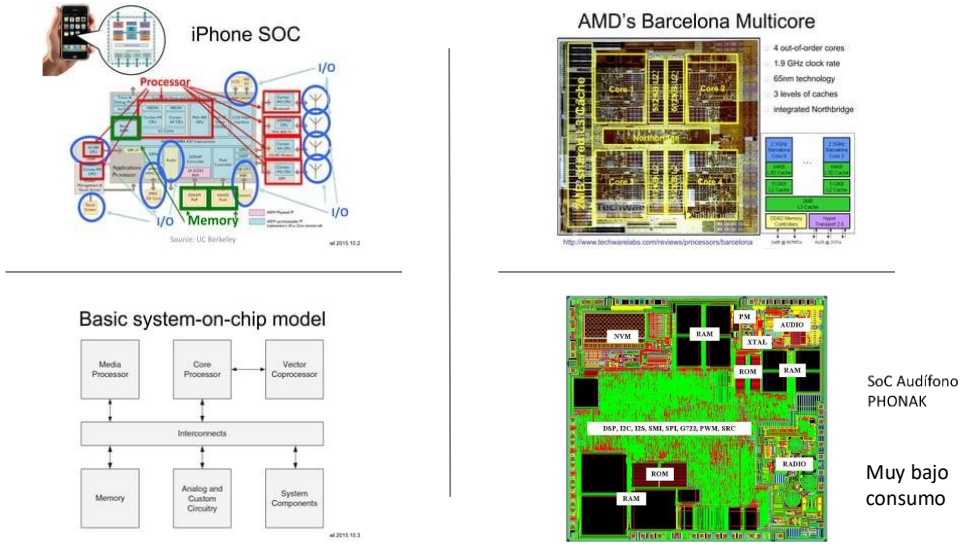


Figura 19. Cuatro ejemplos de SoC: a) Diagrama de bloques de SoC en iPhone; b) Fotomicrografía del Chip "Barcelona Multicore" de AMD (4 cores.)Fuente: AMD Barcelona Architecture Launch: Native Quad-Core | HotHardware ; c) Un diagrama de un modelo de SoC básico (Procesador, Memoria, partes analógicas y ADCs/DACs,…) y d)Fotmicrografía SoC audífono (mixed analog-digital) Phonak de muy bajo consumo.

Características principales de un SiP (Ver figura 20):

- Combina múltiples dados (chips) independientes en un único encapsulado.
- Puede integrar diferentes tecnologías (p. ej., CMOS, RF, MEMS, fotónica, electrónica de potencia).
- Permite un diseño modular y flexibilidad para integrar diversas funcionalidades.
- Se utiliza en aplicaciones como 5G, automoción, aceleradores de IA y wearables. Permite utilizar técnicas de encapsulado avanzadas (p. ej., integración 2.5D/3D, vías a través de silicio).
- Es más fácil combinar componentes que en un SoC.

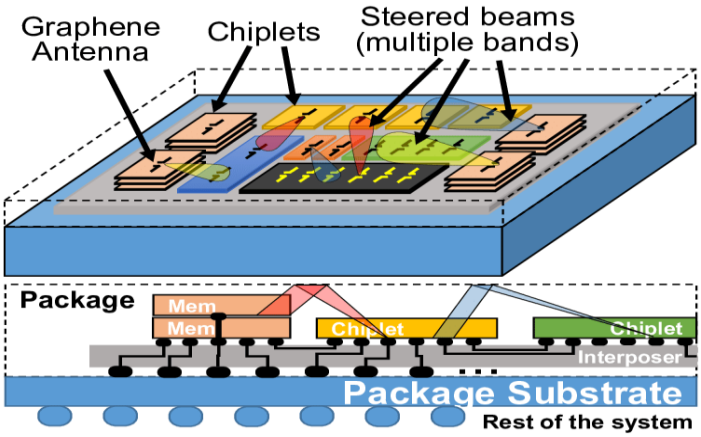


Figura 20. Representación de un SiP que incluye chips, y MEMs.

Característica	SoC (System on Chip)	SiP (System in Package)
Integración	Todos los componentes en una único “dado”	Múltiples “dados” independientes en un encapsulado
Tamaño	Más compacto	Ligeramente más grande debido a múltiples “dados”
Rendimiento	Comunicación de alta velocidad entre componentes	Ligeramente menor debido a la comunicación entre chips
Eficiencia energética	Menor consumo de energía debido a interconexiones cortas	Consumo de energía ligeramente mayor debido a enlaces entre chips
Flexibilidad	Menos flexible; todo el diseño debe optimizarse en conjunto	Más flexible; permite combinar diferentes tecnologías de chip
Complejidad de fabricación	Requiere fabricación avanzada de semiconductores (p. ej., FinFET, EUV)	Utiliza técnicas de encapsulado avanzadas (p. ej., apilamiento 2.5D, 3D)
Aplicaciones	Smartphones, sistemas embebidos, IoT, chips de IA	5G, automoción, aceleradores de IA, módulos de RF
Escalabilidad	Limitada debido a la naturaleza monolítica	Alta escalabilidad mediante la adición de más matrices

Tabla 1. Comparación entre SoC (System on Chip) y SiP (System in Package)

El SoC es ideal para aplicaciones de alto rendimiento y bajo consumo que requieren una integración estrecha, mientras que el SiP proporciona mayor flexibilidad y modularidad para diversos diseños de sistemas.

En algunas aplicaciones, muchas veces, interesa más diseñar un sistema utilizando el concepto de Chiplet que diseñar el sistema en un único circuito integrado monolítico (SoC). Los **chiplets** son procesadores segmentados. En lugar de consolidar cada componente en un solo chip (lo que se conoce como enfoque monolítico), se fabrican secciones específicas como chips separados. Estos chips individuales se ensamblan en un solo encapsulado mediante un complejo sistema de conexión. *(Evanson, 2025)*

Un **chiplet** es un pequeño circuito integrado (CI) monolítico especializado, diseñado para combinarse con otros chiplets dentro de un único encapsulado y formar un sistema completo (Ver Fig. 21). En lugar de diseñar un gran chip monolítico, se pueden fabricar chiplets más pequeños por separado e integrarlos en un encapsulado, lo que mejora el rendimiento y la escalabilidad.

Aspectos clave de los chiplets:

- Enfoque modular para el diseño de chips.
- Optimizado para funciones específicas (p. ej., CPU, GPU, memoria, aceleración de IA).
- Permite la reutilización de diseños probados, lo que reduce los costes de desarrollo.
- Se suelen conectar mediante interconexiones de alta velocidad como UCIe, EMIB o TSV.
- Se encuentra en la computación de alto rendimiento, aceleradores de IA y procesadores avanzados (p. ej., AMD, Intel, NVIDIA).

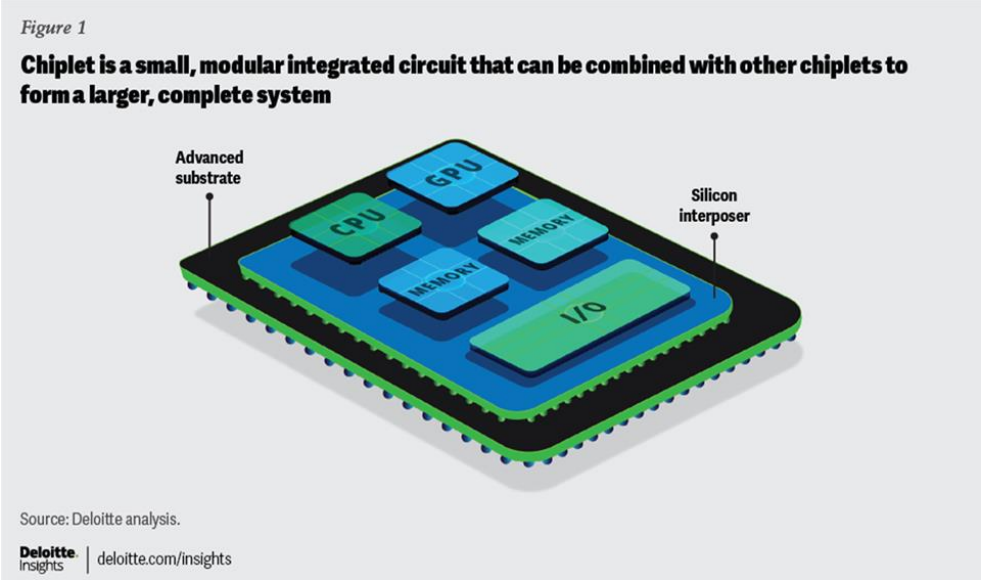


Figura 21. Sistema integrado (SiP) que monta varios chiplets (CPU, GPU,Memorias, I/O). Fuente: Deloitte análisis.

La **Integración Heterogénea** (HI) se refiere al proceso de combinar diversos materiales semiconductores, tecnologías o componentes funcionales en un único encapsulado o sistema. Esta integración permite una funcionalidad avanzada que no se puede lograr con un solo circuito integrado monolítico.
Aspectos Clave de la Integración Heterogénea:

- Combina diferentes tecnologías (p. ej., CMOS, MEMS, fotónica, RF, electrónica de potencia).
- Puede incluir diferentes materiales (p. ej., silicio, GaN, InP, grafeno).
- Se utiliza en encapsulados 2.5D/3D, SiP y sistemas semiconductores avanzados.
- Se encuentra en aplicaciones como 5G, IA, automoción, defensa, aeroespacial y dispositivos médicos.

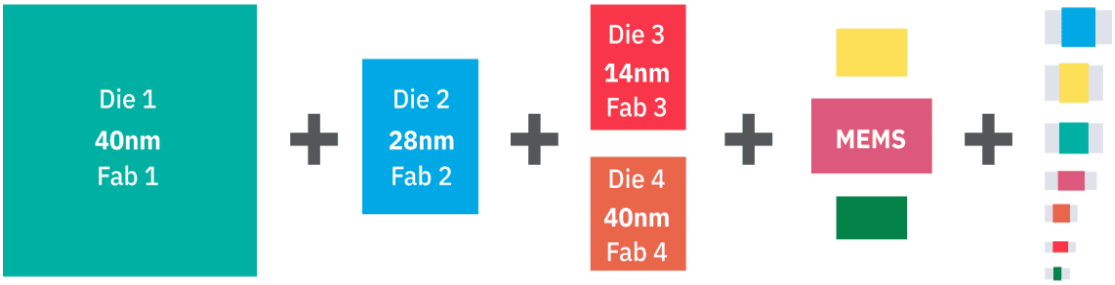


Figura 22a. Chips y MEMs que forman los componentes del sistema integrar Fuente: Heterogeneous Integration (HI) | ASE

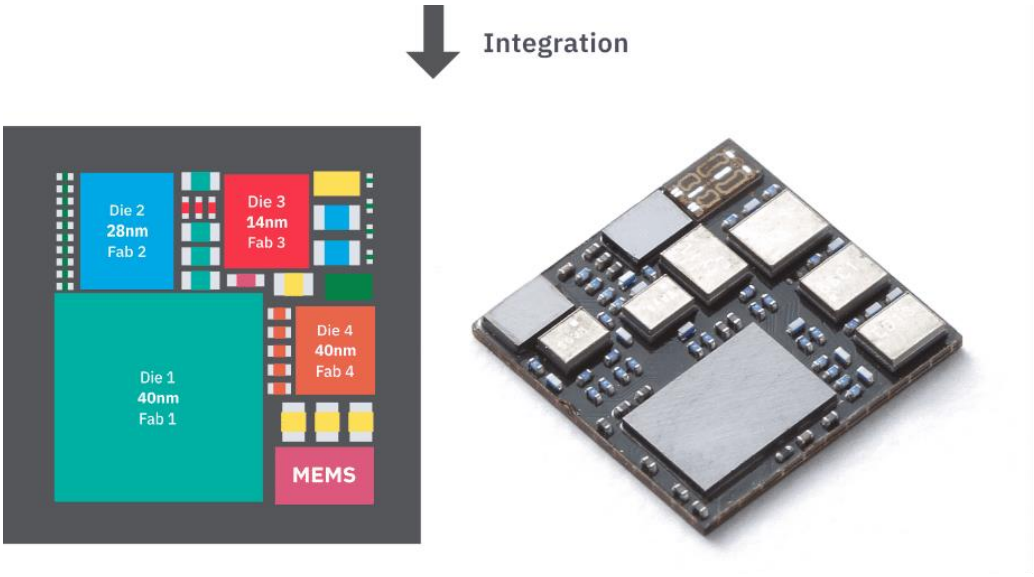


Figura 22b. El sistema con los componentes ya interconectados (integración heterogénea) para encapsularse en un SiP. Fuente: Heterogeneous Integration (HI) | ASE

Las figuras 22a y 22b muestran un esquema de los diferentes circuitos integrados monolíticos y MEMs antes de interconectarse (Fig. 22a) y después de la interconexión para encapsularse en un SiP (Fig. 22b). La figura 23 es una integración heterogénea que utiliza un interposer de silicio para las interconexiones con componentes de fotónica integrada, antena, filtros SAW, MEMS/sensores y circuitos integrados 2,5D y 3D.

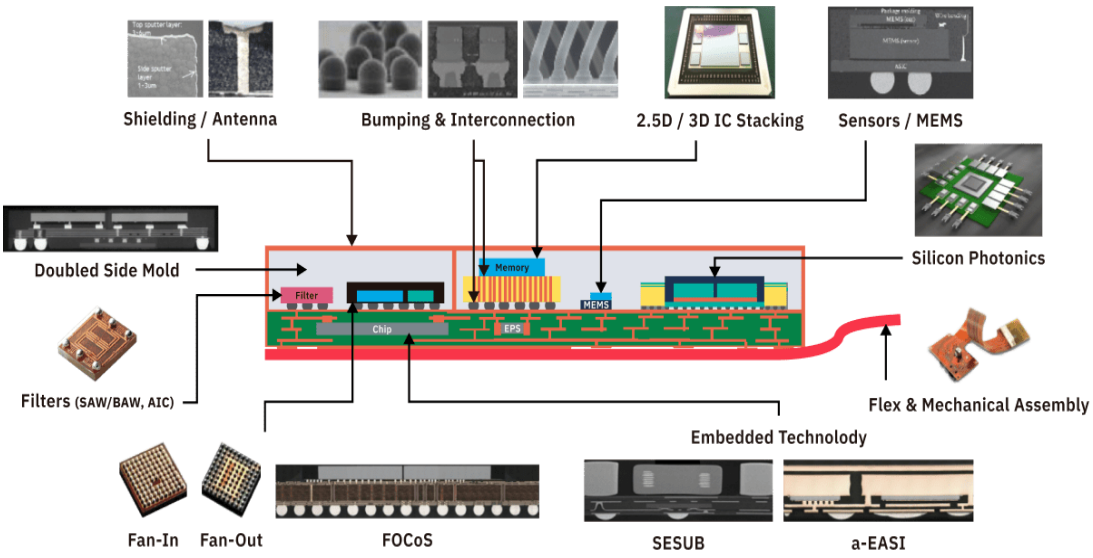


Figura 23. Vista en sección con el interposer que interconecta los diferentes componentes de una Integración heterogénea con diferentes tecnologías. Fuente: Heterogeneous Integration (HI) | ASE

Chips para aplicaciones en Defensa

Los chips para aplicaciones en defensa, incluyendo el espacio para aplicaciones de defensa y seguridad, están diseñados para ofrecer una durabilidad, seguridad y fiabilidad extremas en entornos hostiles. Son mucho más caros, debido a que tienen, generalmente un volumen de producción mucho más bajo que sus equivalentes en el sector civil, además de que su diseño,

proceso de fabricación, prueba, encapsulado y certificación requieren más equipamiento y tiempo que los chips comerciales. Suelen utilizar tecnología más antigua y probada en lugar de nodos de última generación.

En la tabla 4, se muestran las diferencias fundamentales entre los chips para la defensa y el espacio y los chips comerciales.

Característica	Chips de grado militar	Chips Comerciales
Rango de Temperatura	-55ºC a + 125ºC	0ºC a +70ºC
Resistencia a la radiación	Sí (rad-hard)	No
Resistencia a golpes y vibraciones	Alta (MIL-STD.883)	estándard
Características de seguridad	Anti-manipulación, encriptado, cadena de suministro segura	Seguridad Mínima
Nodo del proceso tecnológico	90nm, 130nm, 180nm	5nm, 7nm, 10nm, 28nm
Material	SOI[8], SiC, materiales resistentes a la radiación	Silicio estándard
Coste	10 x a 100x más alto	Bajo costo, producción en masa

Tabla 2. Comparación resumida entre las características clave que diferencian a los chips de grado militar (chips para la defensa, la seguridad y el espacio) y los chips comerciales.

Materiales para la fabricación de chips para la defensa.

Por materiales, aquí se entiende, básicamente los substratos sobre los que se fabrican los elementos activos que, una vez Inter conexionados, forman los sistemas integrados. Estos substratos son los materiales semiconductores. Se distinguen dos grandes grupos: a) los semiconductores simples, como el silicio y el germanio; y b) los semiconductores compuestos, prioritariamente los del grupo III-V.

Afortunadamente, el silicio es un elemento muy abundante en la naturaleza y el elemento clave en la fabricación de chips semiconductores. En efecto, el silicio, el material principal en el que se basan los chips semiconductores, es el segundo elemento más abundante (28%) en la corteza terrestre, por lo que garantiza un suministro relativamente estable de silicio en el futuro cercano.

Silicio (Si):

- Banda prohibida: Aproximadamente 1.1 eV.

- Movilidad de electrones: Moderada (alrededor de 1400 $cm^2/V \cdot s$).

- Temperatura de operación: Adecuado para aplicaciones a temperatura ambiente.

- Usos comunes: Electrónica, fotovoltaica, y circuitos integrados.

- Ventajas: Abundante, bajo costo y bien desarrollado en la industria.

[8] SOI Silicon On Insulator: Es una estructura semiconductora que consta de una capa de silicio monocristalino separada del sustrato principal por una fina capa de aislante.

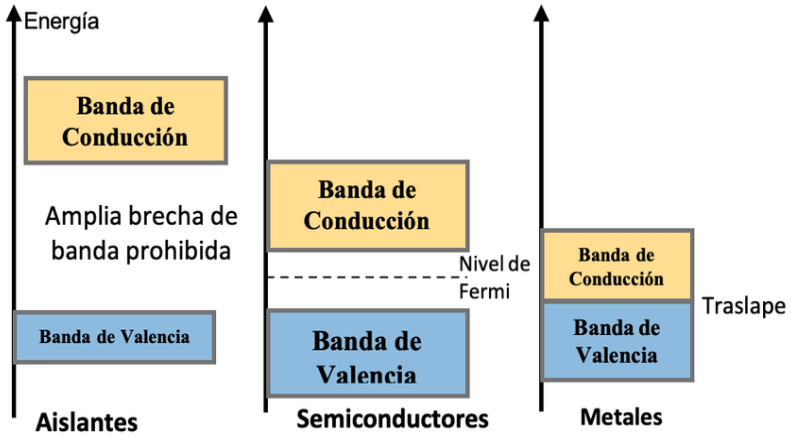


Figura 24. Estructura de bandas de energía de aislantes, semiconductores y metales

Figura 25. Barra de silicio formada a partir de un cristal de silicio (semilla) por el método de Czochvalski. Las obleas se obtienen cortando con sierras especiales la barra en rodajas.

Semiconductores compuestos III-V

Los semiconductores III-V se crean mediante la combinación de elementos del grupo III (como el galio, el aluminio y el indio) y del grupo V (incluido el fósforo, el arsénico y el nitrógeno) de la tabla periódica, lo que da como resultado la generación de una estructura de red cristalina con características electrónicas excepcionales. A diferencia del silicio, los semiconductores III-V ofrecen propiedades de banda prohibida directa, lo que proporciona una conversión altamente eficiente de energía de fotones a electrones. Este atributo único hace que los semiconductores III-V sean particularmente adecuados para aplicaciones optoelectrónicas.

Nitruro de galio (GaN):

- Banda prohibida: Alrededor de 3.4 eV.
- Movilidad de electrones: Alta (aproximadamente 2000 $cm^2/V \cdot s$).
- Temperatura de operación: Excelente para altas temperaturas y alta potencia.
- Usos comunes: Dispositivos de potencia, LED y láseres.
- Ventajas: Alta eficiencia y capacidad para manejar altas tensiones.

El nitruro de galio es un semiconductor compuesto de banda prohibida directa III/V que resulta muy adecuado para transistores de alta potencia capaces de funcionar a altas temperaturas. Desde la década de 1990, se ha utilizado habitualmente en diodos emisores de luz (LED). El nitruro de galio se utiliza en dispositivos de potencia semiconductores, componentes de RF, láseres y fotónica. Se empieza a utilizar GaN en la tecnología de sensores.

En 2006, los transistores GaN, a veces denominados FET de GaN, comenzaron a fabricarse mediante el crecimiento de una fina capa de GaN sobre la capa AIN de una oblea de silicio estándar mediante deposición química en fase de vapor de metal orgánico (MOCVD). La capa AIN actúa como un amortiguador entre el sustrato y el GaN. Este nuevo proceso permitió **que los transistores de nitruro de galio se pudieran producir en las mismas fábricas, ya establecidas, que las de silicio**, utilizando casi los mismos procesos de fabricación. Al utilizar un proceso conocido, esto permite costos de fabricación bajos y similares y reduce la barrera para la adopción de transistores más pequeños con un rendimiento muy mejorado.

La tecnología de GaN es prometedora en aplicaciones como la electrónica de potencia, los sistemas de RF, la industria automotriz, la industria aeroespacial, las telecomunicaciones y las energías renovables.

GaN en defensa

La tecnología **de GaN** ha desempeñado un papel clave en la mejora de la eficacia y la eficiencia de las **operaciones de defensa**, desde los **sistemas de radar avanzados** hasta los **sistemas de comunicación.** Sus características principales son:

- **Alta movilidad de electrones y velocidad de saturación**, lo que permite operaciones de alta velocidad. En comparación con el silicio, GaN tiene propiedades de transporte de electrones superiores, lo que permite velocidades de conmutación más rápidas y posibilita el desarrollo de dispositivos de alta frecuencia.
- **El ancho de banda prohibida de GaN es otra ventaja clave**. Con una banda prohibida aproximadamente tres veces más amplia que el silicio, los semiconductores de GaN exhiben capacidades de alto voltaje de ruptura. Esto permite un manejo eficiente de la energía y los hace adecuados para aplicaciones de alta potencia. Los dispositivos de GaN pueden operar a voltajes más altos sin comprometer el rendimiento, lo que proporciona una mayor confiabilidad y robustez.
- **Los semiconductores de GaN también exhiben una excelente conductividad térmica**. El material disipa eficazmente el calor, lo que permite una gestión térmica eficaz en aplicaciones de alta potencia. Esta propiedad ayuda a prevenir el sobrecalentamiento y mejora la fiabilidad del dispositivo. Además, la compatibilidad del GaN con temperaturas de funcionamiento más altas en comparación con los semiconductores basados en silicio contribuye aún más a mejorar el rendimiento y la fiabilidad, ya que reduce la necesidad de sistemas de refrigeración complejos.
- **Potencial del GaN para dispositivos más pequeños y ligeros**. Debido a su alta densidad de potencia y capacidades de miniaturización, los componentes basados en GaN pueden ofrecer salidas de alta potencia mientras ocupan espacios físicos más pequeños.
- Los semiconductores de GaN también son prometedores en términos de eficiencia energética

Aplicaciones de la tecnología GaN en la industria de defensa.

- La tecnología GaN es valorada por la industria de defensa debido a sus características de rendimiento superiores, que incluyen alta densidad de potencia, eficiencia y confiabilidad. Para empezar, es fundamental para mejorar los sistemas de radar y guerra electrónica (EW), señala "*Military Embedded Systems*"
- Los componentes basados en GaN se utilizan en radares de matriz de barrido electrónico activo (AESA) y otras aplicaciones de alta potencia debido a su capacidad para ofrecer alta potencia de salida, eficiencia y amplio ancho de banda. Estas características hacen que GaN sea ideal para los sistemas de radar y EW de próxima generación, que requieren un tamaño, peso y potencia (SWaP) reducidos.
- GaN también se utiliza en **sistemas de comunicación militar** para mejorar el rendimiento y la confiabilidad. Se emplea en módulos multichip para aplicaciones de alta frecuencia, lo que permite una navegación eficiente y el control del tráfico aéreo en tiempo real. La capacidad de GaN para operar a frecuencias más altas lo hace adecuado para bloqueadores militares y radios terrestres
- La tecnología GaN, según *Efficient Power Conversion Corporation*, es ventajosa para la gestión de energía en aplicaciones de defensa. Ofrece una mayor densidad de potencia y una mejor gestión térmica, lo que es crucial para los sistemas que operan en entornos hostiles. Por ejemplo, los amplificadores de potencia de GaN están reemplazando a los tubos de ondas progresivas (TWT) tradicionales en aplicaciones espaciales debido a su mayor confiabilidad y eficiencia.

En la Tabla 3 se indican las energías de banda prohibida para diferentes materiales semiconductores. Los materiales considerados de anchura de banda prohibida grande (salto de energía en eV > 3) son los de fondo verde. Entre ellos están el SiC y el GaN que, por este motivo, son adecuados para Electrónica de potencia.

MATERIALES SEMICONDUCTORES		
MATERIAL	Símbolo Químico	ENERGÍA BANDA PROHIBIDA (eV)
Germanio	Ge	0,7
Silicio	Si	1,1
Arseniuro de Galio	GaAs	1,4
Carburo de Silicio	SiC	3,3
Óxido de Zinc	ZnO	3,4
Nitruro de Galio	GaN	3,4
Diamante	C	5,5

Tabla 3. Materiales semiconductores y su Energía de banda prohibida en eV (Fogarty, 2019)

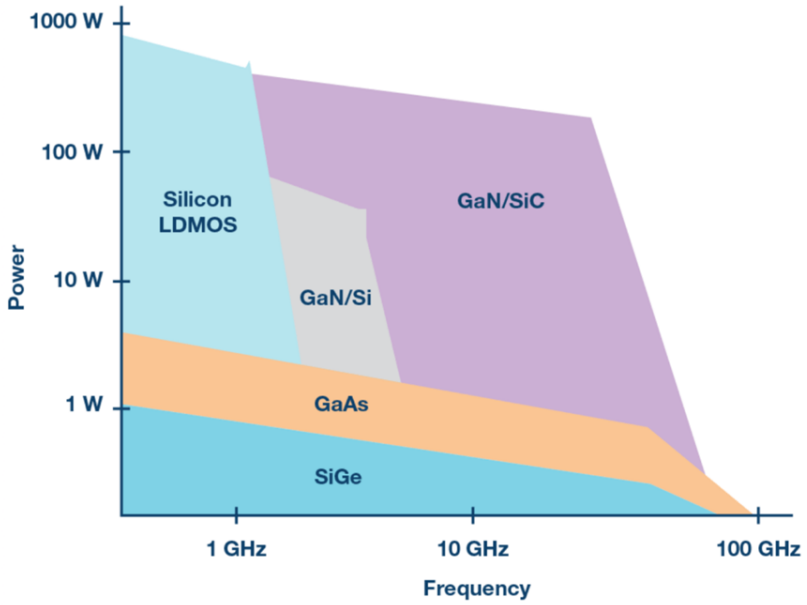


Figura 26. Comparación de potencia y frecuencia de diferentes materiales semiconductores en el rango de microondas, que incluye a las ondas milimétricas. Fuente: Analog Devices (Fogarty, 2019)

En las comunicaciones militares y en el IoT militar en el 5G+ y 6G, se utilizarán cada vez más las ondas milimétricas[9]. En efecto, las ondas milimétricas son de gran interés para la Defensa por las ventajas estratégicas que comportan, entre las que destacan:

- **Alta resolución** → Permite la detección precisa de objetivos.
- **Baja interferencia** → Menos congestión en comparación con frecuencias más bajas.
- **Capacidades de sigilo o secreto** → Difícil de detectar con sistemas de radar convencionales.
- **Penetración de algunos materiales** → Atraviesan la niebla, la lluvia y la ropa.

Aplicaciones de las ondas milimétricas (MMW) en aplicaciones de defensa:

1. Sistemas de Radar

Los radares MMW brindan una resolución superior para la detección y el seguimiento de objetivos.

- Las matrices de escaneo electrónico activo (AESA) utilizan ondas milimétricas para la detección furtiva y la guía de misiles.
- Los radares de control de tiro en aviones de combate operan en frecuencias MMW para apuntar con precisión.

[9] Las **ondas milimétricas** (MMW) se refieren a ondas electromagnéticas con frecuencias entre 30 GHz y 300 GHz y longitudes de onda que van desde 1 mm a 10 mm. Estas ondas se ubican entre las frecuencias de microondas e infrarrojas en el espectro electromagnético.

- Los radares terrestres utilizan MMW para detectar amenazas que vuelan a baja altura, como los drones

Ejemplo: El radar AN/APG-81 en el avión de combate F-35 utiliza MMW para un seguimiento de alta precisión.

2. Armas de energía dirigida (DEW) (*Directed Energy Weapons*)

Las ondas milimétricas de alta potencia se pueden utilizar como armas para desactivar los dispositivos electrónicos enemigos o causar molestias

- Sistemas de denegación activa (ADS) → Arma de control de multitudes no letal que utiliza MMW de 95 GHz para crear una intensa sensación de ardor.
- Tecnología antidrones → La energía MMW puede inutilizar los UAV (vehículos aéreos no tripulados).

Ejemplo: El sistema ADS ("rayo de calor") del ejército de EE. UU. dispersa multitudes sin fuerza letal.

3. Comunicaciones militares seguras

Las frecuencias MMW proporcionan enlaces de comunicación seguros de gran ancho de banda para operaciones militares.

- Las comunicaciones por satélite (SATCOM) utilizan MMW para transmisiones de datos de alta velocidad.
- Las redes de banda ancha tácticas (como las aplicaciones militares 5G) utilizan MMW para compartir datos cifrados de baja latencia.

Ejemplo: Los sistemas de comunicación de banda W (75–110 GHz) se utilizan para la transferencia segura de datos en el campo de batalla.

4. Imágenes y vigilancia

Los sistemas de imágenes MMW pueden detectar armas ocultas y amenazas en condiciones de baja visibilidad.

- Puede ver a través de la ropa y los obstáculos (útil para la lucha contra el terrorismo).
- Funciona en la niebla, el polvo y el humo, donde la óptica tradicional falla.

Ejemplo: Las cámaras pasivas de ondas milimétricas detectan objetos al capturar la radiación natural.

5. Sistemas de guía de misiles

Los buscadores de ondas milimétricas mejoran la precisión de los misiles.

- Los buscadores de misiles utilizan ondas milimétricas para la guía terminal, lo que permite un seguimiento preciso del objetivo.
- La capacidad de penetración en todo tipo de clima hace que las ondas milimétricas sean útiles en todas las condiciones.

Ejemplo: El misil AGM-114R Hellfire utiliza un radar de ondas milimétricas para la adquisición de objetivos.

No obstante, las grandes ventajas estratégicas del uso de las ondas milimétricas (MMW) en aplicaciones de la defensa, su uso plantea una serie de desafíos, entre los que podemos destacar:

- **Absorción atmosférica** → Las señales MMW se debilitan en largas distancias, especialmente con lluvia y niebla.
- **Alto costo del equipo** → La tecnología MMW requiere materiales avanzados y fabricación de precisión.
- **Penetración limitada** → Si bien pueden penetrar la ropa, tienen dificultades con obstáculos densos como paredes y metal.

Entre las tendencias futuras de la utilización de las MMW en Defensa, destaca:

- **Integración con IA** → Los radares MMW impulsados por IA mejoran la detección autónoma de amenazas.
- **Redes de campo de batalla 5G** → Comunicación de alta velocidad y baja latencia para centros de comando militar.
- **Miniaturización** → Los chips MMW compactos mejoran los UAV y los sistemas que usan los soldados.

Las importantes ventajas estratégicas de las MMW en aplicaciones para la defensa dan una relevancia especial a los circuitos diseñados y fabricados con GaN. En efecto, en la Fig. 26, puede apreciarse la ventaja de los circuitos GaN/SiC sobre otras tecnologías como GaAs, SiGe y Silicio en las frecuencias correspondientes a las MMW. GaN es más eficiente energéticamente que el silicio para RF 5G.

De hecho, GaN ha sido el heredero aparente del silicio en los amplificadores de potencia 5G durante años, especialmente cuando se trata de redes 5G MMW. Lo que lo hace tan atractivo es su capacidad para manejar de manera eficiente un voltaje más alto en un área mucho más pequeña que los dispositivos MOSFET de difusión lateral (LDMOS) comparables. Además, puede alimentar una gama mucho más amplia de frecuencias MMW que el silicio estándar. (Fogarty, 2019)

Diseño de chips para la defensa

Sistemas Integrados

Los sistemas electrónicos integrados los podemos definir como un conjunto de componentes electrónicos, normalmente en un único chip o en un SiP, que trabajan juntos de manera coordinada para realizar una función específica como, por ejemplo, controlar un equipo industrial, procesar señales o gestionar datos. En definitiva, es una solución tecnológica eficiente y optimizada que integra hardware y, a menudo, software para cumplir con una función electrónica determinada dentro de un equipo o producto.

Normalmente los sistemas electrónicos integrados necesitan captar datos, que suelen hacer mediante los sensores. Éstos captan señales del mundo físico, por lo que son, generalmente, magnitudes analógicas, que se transforman, por medio del sensor en señales eléctricas analógicas que se pasan a digitales mediante circuitos conocidos como “convertidores analógico-digitales”, que son circuitos que realizan una función esencial y su diseño depende de varios factores como la precisión requerida por el sistema, la velocidad de conversión, el consumo de

energía, etc. Todos estos parámetros determinan el tipo de convertidor A/D utilizado. Una vez convertidos los parámetros físicos captados por los sensores en palabras de n bits, estos son procesados (DSP, microcontrolador, etc. y memorias). Mediante circuitos de comunicaciones se envían (y reciben) los resultados del procesado y se activan (si procede) los activadores (efecto). (ver figura 27)

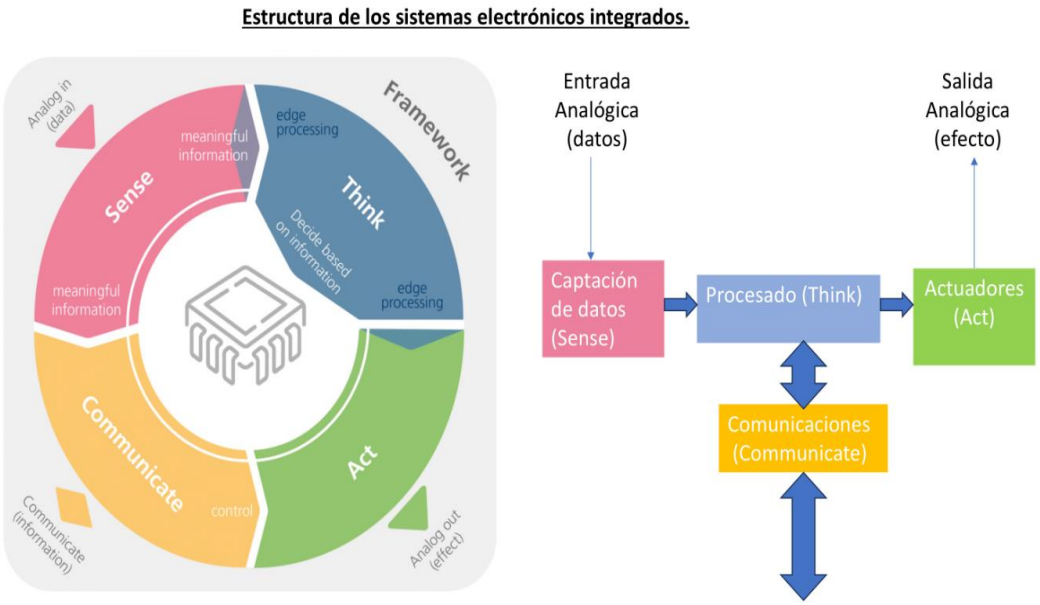


Figura 27. Sistemas integrados: Captación de datos (sensores), procesado(CPUs, GPUs, memorias,…), Actuadores y comunicaciones. Fuente: IPCEI-Microelectronics/CT Chapeau Text y Elaboración propia.

Sensores en la Defensa

Uno de los puntos fundamentales para los sistemas de IA de computación en el borde es el suministro de datos, tanto para entrenamiento como para inferencia. El suministro de datos para inferencia en computación en el borde, por ejemplo, para un vehículo autónomo, o para un sistema de guiado de misiles, debe ser en tiempo real. Esta captación de datos, en tiempo real, lo realizan los sistemas de sensores. En este sentido, los sensores son elementos clave en el funcionamiento de los sistemas de defensa.

Uno de los dispositivos más esenciales en la industria de defensa es el sensor para capturar datos Las innovaciones en vigilancia, comunicación y transporte dependen de estos sensores simples para crear una observación de red intrincada. Las FFAA. utilizan, principalmente, los siguientes:

- **Los sensores activos** pueden localizar objetos en su vecindad mediante el uso de señales o longitudes de onda de luz. Se utilizan en navegación, defensa aérea, vigilancia, búsqueda y rescate y tecnología de sonar. A menudo utilizan una fuente de radiación interna para iluminar su entorno.
- **Los sensores portátiles** se pueden usar directamente en el cuerpo y transmitir datos tanto para entrenamiento como para fines del mundo real. Se pueden usar para rastrear soldados y mejorar las comunicaciones. Son compactos y duraderos para acomodar a un individuo en movimiento.

- **Los sensores de cámara** ayudan a detallar los cambios en el entorno y complementan los sistemas de vigilancia. Pueden ayudar a rastrear objetivos e identificar rostros o movimiento.
- **Los sensores MEMS** (sistemas microelectromecánicos) son un tipo de sensor que detecta los cambios de presión dentro de las aeronaves, acumula datos en satélites e identifica si los vehículos desconocidos son amigos o enemigos. Son compactos, fiables y rentables.
- **Los sensores infrarrojos** pueden ayudar a detectar distintos tipos de armas que pueden pasar desapercibidas a simple vista. Ciertas sustancias químicas y explosivos pueden resultar difíciles de detectar sin la ayuda de la tecnología infrarroja.

Semiconductores y su uso en Defensa

Los semiconductores desempeñan un papel crucial en diversas aplicaciones militares, contribuyendo a los **sistemas de comunicación**, **tecnologías de cifrado**, **sistemas de radar**, **guía de misiles** y **guerra electrónica**. Su pequeño tamaño, bajo consumo de energía y alta confiabilidad los hacen ideales para tecnologías militares que requieren compacidad, eficiencia y durabilidad.

En los **sistemas de comunicación**, los semiconductores se utilizan en el desarrollo de radios avanzadas, comunicaciones por satélite e infraestructura de red. Permiten una transmisión de datos más rápida, una calidad de señal y una conectividad mejoradas.

Los semiconductores también desempeñan un papel vital en **las tecnologías de cifrado**, garantizando una comunicación segura y protegiendo la información militar confidencial del acceso no autorizado.

En los **sistemas de radar**, los semiconductores se utilizan en la creación de amplificadores de alta frecuencia y componentes de procesamiento de señales. Estos componentes permiten la detección y el seguimiento precisos de los objetivos, mejorando el conocimiento de la situación y ayudando en la identificación de amenazas.

Los semiconductores también se utilizan en la **guerra electrónica**, donde permiten la creación de contramedidas electrónicas para interrumpir las comunicaciones enemigas y los sistemas de radar. Los **sistemas de guía de misiles** basados en semiconductores utilizan sensores y procesadores para calcular las trayectorias de los objetivos y guiar los misiles hacia sus objetivos previstos con precisión y exactitud.

Condiciones que deben cumplir los chips de Defensa

En el diseño de chips para defensa hay que tener en cuenta una serie de condiciones especiales debido a su utilización en ambientes hostiles, con dificultades de recambio y mantenimiento, estableciéndose un conjunto de especificaciones que estos chips deberán cumplir para ser certificados en grado militar.

Rango de temperatura y tolerancia ambiental

- ***Temperatura de operación***: -55 ºC a +125 ºC (algunas veces puede llegar hasta + 200 ºC)
- ***Resistencia a la radiación:*** Alta (rad-hard)
- ***Resistencia a golpes y vibraciones:*** Resiste condiciones extremas (Mil-STD 883)
- ***Resistencia a la humedad y a la corrosión:*** Protegido contra la humedad y la niebla salina.

Confiabilidad y Longevidad

- ***Tasa de fallos:*** Muy baja (diseñada para sistemas de misión crítica)
- ***Vida útil:*** 15-30 años
- ***Estándares de prueba:*** MIL-STD-883 (pruebas de fiabilidad de estándar militar)

Seguridad y Anti manipulación

- ***Características Anti manipulación:*** Arranque seguro, mecanismos de autodestrucción y de cifrado en hardware.
- ***Protección electromagnética (EMI/EMC):*** Protegido para evitar ataques electromagnéticos.
- ***Seguridad de la cadena de suministro:*** Seguimiento estricto, a menudo integrado en fábricas seguras.

Endurecimiento por Radiación (rad-hard)

- ***Resistencia a la Radiación:*** Diseñado para soportar rayos gamma, rayos cósmicos y bombardeo de neutrones.
- ***Dosis ionizante total (TID) Tolerancia:*** 100 Krad- 1Mrad

Proceso de fabricación y materiales

- ***Nodo tecnológico del proceso:*** Más grandes que los chips comerciales para robustez (90 nm, 120 nm, 180 nm).
- ***Material del sustrato:*** Silicio, SOI, SiC, GaN on SiC
- ***Encapsulado:*** Cerámico, sellado con metal.

Los chips militares evitan los nodos ultrapequeños (5 nm o menos) debido a la sensibilidad a la radiación, así como a la mayor robustez de los nodos más grandes.

Costo y disponibilidad

- ***Costo:*** Cuestan entre 10 y 100 veces más que los chips comerciales equivalentes.
- ***Volumen de producción:*** Volumen bajo (Pedidos personalizados, fábricas clasificadas).
- ***Disponibilidad:*** Restringida, controlada por los gobiernos.

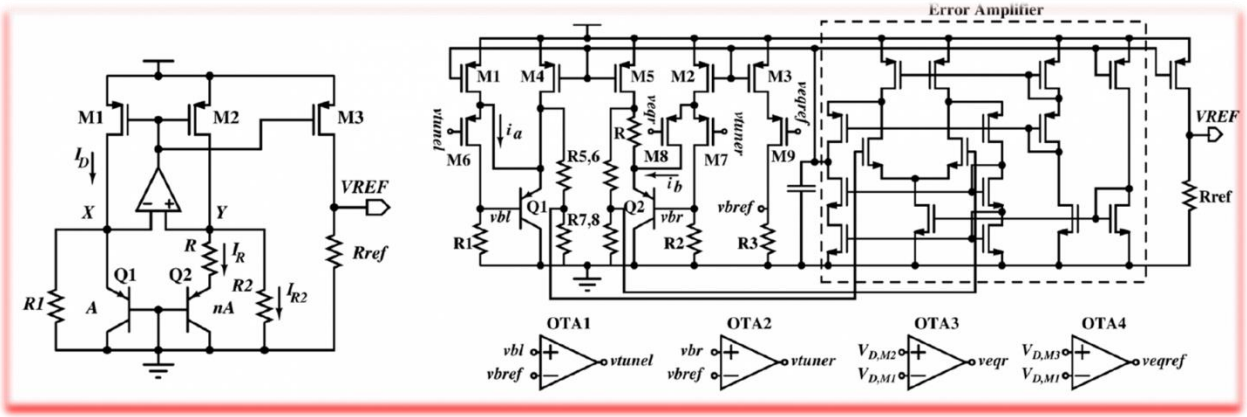


Figura 28. Generador de tensión de referencia: a) el de la izquierda es para uso comercial; b) el de la derecha es un generador de tensión de referencia diseñado para cumplir con las especificaciones militares.

Como puede apreciarse en la figura 28 el diseño de un generador de tensión de referencia es mucho más complejo en el caso militar que en el caso comercial (circuito de la izquierda) para poder cumplir con los requerimientos de mayor estabilidad con una variación mayor de la temperatura (Rango militar) y la certificación MIL-STD.

Hardware Abierto (RISC V)

RISC-V es una arquitectura de conjunto de instrucciones (ISA) abierta y modular que ha ganado interés en aplicaciones de defensa por su flexibilidad y ausencia de restricciones de propiedad intelectual. Sin embargo, su uso en entornos militares tiene ventajas y desafíos.

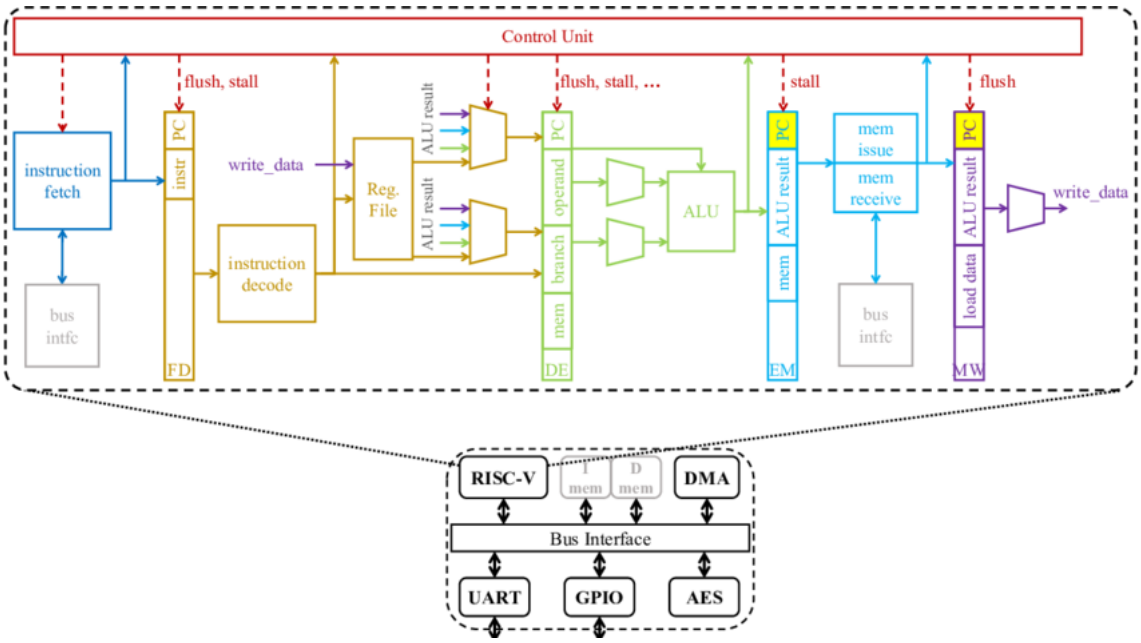


Figura 29. Diagrama simplificado arquitectura RISC-V. Fuente: Kiaei,P & Schaumont, P., 2022, IACR transactions on Cryptographic hardware and Embedded Systems.

Ventajas de RISC-V para Defensa:

1. **Soberanía Tecnológica y Reducción de Dependencia**

- Al ser una arquitectura abierta, los gobiernos y empresas pueden desarrollar procesadores personalizados sin depender de empresas como Intel, AMD o ARM.

China y Rusia están invirtiendo en RISC-V para evitar sanciones y restricciones de exportación de chips avanzados.

2. Seguridad y Personalización

- Los procesadores RISC-V pueden diseñarse con medidas de seguridad específicas, como criptografía avanzada y protección contra ataques de hardware.

En efecto, en defensa, se pueden crear chips resistentes a ciberataques o adaptados a sistemas clasificados sin depender de proveedores extranjeros.

3. Eficiencia Energética y Bajo Consumo

- Su diseño permite optimizaciones en consumo energético, lo cual es clave para drones, satélites militares y dispositivos portátiles para soldados.

4. Escalabilidad y Modularidad

- RISC-V permite personalizar la arquitectura según las necesidades de defensa, desde procesadores de bajo consumo hasta chips de alto rendimiento para supercomputación militar.

5. Costos Reducidos

• Al no pagar licencias a ARM o x86, los costos de desarrollo pueden ser menores en el largo plazo, favoreciendo la adopción en equipos de defensa con grandes volúmenes de producción.

Inconvenientes de RISC-V para Defensa:

1. Ecosistema Inmaduro

• En comparación con x86 y ARM, RISC-V aún tiene menos soporte en software y herramientas de desarrollo.

Así muchos sistemas de defensa usan software optimizado para x86 o ARM, lo que dificultaría una transición rápida.

2. Falta de Soporte para Aplicaciones Críticas

• Actualmente, las versiones comerciales de RISC-V no tienen el mismo nivel de rendimiento que las CPUs de Intel o AMD utilizadas en supercomputadoras militares o sistemas avanzados de radar.

3. Vulnerabilidad a Control Chino

• China ha invertido fuertemente en RISC-V, y empresas chinas son líderes en su desarrollo (como Alibaba y SiFive), lo que genera preocupaciones en países occidentales sobre posibles puertas traseras o espionaje.

Actualmente EE.UU. está considerando restricciones al acceso de China a tecnología RISC-V, lo que podría afectar su adopción global.

4. Falta de Procesos de Certificación Militar

• Los chips RISC-V aún no tienen la misma cantidad de certificaciones de seguridad y fiabilidad requeridas por el sector militar en comparación con ARM y x86.

Procesos Tecnológicos para Defensa

Materiales y Tecnología de Fabricación

Militar:

- ✔ Uso de materiales más resistentes, como carburo de silicio (SiC) y nitruro de galio (GaN), en lugar de silicio convencional, para soportar altas temperaturas y radiación.
- ✔ Procesos de fabricación más antiguos y robustos (180 nm, 90 nm, 65 nm) en lugar de nodos ultrafinos como 5 nm o 3 nm.
- ✔ Mayor uso de back-end packaging especializado para encapsular los chips y protegerlos de interferencias electromagnéticas (EMI) y ataques de hardware.

Comercial:

- ✔ Uso de silicio puro y tecnologías de vanguardia para maximizar el rendimiento y reducir el consumo energético.
- ✔ Enfoque en miniaturización con nodos avanzados (5 nm, 3 nm, 2 nm).
- ✔ Procesos de fabricación optimizados para costos y producción masiva, sin priorizar durabilidad extrema

Aumento de la Fiabilidad: Eliminar la Electromigración

La electromigración es el transporte de masa de átomos de metal causado por el flujo de electrones de la corriente que fluye a través de un conductor, generalmente cobre. Cuando la densidad de corriente es lo suficientemente alta, el metal se difundirá en la dirección del flujo de corriente, creando pequeños montículos aguas abajo y dejando vacantes o huecos. Con suficiente electromigración, se producen fallas debido al adelgazamiento severo de la línea, lo

que causa aberturas, o debido a montículos que unen líneas adyacentes, lo que causa cortocircuitos. (Ver figura 30).

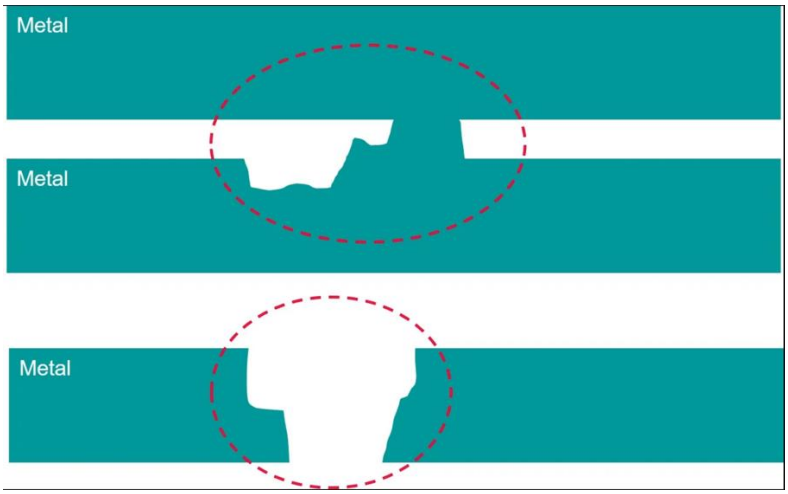


Figura 30. Cortocircuitos y circuitos abiertos causados en las pistas de metal de un chip por la electromigración. Fuente: Peters, Laura 2024

La ***electromigración*** en el propio chip es un problema que ha requerido atención y soluciones por parte de los ingenieros de fiabilidad a lo largo del tiempo con la reducción de dimensiones de los dispositivos y las interconexiones. Sin embargo, con el escalado y los rápidos avances en encapsulados avanzados – implementación de TSV, encapsulado en abanico con capas de redistribución y protuberancias de pilares de cobre- la *electromigración* se ha convertido en un problema importante a nivel de encapsulado.

La corriente que fluye a través de la soldadura provoca calentamiento Joule, y el calor de otras partes del encapsulado también puede disiparse a través de las soldaduras. La *electromigración* (EM) puede convertirse en un problema para las conexiones de las juntas de soldadura entre el chip y el intercalador, o entre el chip y la PCB, así como en las RDL. Las fallas en las juntas de soldadura suelen manifestarse como huecos o grietas.

La *electromigración* progresa más rápidamente a temperaturas más altas, a corrientes más altas, bajo mayor tensión mecánica y en presencia de defectos o impurezas en el metal. La ecuación de Black[10] describe el tiempo medio hasta el fallo de una interconexión con respecto a su temperatura, densidad de corriente y la energía de activación necesaria para desalojar un átomo metálico.

[10] La ecuación de Black se expresa como: $MTTF = \frac{A}{J^N} . e^{\frac{Ea}{K.T}}$.

J es la densidad de corriente, k es la constante de Boltzmann, T es la temperatura, Ea es la energía de activación y N es un factor de escala que depende de las propiedades del metal. La ecuación de Black es útil porque muestra fácilmente cómo las interconexiones más cortas y anchas tienden a tener un MTTF (*Tiempo Medio Hasta el Fallo*) más largo. Además, el tiempo hasta el fallo por electromigración depende en gran medida de **la temperatura de la interconexión**. Esta temperatura se debe principalmente a la temperatura ambiental del chip, el autocalentamiento del conductor causado por el flujo de corriente, el calor de las interconexiones o transistores vecinos y la conductividad térmica del material circundante.

Tipos de chips para su uso en IA

Por **"Chips de IA"** nos referimos a ciertos tipos de chips procesadores de datos que alcanzan alta eficiencia y velocidad para cálculos específicos de IA a expensas de baja eficiencia y velocidad para otros cálculos.

El desarrollo de **chips IA** (también conocidos como "*aceleradores de IA*"), ligado al avance de las tecnologías de semiconductores que han permitido el procesado eficiente de cantidades enormes de datos, ha jugado, juega y jugará un papel muy importante en el avance de la IA.

Los chips para IA han sido optimizados para satisfacer las necesidades de computación de los algoritmos de IA, habilitando un procesado más rápido y eficiente desde el punto de vista de consumo de energía que los procesadores tradicionales de propósito general (CPUs).

Los chips para IA se orientan a acelerar varias tareas de IA, como procesado de los datos, reconocimiento de patrones, entrenamiento de redes neuronales, e inferencia. Son muy adecuados para realizar tareas que involucran operaciones matriciales a gran escala y **computación paralela** propias de los **algoritmos de aprendizaje profundo.** (Li D. , 2023)

.

El paralelismo es un concepto fundamental en el diseño de los chips de Inteligencia Artificial, jugando un papel crucial en acelerar la ejecución de las tareas de IA. Estos conceptos han sido optimizados para satisfacer las grandes necesidades computacionales de los algoritmos de IA, especialmente los que involucran a redes neuronales y modelos de aprendizaje profundo. Los chips de IA responsables de ejecutar las tareas de IA utilizan una arquitectura SIMD[11], mientras realizan las mismas operaciones sobre múltiples puntos de datos, acelerando de forma significativa la velocidad de cómputo. Para acelerar la velocidad de procesamiento general, se considerará, en el diseño de chips IA, el paralelismo, lo que significa procesar múltiples puntos de datos o cálculos *simultáneamente*.

Según su construcción los chips para IA pueden clasificarse en: a) GPUs (Graphic Processing Units); b)FPGAs (Field Programmable Gate Arrays) c) ASICs (Circuitos Integrados de Aplicación Específica) que incluyen Redes Neuronales Profundas, TPUs (Tensor Processing Units) y d) Circuitos neuromórficos, AIMC (Artificial Intelligence Memory in Computing).

Así pues, los chips IA, no solo incluyen GPUs, que, originalmente, fueron concebidos para procesado gráfico (para videojuegos), y FPGAs, sino chips especializados (ASICs) que ya son el mayor componente de la computación en IA debido a sus altas capacidades de procesamiento paralelo. Estos chips se personalizan para cargas de trabajos IA específicas o aplicaciones que proporcionan un rendimiento excelente y una gran eficiencia energética optimizando su diseño para un conjunto específico de tareas (Pang, March-April 2022)

[11] SIMD (Single Instruction Multiple Data) : Las instrucciones SIMD son un tipo especial de instrucciones de procesador que pueden funcionar con varios elementos de datos al mismo tiempo. Por ejemplo, en lugar de agregar dos números a la vez, una instrucción SIMD puede agregar cuatro u ocho números a la vez. Esto puede ahorrar tiempo y recursos, ya que el procesador puede ejecutar menos instrucciones y utilizar menos registros. Las instrucciones SIMD también se conocen como instrucciones vectoriales, porque pueden manipular vectores de datos, que son colecciones ordenadas de números.

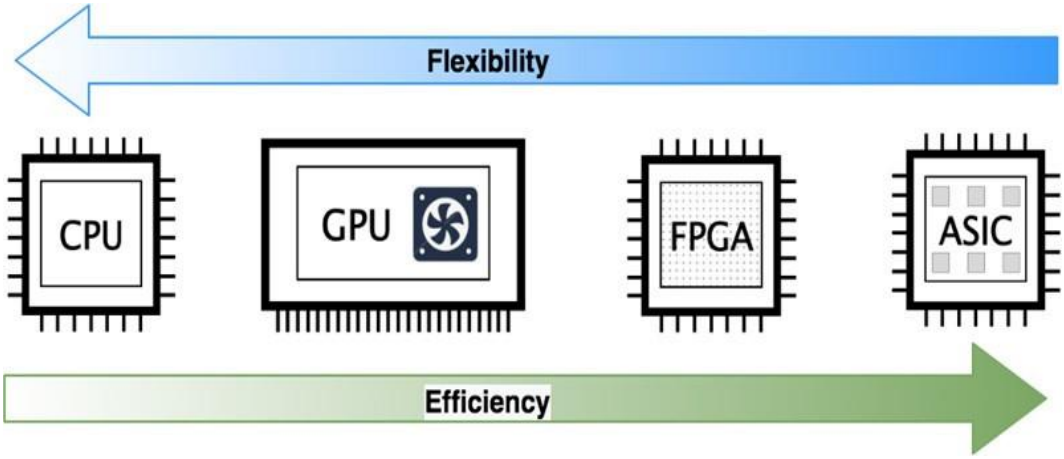


Figura 31. Clasificación general de los chips IA según su construcción: a) CPU; b)GPUs; c) FPGAs y d) ASICs. De izquierda a derecha aumentan su eficiencia energética; de derecha a izquierda aumentan su flexibilidad (la arquitectura puede utilizarse para diversas aplicaciones de IA). Fuente: (Li D. , 2023)

Clasificación de los chips IA según su propósito de uso:

- **Chips de entrenamiento** (Training chips): tienen gran potencia de cálculo y gran ancho de banda de memoria, para manejar conjuntos de datos enormes y arquitecturas de modelos complejas.
- **Chips de Inferencia** : se diseñan para desplegar modelos de IA entrenados en aplicaciones del mundo real

	Training		Inference		Generality[88]	Inference accuracy[89]
	Efficiency	Speed	Efficiency	Speed		
CPU	1x baseline				Very High	~98-99.7%
GPU	~10-100x	~10-1,000x	~1-10x	~1-100x	High	~98-99.7%
FPGA	-	-	~10-100x	~10-100x	Medium	~95-99%
ASIC	~100-1,000x	~10-1,000x	~100-1,000x	~10-1,000x	Low	~90-98%

Tabla 4. Comparación de prestaciones entre los grandes tipos de chips para IA. Fuente: (Khan & Mann, April 2020)

Los ASICs tienen una mejor eficiencia que las FPGAs y la GPUs. Tanto en entrenamiento como en inferencia, pero tienen mucha menos flexibilidad que las GPUs y las FPGAs. Ver Fig. 31.

Todas las aplicaciones de IA deben ser capaces de entrenarse e inferir. Durante la fase de entrenamiento, los desarrolladores presentan imágenes a la red neuronal (por ejemplo, las de perros o peatones, para, por ejemplo, los evite un coche autónomo) y realizan pruebas de reconocimiento. Luego, refinan los parámetros de la red hasta que la red neuronal muestra alta precisión en la detección visual. Una vez que la red ha visto millones de imágenes y está completamente entrenada, permite el reconocimiento de perros y peatones **durante la fase de inferencia**.

La nube es un lugar ideal para el entrenamiento porque proporciona acceso a vastos almacenes de datos de múltiples servidores, y cuanta más información revise una aplicación de IA durante el entrenamiento, mejor será su algoritmo. Además, la nube puede reducir los gastos porque permite que las unidades de procesamiento gráfico (GPU) y otro hardware costoso entrenen múltiples modelos de IA. Dado que el entrenamiento se produce de forma intermitente en cada modelo, la capacidad no es un problema.

Con la inferencia, los algoritmos de IA manejan menos datos, pero deben generar respuestas más rápidamente. Por ejemplo, un automóvil autónomo no tiene tiempo de enviar imágenes a la nube para su procesamiento una vez que detecta un objeto en la carretera, ni las aplicaciones médicas que evalúan a pacientes críticos tienen margen de maniobra para interpretar los escáneres cerebrales después de una hemorragia, ni las aplicaciones militares que detectan un misil que se dirige a su objetivo y debe actuarse con el escudo anti-misiles antes de que impacte. Y eso hace que la informática de borde, o dentro del dispositivo, sea la mejor opción para la inferencia. (Batra, Jacobson, Madhav, Queirolo, & Santhanam, December 2018)

Comparación entre Arquitecturas Von Neuman (CPUs) y Paralelas (GPUs,ASICS)

El algoritmo, la potencia computacional y los datos masivos (big data) son los tres elementos principales que impulsan el crecimiento de la IA. El desarrollo de Internet, del IoT y de grandes bases de datos ha propiciado la disponibilidad de datos masivos, igualmente, el desarrollo de la algorítmica ha sido importante en las últimas décadas, sin embargo, sin el avance en la potencia computacional no se habría logrado el nivel actual de la IA.

El avance en la potencia computacional está ligado al desarrollo continuo en la industria de semiconductores (ley de Moore y nuevos dispositivos como FinFET y GAA) y al diseño de nuevas arquitecturas (paralelismo) que soportan las necesidades del aprendizaje profundo.

Los primeros sistemas que pueden considerarse dentro de lo que se denomina IA, llamados sistemas expertos, se implementaban sobre chips que se basaban en una CPU (Central Processor Unit) que es la arquitectura Von Neuman por antonomasia.

Una CPU Von Neuman se basa en una memoria donde se comparten los datos y las instrucciones, con una unidad de control (donde está el contador de programa y el decodificador de instrucciones), registros y una ALU (Unidad Aritmético-Lógica). La GPU (Graphics Processing Unit) realmente, es un procesador de imágenes y muestra la información que se va a visualizar, proporciona señales de escaneo a la pantalla y la controla. Es un componente esencial que conecta la pantalla y la placa base del ordenador. Comparada con una CPU tradicional, la GPU tienen una arquitectura con mucho más paralelismo y es más eficiente en el procesado de datos gráficos y algoritmos complejos. Si se compara la diferencia en estructura entre GPU y CPU, la mayor parte del área de una CPU es el controlador y los registros, mientras que en la GPU tiene mucha más área dedicada a ALUs para el procesado paralelo a través de instrucciones SIMD (Single Instruction Multiple Data) para ejecutar el procesado sobre múltiples datos de forma simultánea. (Li, Gu, & Jiang, 2019). Véase la figura 32.

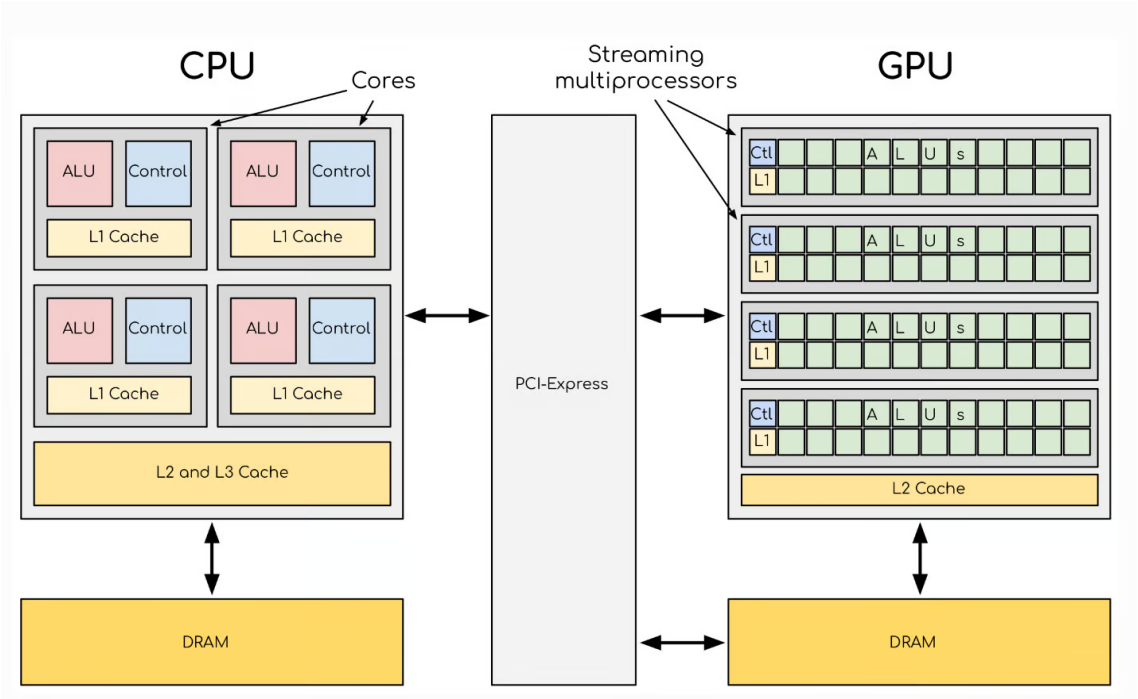


Figura 32. Comparación entre una arquitectura CPU y una arquitectura GPU (paralelismo). Fuente: The GPU hardware and software ecosystem — GPU programming: why, when and how? documentation

Comparadas con CPUs de núcleo único, la velocidad de procesado de los programas sobre un sistema GPU, frecuentemente, es de decenas o incluso millares de veces mayor. Las GPUs tienen una gran versatilidad, alta velocidad y eficiencia, y están especialmente indicadas para aprendizaje profundo (Deep Learning). Deep Learning GPU: Making the Most of GPUs for Your Project

Como se ha comentado, la intención original del diseño de las GPUs era hacer frente a la necesidad de la computación paralela a gran escala en el procesamiento de imágenes. Por todo esto, cuando se aplica a algoritmos de aprendizaje profundo tiene algunas limitaciones: a) Las ventajas del procesado paralelo no se pueden aprovechar completamente. El aprendizaje profundo tiene dos funciones: **entrenamiento** e **inferencia.** Las GPUs son muy eficientes en el proceso de entrenamiento, sin embargo, en el proceso de inferencia su eficiencia es menor que otras alternativas. b) La estructura hardware no puede ser configurada de forma flexible. La estructura hardware de las GPUs está relativamente fijada y no puede ser configurada de forma flexible como en una FPGA. c) Cuando se ejecuta un algoritmo sobre la GPU es menos eficiente energéticamente que cuando se corre sobre una FPGA. En la Tabla 5, puede verse una comparativa de características entre las CPUs y las GPUs.

Característica	CPU	GPU
Arquitectura	Secuencial, un solo núcleo o pocos núcleos	Paralelismo masivo, múltiples núcleos
Velocidad de procesamiento	Ideal para tareas secuenciales	Ideal para tareas paralelizadas

Versatilidad	Más versátil, maneja una variedad de tareas	Especializada en gráficos y cálculos paralelos
Memoria	Memoria compartida entre datos e instrucciones	Memoria especializada (VRAM)
Tareas	Tareas generales, control de software, lógica	Procesamiento gráfico, simulaciones, IA
Uso común	Navegadores, software, sistemas operativos	Juegos, edición de video, IA, aprendizaje automático

Tabla 5. CPU vs GPUs

GPUs: Tipos, Prestaciones. NVIDIA H100 Tensor Core.

Hoy en día las GPUs son el tipo de Chip IA predominante en los centros de datos IA en la nube, aunque las combinaciones de CPU+Aceleradores de AI (chips con arquitecturas especializadas (ASICs)) van imponiéndose.

Las GPUs, que están más extendidas hoy en día son las de NVIDIA que, inicialmente, como se ha comentado, se diseñaron para procesadores gráficos. En efecto, en la mayoría de los centros de datos en la nube se encuentran el H100 y el H200. A principio de este año 2025 está previsto que se empiece a instalar el B200, que casi cuadruplica el rendimiento del H100. (ver figura 33)

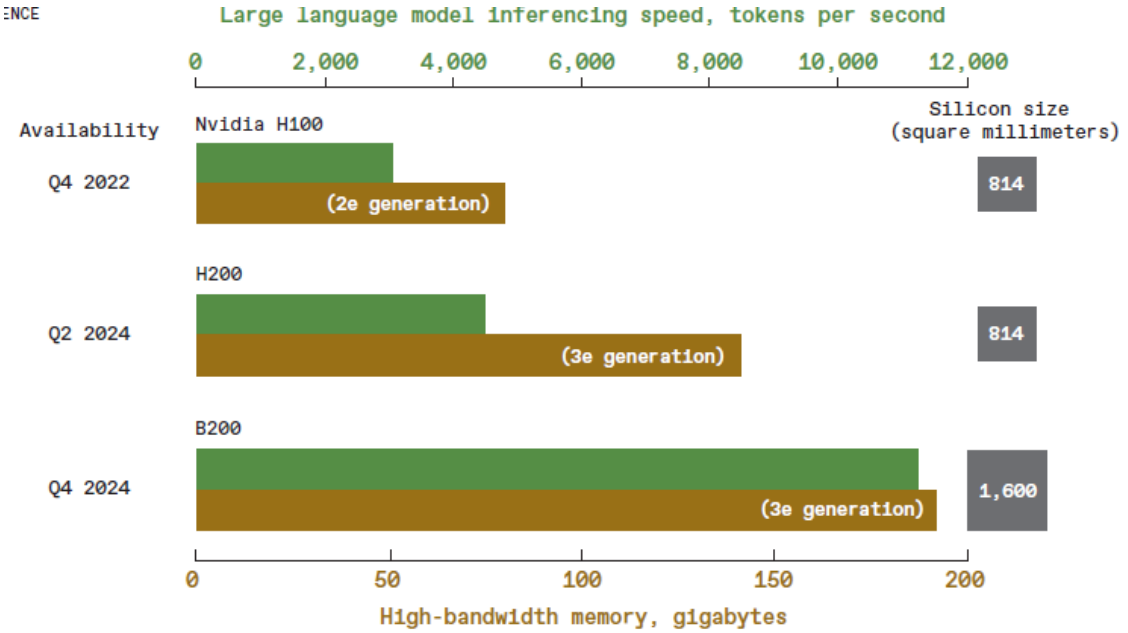


Figura 33. Comparación del rendimiento de las GPUs de Nvidia, H100, H200 y B200. Fuente: (Smith, October 2024)

De acuerdo con (Smith, October 2024)*:* "El dominio de la IA de Nvidia, como la explosión del aprendizaje automático en sí, es un giro reciente de los acontecimientos. Pero tiene sus raíces en el esfuerzo de décadas de la compañía para establecer las GPU como hardware de computación general que es útil para muchas tareas además de la representación de gráficos.

Ese esfuerzo abarca no solo la arquitectura de GPU de la empresa, que evolucionó para incluir **"núcleos tensores"** expertos en acelerar las cargas de trabajo de IA, sino también, fundamentalmente, su plataforma de software, llamada **CUDA** (Compute Unified Device Architecture), para ayudar a los desarrolladores a aprovechar el hardware". Es esencial la existencia de herramientas software eficientes que permitan a los desarrolladores aprovechar al máximo los recursos hardware que se les ofrecen.

Lanzada en 2006, CUDA ayuda a los desarrolladores a utilizar los numerosos núcleos de una GPU de Nvidia. Eso ha demostrado ser esencial para acelerar las tareas de cómputo altamente paralelizadas, incluida la IA generativa moderna. El éxito de Nvidia en la creación del ecosistema CUDA hace que su hardware sea el camino de menor resistencia para el desarrollo de IA.

Mientras que empresas con décadas de antigüedad como Advanced Micro Devices (AMD) e Intel buscan utilizar sus propias GPU para competir con Nvidia, empresas emergentes como Cerebras y sambaNova han desarrollado arquitecturas de chips radicales que mejoran drásticamente la eficiencia del entrenamiento y la inferencia de la IA generativa. Estos son los competidores con más probabilidades de desafiar a Nvidia (Smith, October 2024).

NVIDIA H100

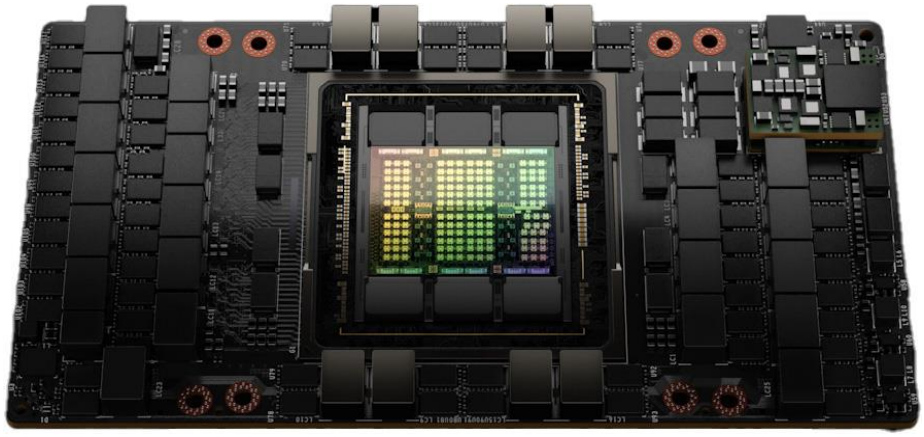


Figura 34. GPU Nvidia H100. Fuente: Nvidia

La **GPU NVIDIA® H100 Tensor Core**, basada en la arquitectura de **GPU NVIDIA Hopper,** ofrece el siguiente gran salto en el rendimiento informático acelerado para las plataformas de centros de datos de NVIDIA. H100 acelera de forma segura diversas cargas de trabajo, desde cargas de trabajo de pequeñas empresas, hasta la exaescala y modelos de IA de billones de parámetros.

Implementado mediante **el proceso 4N de TSMC** personalizado para NVIDIA con 80 mil millones de transistores y que incluye numerosos avances arquitectónicos, H100 es el chip más avanzado del mundo jamás construido hasta el año 2024.

NVIDIA H200

Basada en la arquitectura de NVIDIA Hopper™, la NVIDIA H200 es la primera GPU que ofrece 141 gigabytes (GB) de memoria HBM3e a 4,8 terabytes por segundo (TB/s), lo que supone casi el doble de capacidad que la GPU NVIDIA H100 Tensor Core, con 1,4 veces más ancho de banda de memoria. La H200 cuenta con una memoria más grande y rápida que acelera la IA generativa y

los LLM, a la vez que hace avanzar la computación científica para cargas de trabajo de HPC con una mejor eficiencia energética y un menor coste total de propiedad.

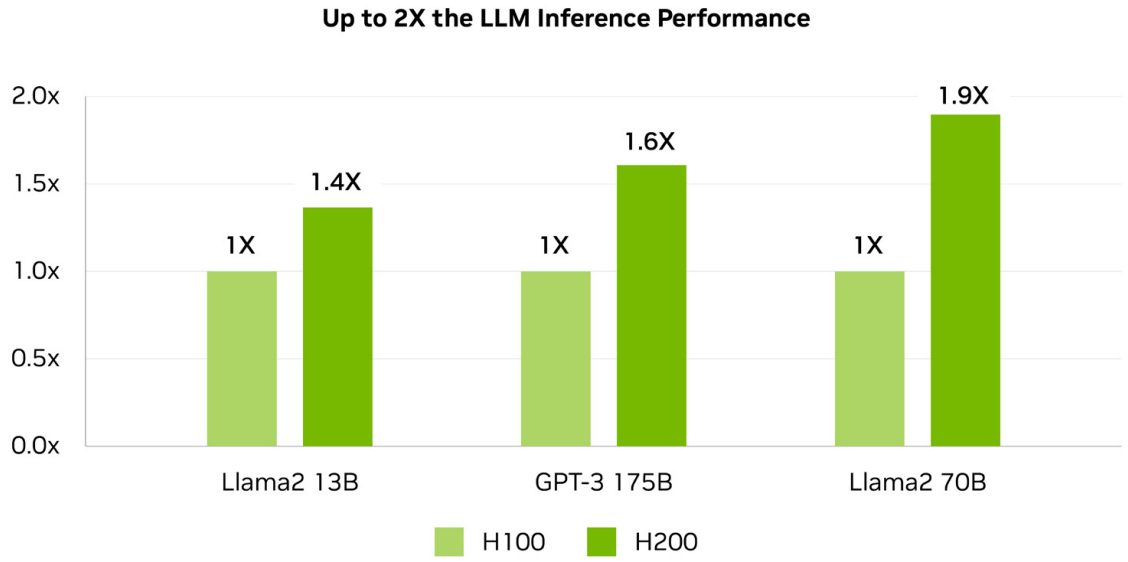


Figura 35. Comparación entre H100 y H200 en prestaciones de inferencia para LLM. Fuente: Nvidia

GPU Blackwell

Construida con 208 mil millones de transistores, más de 2,5 veces la cantidad de transistores en las GPU NVIDIA Hopper, y utilizando el proceso 4NP de TSMC diseñado para NVIDIA, Blackwell es la GPU más grande jamás construida. NVIDIA Blackwell logra el mayor cómputo jamás creado en un solo chip, 20 petaFLOPS.

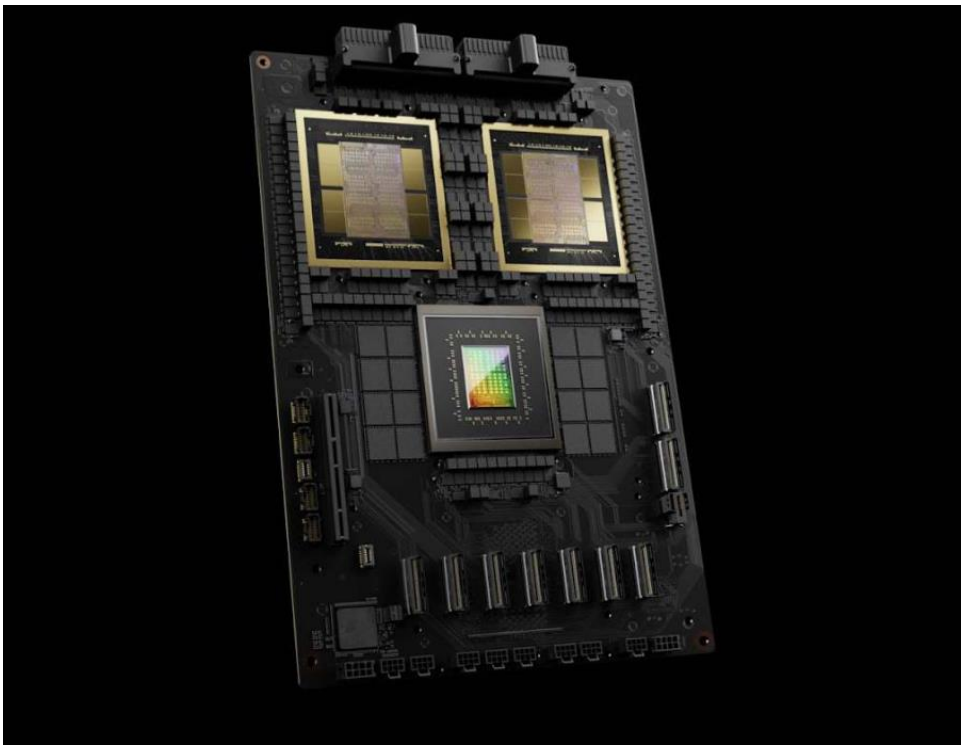


Figura 36. Superchip NVIDIA GB200, incluye dos GPU Blackwell y una CPU Grace. Fuente: Nvidia

Esta arquitectura puede incorporar una cantidad significativa de potencia de cómputo al fusionar dos matrices en una única GPU unificada. Cada una de las dos matrices es la matriz más grande posible dentro de los límites del tamaño de la retícula, tan grande como se puede construir actualmente. Los dos chips están conectados y unificados con una única interfaz de alto ancho de banda (NV-HBI) de chip a chip de NVIDIA de 10 terabytes por segundo (TB/s), lo que proporciona una GPU unificada y totalmente coherente. La arquitectura Blackwell es mucho más que un chip con altas tasas de cálculo de operaciones de punto flotante por segundo (FLOPS). Sigue aprovechando y beneficiándose del rico ecosistema de herramientas de desarrollo de NVIDIA, bibliotecas CUDA-X™, más de cuatro millones de desarrolladores y más de 3000 aplicaciones que escalan el rendimiento en miles de nodos.

Transformer Engine de segunda generación

Blackwell presenta el nuevo Transformer Engine de segunda generación. El Transformer Engine de segunda generación utiliza la tecnología Blackwell Tensor Core personalizada combinada con las innovaciones de TensorRT-LLM y Nemo Framework para acelerar la inferencia y el entrenamiento para los modelos LLM y Mixture-of-Experts (MoE) NVIDIA Blackwell Architecture Technical Overview

AMD

AMD ha luchado contra Nvidia en el campo de los chips gráficos durante casi dos décadas. Ha sido, a veces, una lucha desequilibrada. Cuando se trata de gráficos, las GPU de AMD rara vez han superado a las de Nvidia en ventas o reconocimiento de marca. Aun así, el hardware de AMD tiene sus puntos fuertes. La amplia cartera de GPU de la empresa se extiende desde gráficos integrados para computadoras portátiles hasta GPU para centros de datos enfocadas en IA con más de 150 mil millones de transistores. La empresa también fue una de las primeras en apoyar y adoptar la memoria de gran ancho de banda (HBM), un tipo de memoria que ahora es esencial para las GPU más avanzadas del mundo. “Si nos fijamos en el hardware… se compara favorablemente” con Nvidia, dice Kimball, refiriéndose al Instinct MI325X de AMD, un competidor del H100 de Nvidia. “AMD hizo un trabajo fantástico al diseñar ese chip”. (Smith, October 2024)

La GPU discreta AMD Instinct MI325X (Ver Fig. 37) ofrece un rendimiento superior en un amplio conjunto de tipos de datos necesarios para el software de IA, incluidos FP16, BF16, FP8 e INT8 utilizados tanto en inferencia como en entrenamiento de alta precisión. Una memoria HBM3E líder en la industria de **256 GB** MI325-001A y **un ancho de banda de 6 TB/s** permiten que un solo acelerador contenga y procese un **modelo de 1 billón de parámetros**, al tiempo que reduce el costo total de propiedad para modelos de selectos “Large-Language”. La compatibilidad con “Matrices dispersas”[12] economiza aún más el uso de memoria y aumenta la velocidad computacional, lo que ayuda a permitir un escalamiento sostenible de soluciones de IA en los centros de datos, acelerando el tiempo de comercialización y mejorando el rendimiento.

[12] “Matrices dispersas” o “esparsidad de la matriz”, en inglés, “matrix sparsity” se refiere a la proporción de elementos cero o muy cercanos a cero en una matriz. Se usa en redes neuronales, procesamiento de datos y optimización para reducir la carga computacional. En **Redes neuronales profundas**: Se buscan **matrices dispersas** para acelerar cálculos.

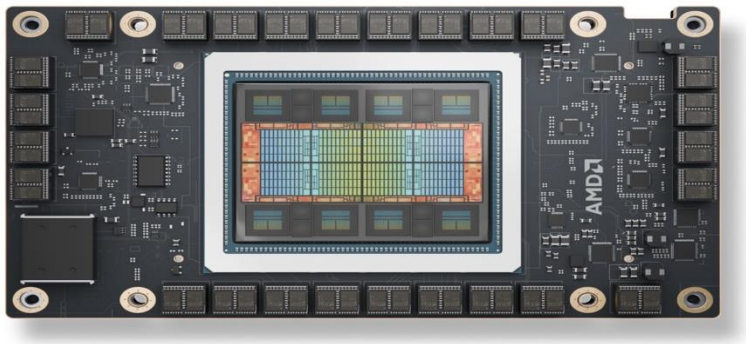


Figura 37.GPU Instinct MI325X AI Accelerator. Fuente: AMD

GPUs: Potencia computacional.

El aprendizaje profundo es un campo dentro del conjunto más amplio de disciplinas conocidas como aprendizaje automático, que a su vez es un subconjunto de la inteligencia artificial.

Implica el uso de modelos matemáticos complejos, conocidos como redes neuronales, que extraen información de datos suministrados.
Para extraer dicha información las redes neuronales necesitan ser "entrenadas" con patrones.

En la década de 2000, era evidente que las GPU eran ideales para tales tareas. Sin embargo, Nvidia apostó por una expansión significativa del mercado del aprendizaje profundo y agregó **una característica adicional** a su arquitectura Volta para destacar en este campo. Comercializados como **núcleos tensoriales**, estos eran bancos de unidades lógicas FP16, que operaban juntas como una gran matriz, pero con capacidades muy limitadas. De hecho, eran tan limitados que solo realizaban una función: multiplicar dos matrices FP16 4x4 y luego agregar otra matriz FP16 o FP32 4x4 al resultado (un proceso conocido como operación GEMM[13]).

[13] La **operación GEMM (General Matrix Multiplication)** es un **proceso fundamental** en inteligencia artificial (IA), aprendizaje profundo y computación de alto rendimiento. Se refiere a la **multiplicación de matrices generalizada**, una de las operaciones más intensivas en cómputo utilizadas en redes neuronales.
GEMM realiza la siguiente operación matricial:
C=α·A·B+β·C, Donde:
A y B son matrices de entrada.
C es la matriz de salida (puede ser inicializada con valores previos).
α y β son escalares que controlan la contribución de cada término.
Esto permite una multiplicación de matrices más flexible y eficiente en hardware optimizado.
Las redes neuronales profundas (DNN) y modelos como GPT, CNNs y Transformers dependen fuertemente de multiplicaciones de matrices. GEMM se usa para:

- Cálculo de capas densas (fully connected layers) en redes neuronales.
- Multiplicación de tensores en Transformers (como en la atención de modelos como GPT).
- Operaciones en GPU y TPUs (Tensor Processing Units) optimizadas con BLAS/cuBLAS.

El resto, sin embargo, es esencialmente un chip **SIMD** masivamente paralelo, respaldado por un sistema de memoria/caché robusto e intrincado. - SIMD = Single Instruction Multiple Data

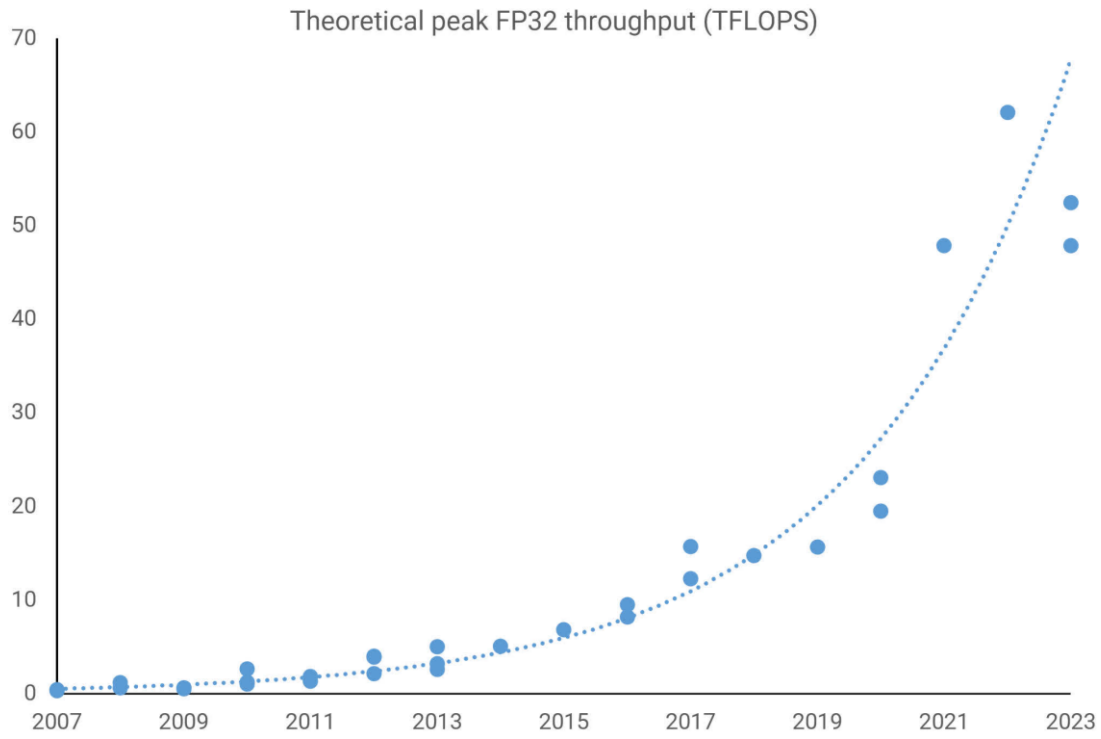


Figura 38. Evolución en el tiempo del "Rendimiento máximo teórico en FP32" en TFLOPs[14].Fuente: (Evanson, N., 2024, Techspot) Goodbye to Graphics: How GPUs Came to Dominate AI and Compute | TechSpot

FPGAs: Xilinx VITIS (AMD)

Las FPGA se utilizan para soportar aplicaciones de IA en la nube, en el centro de datos y en el borde

Una FPGA es un circuito hardware con puertas lógicas reconfigurables. Por tanto, puede mapearse cualquier circuito lógico sobre el array de puertas, permite a los usuarios crear un circuito personalizado mientras el chip se implementa en el campo (no solo durante la fase de diseño o fabricación), sobrescribiendo las configuraciones de un chip.

Por ejemplo, en una red neuronal, Supongamos una capa completamente conectada (fully connected layer): **Y = X.W + B**; donde **X** es la matriz de entrada (batch de datos); **W** es la **matriz de pesos** de la red neuronal; **B** es el vector de sesgo, **Y** es la salida.

[14] **TFLOP** (Tera Floating Point Operations Per Second) es una unidad de medida del rendimiento computacional, especialmente en GPUs, CPUs y TPUs utilizadas en inteligencia artificial, videojuegos y supercomputación. **TFLOP = 1 billón de operaciones en punto flotante por segundo**. Para una GPU o CPU, el cálculo aproximado de TFLOPS se hace así: TFLOPS = núcleos x Frecuencia (GHZ) x Instrucciones FLOP por ciclo

TFLOPS = núcleos x Frecuencia (GHZ) x Instrucciones FLOP por ciclo

Con un chip FPGA, puede crearse casi todo, desde puertas lógicas simples de una sola función hasta procesadores de múltiples núcleos. El proceso de diseño lógico y simulación es muy similar a un ASIC, pero en lugar de mandar el layout a la foundry, se mapea sobre el circuito físico en campo (Field Programmable)

¿Para qué se utilizan las FPGA?

Existe una amplia gama de aplicaciones de FPGA. Puede configurarse una FPGA con miles de unidades de memoria. **Esto permite que los circuitos funcionen en un modelo de computación masivamente paralela, como las GPU**. Con las FPGA, se obtiene acceso a una arquitectura adaptable que permite optimizar el rendimiento. **Esto significa que puede usarse FPGA para igualar o superar el rendimiento de las GPU**

VENTAJAS	INCONVENIENTES
FLEXIBILIDAD: la **reprogramabilidad** es el mayor beneficio de FPGA para el aprendizaje profundo y agrega una flexibilidad significativa a las operaciones. Pueden programarse bloques individuales o el circuito completo para que se ajuste a los requisitos de un algoritmo particular. Si la programación no se ajusta tan bien como esperaba, se puede modificar según sea necesario.	**PROGRAMACIÓN:** la programación de circuitos FPGA requiere una experiencia significativa y no es fácil de obtener. Por ejemplo, los programadores deben estar familiarizados con el lenguaje descriptivo de hardware (HDL). La falta de programadores experimentados puede dificultar la adopción confiable de FPGA.
PARALELISMO: Se puede cambiar entre programas para adaptarse a cargas de trabajo cambiantes con una FPGA. También se pueden manejar múltiples cargas de trabajo sin sacrificar el rendimiento. Esto permite trabajar en diferentes etapas de tareas simultáneamente, lo que no puede hacer con GPU.	**COMPLEJIDAD DE IMPLEMENTACIÓN:** la implementación de FPGA para aprendizaje profundo es relativamente inédita y puede tener demasiado riesgo para organizaciones conservadoras. La falta de apoyo y el conocimiento mínimo de la comunidad significa que los FPGA aún no son ampliamente accesibles para DL.
MENOR LATENCIA: los anchos de banda de memoria más grandes dan como resultado una latencia menor que las GPU. Esto permite procesar cantidades significativas de datos en tiempo real, incluidos los datos de transmisión	**COSTO**: el costo de los FPGA en sí mismos en combinación con los costos de implementación y programación hacen que los circuitos sean una inversión considerable. Especialmente para proyectos pequeños sensibles al coste.
EFICIENCIA ENERGÉTICA: los requisitos de energía más bajos para FPGA pueden ayudar a reducir el consumo de energía general para las implementaciones de aprendizaje automático y aprendizaje profundo.	**FALTA DE BIBLIOTECAS:** actualmente, hay muy pocas bibliotecas de aprendizaje profundo que admitan FPGA sin modificaciones. LeFlow es un proyecto que intenta crear compatibilidad entre FPGA y TensorFlow.

Tabla 6. Ventajas e inconvenientes de las FPGAs aplicadas a la IA. Fuente: FPGA for Deep Learning: Build Your Own Accelerator

La reconfigurabilidad, el bajo consumo y el tiempo real hacen que los FPGA se destaquen en las tareas de inferencia. El chip FPGA debe rediseñarse para implementar mejor los diferentes requisitos cambiantes de las redes neuronales profundas (Li, Zhang, & Wang, 2020)

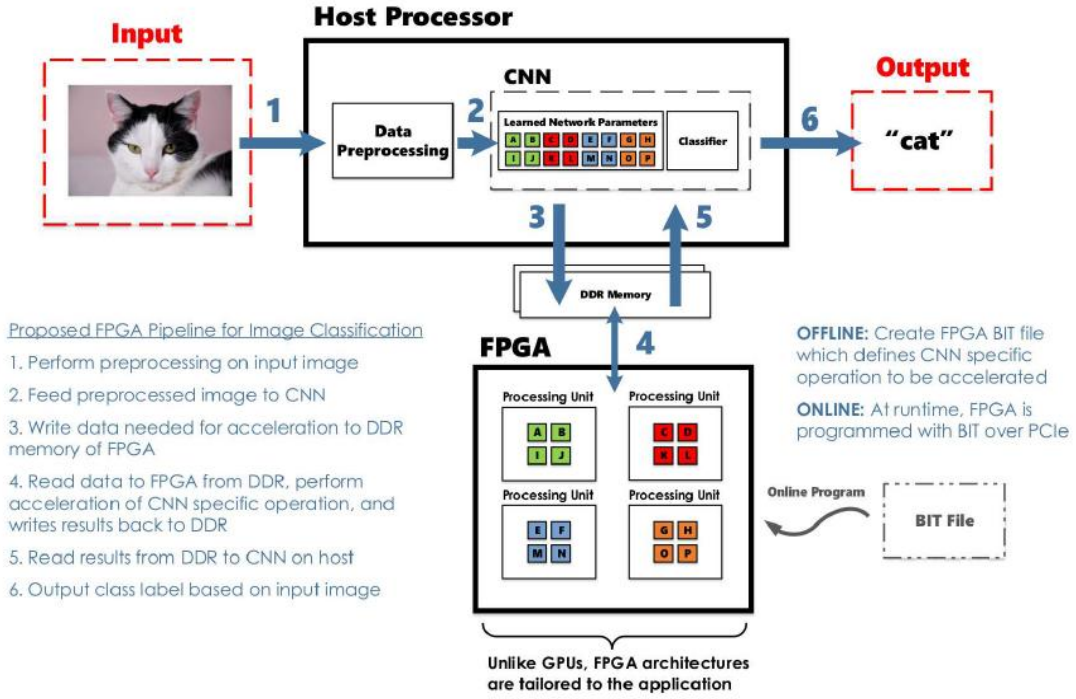


Figura 39. Flujo de implementación propuesto en (Lacey, Taylor, & Areibi, 2016) para la clasificación de imágenes utilizando FPGAs.

Tradicionalmente, al evaluar plataformas de hardware para aceleración IA, debemos inevitablemente considerar el equilibrio entre flexibilidad y rendimiento. En un extremo del espectro, los procesadores de propósito general (GPP) brindan un alto grado de flexibilidad y facilidad de uso, pero funcionan de manera relativamente ineficiente. Estas plataformas tienden a ser más accesibles, pueden producirse de manera económica y son apropiadas para una amplia variedad de usos y reutilizaciones. En el otro extremo del espectro, los circuitos integrados para aplicaciones específicas (ASIC) ofrecen un alto rendimiento a costa de ser muy poco flexibles y más difíciles de producir. Estos circuitos están dedicados a una aplicación específica y su producción es costosa y requiere mucho tiempo.

Los FPGA sirven como un compromiso entre estos dos extremos. Pertenecen a una clase más general de dispositivos lógicos programables (PLD) y son, en el sentido más simple, un circuito integrado reconfigurable. Como tal, brindan los beneficios de rendimiento de los circuitos integrados, con la flexibilidad reconfigurable de los GPP. En un nivel bajo, los FPGA pueden implementar lógica secuencial mediante el uso de flip-flops (FF) y lógica combinacional mediante el uso de tablas de búsqueda (LUT). Los FPGA modernos también contienen componentes reforzados para funciones de uso común, como núcleos de procesador completos, núcleos de comunicación, núcleos aritméticos y RAM de bloque (BRAM). Además, las tendencias actuales de FPGA tienden hacia un enfoque de diseño de sistema en chip (SoC), donde los coprocesadores ARM y los FPGA se encuentran comúnmente en la misma estructura.

Para el aprendizaje profundo, los FPGA brindan un potencial obvio para la aceleración más allá de lo que es posible en los GPP tradicionales. La ejecución a nivel de software en los GPP se basa

en la arquitectura tradicional de Von Neumann, que almacena instrucciones y datos en la memoria externa para ser recuperados cuando sea necesario. Esta es la motivación para los cachés, que alivian gran parte de las costosas operaciones de memoria externa [8]. El cuello de botella en esta arquitectura es la comunicación entre el procesador y la memoria, que perjudica gravemente el rendimiento de GPP, especialmente para las técnicas limitadas a la memoria que se requieren con frecuencia en el aprendizaje profundo. En comparación, las celdas lógicas programables en los FPGA se pueden utilizar para implementar la ruta de datos y control que se encuentra en las funciones lógicas comunes, que no dependen de la arquitectura de Von Neumann. (Lacey, Taylor, & Areibi, 2016). También son capaces de explotar la memoria distribuida en chip, así como altos grados de paralelismo de canalización, que encajan naturalmente con los métodos de aprendizaje profundo. En la Fig. 39 se da un esquema de clasificación de imágenes que se entrena en el host y una vez entrenado se mapea sobre una FPGA donde se ejecutarán las clasificaciones en inferencia.

Los FPGA modernos también admiten la **reconfiguración dinámica parcial**, donde parte del FPGA se puede reprogramar mientras se utiliza otra parte del FPGA. Esto puede tener implicaciones para los grandes modelos de aprendizaje profundo, donde las capas individuales se pueden reconfigurar en el FPGA sin interrumpir el cálculo en curso en otras capas. Esto permitiría adaptar modelos que pueden ser demasiado grandes para caber en un solo FPGA y también aliviaría las costosas lecturas de memoria global al mantener los resultados intermedios en la memoria local.

En la Fig. 40 se muestra un esquema del diseño de un clasificador de imágenes con IA (DNN). El entrenamiento se realiza con un procesador “host” que puede ser un sistema con GPUs y una vez entrenado se traslada la CNN a la FPGA mediante las herramientas EDA comerciales que existen al efecto. Las principales FPGAs del mercado son AMD (Xilinx) e Intel (Altera). Xillinx y Altera han dominado más del 85 % del mercado de FPGAs durante tres décadas hasta que fueron adquiridas por los principales fabricantes de procesadores.

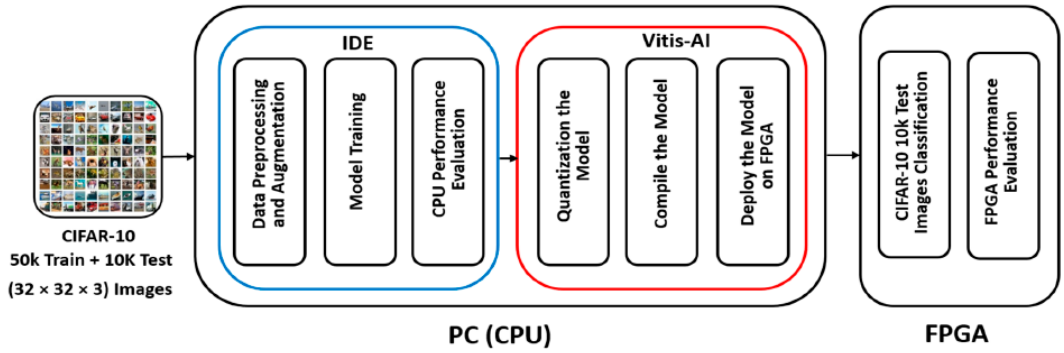


Figura 40. Flujo de diseño para implementar modelos VGG16 y VGG19 sobre FPGA con Xilinx-Vitis IA.
Fuente: : (Khaki & Choi, 2025)

La Figura 41 ilustra el flujo de trabajo propuesto por (Khaki & Choi, 2025) para implementar modelos CNN preentrenados, específicamente VGG16 y VGG19, en FPGA utilizando el marco Xilinx Vitis-AI (Van Maarseveen, Julio 2023).

Para poder trasladar los diseños desde los entornos de IA, como Jupyter, utilizando Tensor Flow en el PC del usuario, se puede utilizar la plataforma Xilinx Vitis-AI.

Xilinx Vitis-AI es una plataforma para implementar modelos de aprendizaje automático en hardware de Xilinx, como los FPGA

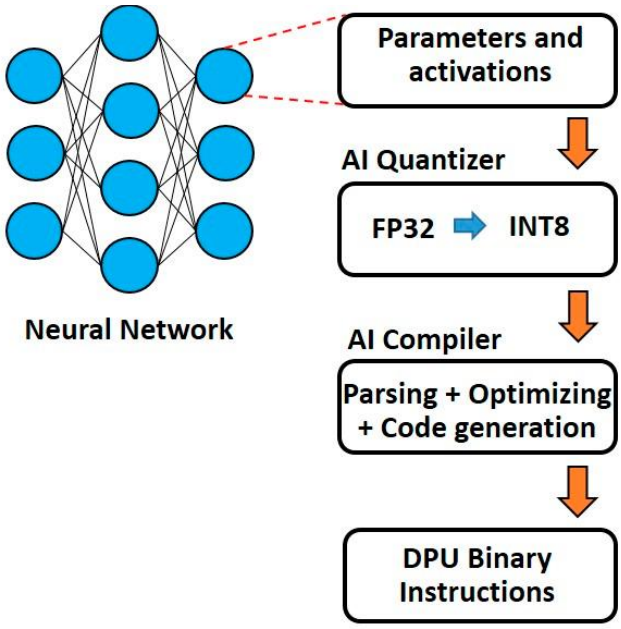


Figura 41. Proceso conceptual de Xilinx Vitis-AI. Fuente: (Khaki & Choi, 2025)

ASICS: Chips dedicados.

Un ASIC, en general, es un circuito integrado (chip) optimizado y personalizado para una aplicación específica. Es el acrónimo de **A**pplication **S**pecific **I**ntegrated **C**ircuit.
ASICs para IA son circuitos específicamente diseñados para cargas de trabajo de Inteligencia Artificial.
Los ASICs pueden optimizarse para tareas de IA específicas, alcanzando mejores prestaciones y eficiencia energética comparados con la GPUs y CPU.

Los ASICs suelen optimizarse a nivel de hardware, procurando que ocupen la mínima área de silicio, el menor consumo de energía posible, altas prestaciones y mínimo coste.

Los ASICs son ideales para las tareas de inferencia. En efecto, son chips, especializados para desplegar y ejecutar modelos de aprendizaje automático entrenados, habilitando las predicciones en tiempo real y toma de decisiones en varias aplicaciones.
A diferencia de los chips de entrenamiento que se enfocan en optimizar el proceso de entrenamiento de las redes neuronales, los chips de inferencia se adecúan a procesar de forma eficiente los datos de entrada a través de modelos pre-entrenados para **generar predicciones rápidas y fiables.**

Para el diseño de **chips de inferencia**, es necesario considerar principalmente: **Baja latencia, eficiencia energética, y flexibilidad de despliegue.**

Al personalizar los chips para tarea específicas, tanto los ASICs como las FPGAs implican ciclos de diseño largos y costosos, además los ASICs pueden resultar en mejores precios al tener un hardware optimizado, pero con poca flexibilidad para implementar cambios.

DNN

Una de las principales razones del auge de las DNN es que la capacidad de procesamiento es más rápida y barata. La capacidad de procesamiento ha marcado la diferencia a la hora de lograr una convergencia rápida.

Otro factor clave es la **disponibilidad de grandes conjuntos de datos**, que las redes neuronales profundas necesitan para aprender de forma eficaz. A medida que las empresas generan más datos, las DNN pueden descubrir patrones complejos que los modelos tradicionales no pueden manejar.

Su capacidad para procesar datos no estructurados como texto, imágenes y audio también ha abierto nuevas aplicaciones en áreas como chatbots, sistemas de recomendación y análisis predictivo. En la Fig. 42 puede verse un esquema de una red neuronal profunda (DNN) (con varias capas "escondidas" (capas entre la capa de entrada y la de salida)). Los chips Aceleradores de AI están formados mayoritariamente por DNN.

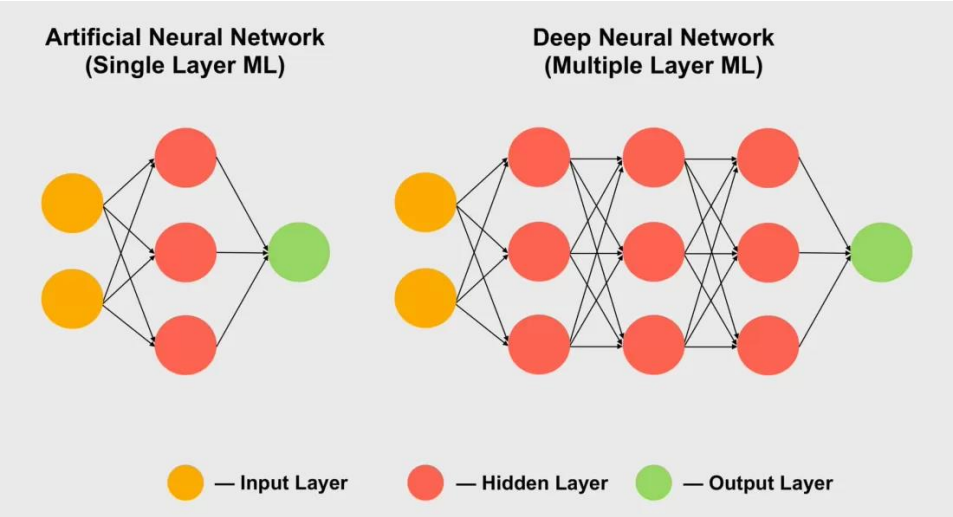


Figura 42. Representación de una Red Neuronal Artificial (una sola capa entre la capa de entrada y la capa de salida) y una Red Neuronal Profunda (DNN) con múltiples capas. Fuente: Deloitte

CHIPS NEUROMÓRFICOS

La computación neuromórfica es un modelo de computación de última generación que aprovecha neuronas artificiales inspiradas en la biología para realizar tareas complejas. Como es natural, para diseñar los circuitos neuromórficos es importante comprender bien cómo funcionan las neuronas biológicas. Hay cientos de miles de millones de neuronas en un cerebro humano. Cada neurona está conectada a otras neuronas a través de sinapsis[15], que forman un enorme bucle de neuronas, transmitiendo señales de manera distribuida y concurrente. La potencia computacional es extremadamente fuerte. La estructura de una neurona del cerebro humano puede verse en Fig. N+20A. Así pues, la estrategia de diseño de los chips neuromórficos es la de concebir un hardware que imite, de alguna forma, las sinapsis del cerebro humano, que es muy diferente de la arquitectura Von Neuman de la mayoría de CPUs, integrando plenamente

[15] La **sinapsis** es el proceso mediante el cual las neuronas se comunican entre sí para transmitir información. Conocemos dos tipos de sinapsis en función de cómo se envía el mensaje:

- **La sinapsis eléctrica:** es una conexión física entre las neuronas que permite que las señales eléctricas se propaguen directamente de una neurona a otra.
- **La sinapsis química**: es el tipo más común de sinapsis, en el que los neurotransmisores se liberan en la hendidura sináptica para comunicar información entre las neuronas.

la memoria, la unidad de procesamiento y los componentes de comunicaciones. El procesamiento de la información se realiza de forma totalmente local. Siempre que las neuronas reciban pulsos de otras neuronas, actuarán simultáneamente para comunicarse entre sí de forma fácil y rápida. Al utilizar el método de cálculo neuronal similar al cerebro humano, el consumo de energía es bajo y la tolerancia a fallos es alta. En comparación con la computadora digital tradicional, la inteligencia será más fuerte y el aprendizaje cognitivo, la organización automática y el procesamiento integral de información difusa también avanzarán un gran paso

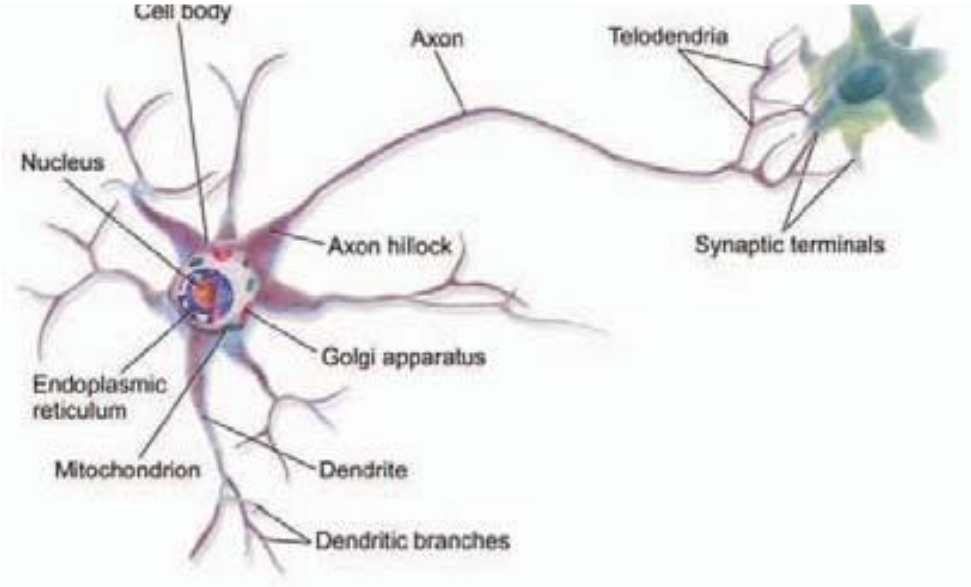


Figura 43. Estructura de una neurona del cerebro humano. Fuente: (Li, Gu, & Jiang, 2019)

A diferencia de las neuronas que se encuentran dentro de las redes neuronales artificiales (ANN) tradicionales, que generan en un rango continuo, las neuronas neuromórficas que se encuentran dentro de las **redes neuronales de picos** (**SNN**) generan picos discretos a lo largo del tiempo, de manera similar a la activación individual de las neuronas biológicas. Este modelo de computación de bajo tamaño, peso y potencia (SWaP), cuando se implementa en hardware neuromórfico especializado y se combina con sensores neuromórficos de bajo consumo, hace que la computación neuromórfica sea un candidato ideal para aplicaciones de computación de borde. Sin embargo, entrenar modelos neuromórficos es un desafío debido a la escasez de conjuntos de datos neuromórficos de calidad. (Baietto & Bihl, 2025)

(Baietto & Bihl, 2025) presentan un enfoque independiente de la plataforma para crear conjuntos de datos neuromórficos sintéticos con una red generativa antagónica condicional (CGAN). Los modelos neuromórficos entrenados en los conjuntos de datos generados tienen un rendimiento comparable al de aquellos entrenados en el conjunto de datos IBM DVSGesture original. Muestran que la generación de conjuntos de datos neuromórficos produce muestras de calidad que pueden ayudar aún más al desarrollo y la implementación de modelos de computación neuromórfica

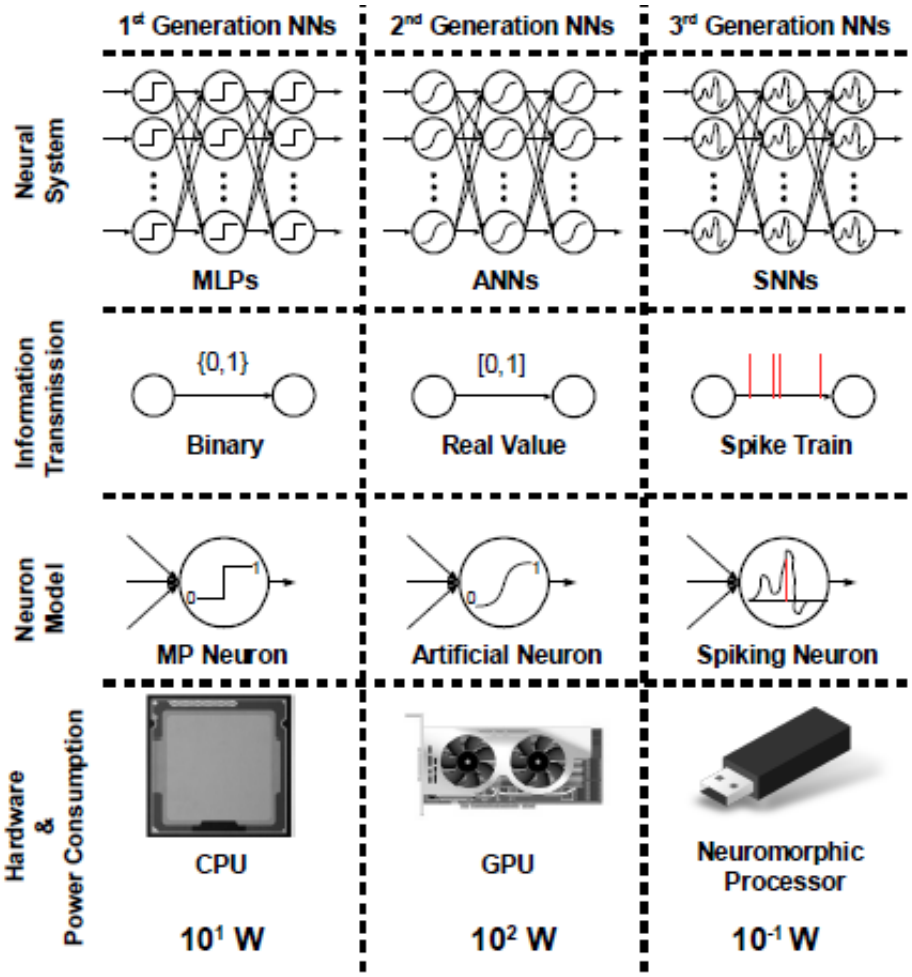


Figura 44. Comparación de las características de las diferentes generaciones de redes neuronales.
Fuente: (Baietto & Bihl, 2025)

Las ANN se implementan tradicionalmente en unidades de procesamiento gráfico (GPU) y unidades de procesamiento tensorial (TPU) de propósito general, y si bien este hardware destaca por sus cálculos de operaciones matriciales considerables requeridos para las ANN, su **consumo de energía sustancial** hace que la implementación en el borde sea un desafío. Otro obstáculo para la combinación de computación en el borde y las ANN son los sensores y los datos posteriores utilizados para entrenar y ejecutar las ANN. Los sensores tradicionales, como las cámaras, están desconectados del cálculo de ANN porque los datos deben primero capturarse, procesarse y luego, finalmente, calcularse, donde cada uno de estos pasos genera un consumo de energía adicional. Estos datos capturados suelen procesarse previamente con varias transformaciones, incluidas rotaciones, filtrado y escalado, en hardware separado tanto del dispositivo de captura como de la ANN. Finalmente, los datos procesados deben enviarse a la ANN, donde se pueden calcular. Estos pasos individualizados ralentizan drásticamente el entrenamiento y la inferencia de la ANN, así como también aumentan el consumo total de energía.

La computación neuromórfica se inspira en la intrincada dinámica de las neuronas biológicas. A diferencia de la computación de IA tradicional que opera en la aproximación de funciones utilizando álgebra lineal, la computación neuromórfica opera según el principio de impulsos

discretos o pulsos de actividad, imitando la comunicación asíncrona observada en el cerebro por los fisiólogos.

Una **red neuronal de impulsos (SNN)** es un tipo de red neuronal artificial que imita la forma en que las neuronas biológicas se comunican mediante impulsos eléctricos discretos o picos. A diferencia de las redes neuronales artificiales (ANN) tradicionales, que se basan en valores numéricos continuos, las SNN procesan y transmiten información de forma asíncrona a través de impulsos, lo que las hace biológicamente más realistas y energéticamente eficientes.

¿Cómo funcionan las redes neuronales de impulsos? Las SNN se basan en neuronas que se comunican mediante impulsos (spikes), al igual que en el cerebro humano. Esto es diferente de las ANN estándar, que utilizan multiplicaciones de matrices con números de punto flotante.

- Neuronas con impulsos: Cada neurona en una SNN integra las señales entrantes a lo largo del tiempo y dispara un impulso (spike) solo cuando se alcanza un cierto umbral. Este procesamiento impulsado por eventos hace que las SNN sean más eficientes en términos de computación y consumo de energía.
- Tres fases de activación neuronal:
    - **Integración** → La neurona acumula señales de entrada (excitatorias e inhibidoras).
    - **Activación** → Si el voltaje acumulado cruza un umbral, la neurona emite un impulso.
    - **Período refractario y de reinicio** → Después de la activación, la neurona se reinicia y se vuelve temporalmente inactiva.

¿Cómo las neuronas generan los impulsos? Hay varios modelos matemáticos de las neuronas que generan impulsos.

Una tecnología de inteligencia artificial bio inspirada llamada SNN se ha convertido en una tecnología de vanguardia en los últimos años. Por lo tanto, las implementaciones de neuronas que generan impulsos también imitan las células neuronales naturales reales. Con base en el modelado de la estructura celular, hay **cuatro modelos de neuronas que generan impulsos ampliamente explorados en implementaciones de aceleradores IA**: i) Modelo Hodgkin-Huxley (HH) ii) Modelo Izhikevich , iii) Modelo Leaky Integrate-and-Fire (LIF) y iv) Modelo Integrate-and-Fire (IF). (Isik, 2023)

i) Modelo Hodgkin-Huxley (HH)

Este modelo simplificó la estructura de la célula neuronal como un modelo de circuito RC como se muestra en la Figura 45

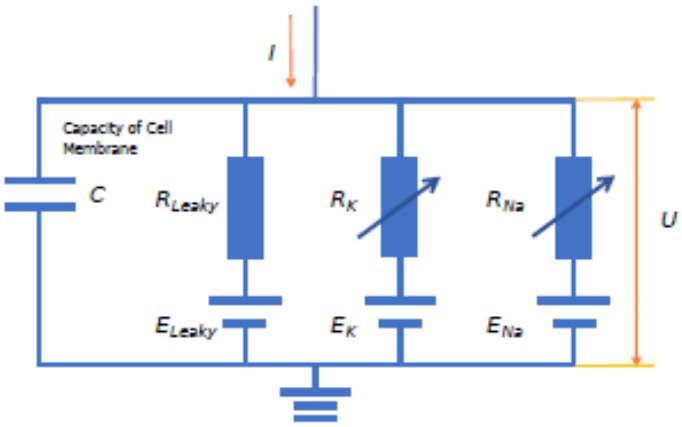


Figura 45. Modelado de circuito RC del modelo de neurona generadora de impulsos de Hodgkin-Huxley. Fuente: (Isik, 2023)

En este modelo, C representa la capacidad de la membrana celular. RK y RNa imitan los canales de iones de sodio y potasio en las neuronas. RLeaky simula el canal con fugas para perder la carga en la membrana.

Modelo Leaky Integrate-and-Fire (LIF)

Para simplificar aún más el modelo de neurona generadora de impulsos, se ha realizado la simplificación del modelo HH como modelo Leaky Integrate-and-Fire (LIF), que ha sido el modelo más utilizado en la investigación reciente. Como se muestra en la Figura 46, LIF simplificó aún más el modelo HH, la fórmula de modelado RC es la Ecuación

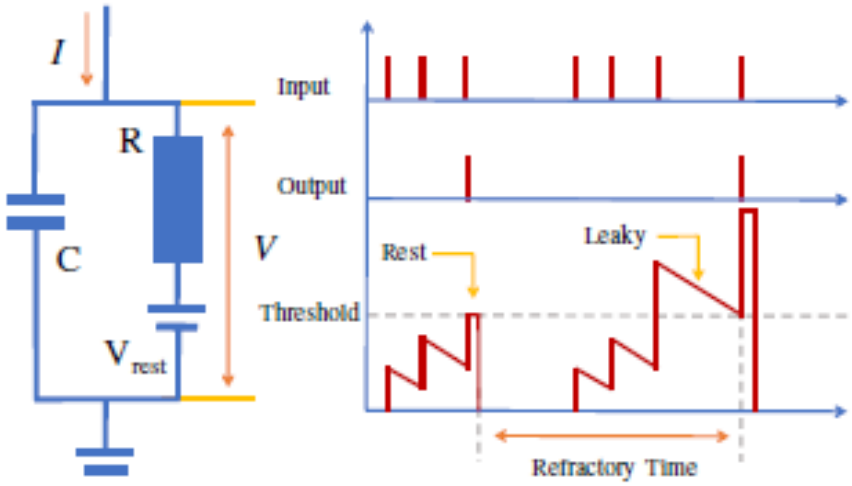


Figura 46. Modelado de circuito RC del modelo de neurona generadora de impulsos de integración y disparo con fugas (LIF). Fuente: (Isik, 2023)

$$C\frac{\mathrm{d}U}{\mathrm{d}t} = I\,(t) - \frac{V - V_{rest}}{R}$$

Reordenando la ecuación,

$$\tau_m \frac{dV}{dt} = -(V - V_{rest}) + I(t)$$

Donde I(t) es la corriente de entrada, V_{rest} es el potencial de reposo y τ_m = RC, la constante de tiempo de la membrana. Este modelo es más realista puesto que introduce "fugas" o “pérdidas”, lo que significa que la neurona pierde gradualmente el potencial almacenado si no se activa.

La Figura 48 muestra la estructura de los elementos de procesamiento neuronal en la implementación de hardware de las SNN. Para las SNN, cuando las redes aplican diferentes modelos neuronales y esquemas de codificación de señales, las estructuras de los elementos de procesamiento neuronal serán diferentes. Por ejemplo, en las implementaciones basadas en el modelo LIF y esquemas de codificación de velocidad, las neuronas pueden constar de los siguientes submódulos: 1) Multiplicador, 2) Acumulador, 3) Umbrales y 4) Codificador de impulsos.

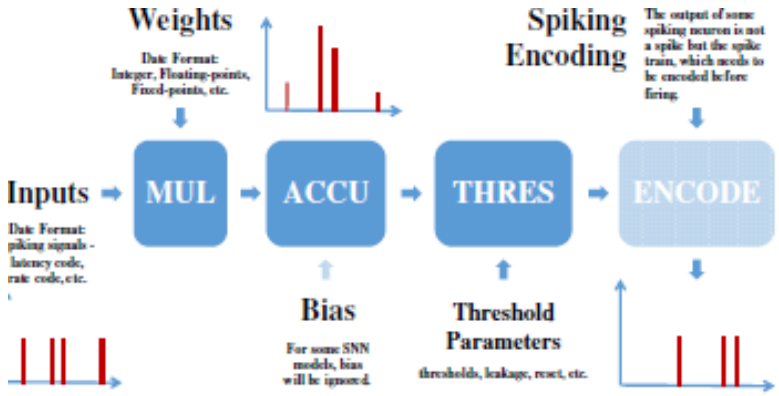


Figura 48. Estructura de neurona de las redes neuronales generadoras de impulsos. Fuente: (Isik, 2023)

La implementación electrónica de las SNN, puede hacerse mediante: a) una implementación digital (FPGA o ASIC) (Las neuronas de la SNN pueden implementarse como en la Figura 48 y b) Un circuito analógico neuromórfico (CMOS/memristor).

1. **Implementación digital con FPGA**.

- Las neuronas están representadas por circuitos lógicos que actualizan el potencial de membrana.
- La propagación de impulsos se maneja mediante señales digitales.
- Los pesos sinápticos se almacenan en la memoria y se actualizan digitalmente.

Ventajas:

1. Altamente flexible (arquitectura SNN programable).
2. Se puede optimizar para la velocidad (computación paralela).
3. Más fácil de integrar con el software de IA moderno.

Desventajas:

1. Mayor consumo de energía que los circuitos analógicos
2. Requiere memoria externa para redes a gran escala.

b. **Implementación con circuitos analógicos neuromórficos**.

Los componentes clave de los circuitos analógicos neuromórficos son:

- Transistores y condensadores CMOS → Modelan el potencial de membrana de las neuronas.
- Memristores/MOSFET de puerta flotante → Implementan pesos sinápticos.
- Espejos e integradores de corriente → Manejan la acumulación de carga.
- Circuito comparador → Determina cuándo se activa una neurona.

La funcionalidad de los componentes y arquitectura es la siguiente:

- Los circuitos neuronales funcionan mediante flujo de corriente y acumulación de carga, imitando a las neuronas biológicas reales.
- Los memristores almacenan y ajustan los pesos sinápticos, lo que proporciona una memoria no volátil.
- Los impulsos se transmiten como pulsos de voltaje en lugar de señales digitales binarias.

Ventajas

1. Consumo de energía ultrabajo (~10–100 veces menor que el digital).

2. Altamente eficiente para computación paralela basada en picos.

3. Puede funcionar sin un reloj (completamente asíncrono)

Desventajas

1. Menos flexible (el hardware está especializado para SNN).

2. Más difícil de integrar con pilas de software modernas.

3. Complejidad de fabricación (requiere procesos de semiconductores especializados).

En algunos casos puede hacerse una implementación mixta de las redes neuronales generadoras de impulsos (SNN). En efecto, los circuitos neuromórficos analógicos/digitales de señal mixta representan un medio ideal para reproducir dinámicas biofísicamente realistas de sistemas neuronales biológicos en tiempo real. Sin embargo, al igual que sus contrapartes biológicas, estos circuitos tienen una resolución limitada y se ven afectados por un alto grado de variabilidad. Se han propuesto esquemas de implementaciones mixtas que minimicen dicha variabilidad. En (Baruzzi, Indiveri, & Sabatini, January 2025) se propone un sistema cuya configuración general consiste en un sensor de visión basado en eventos conectado directamente a un procesador de red neuronal de impulsos neuromórficos que emula la etapa cortical.

El **procesador de red neuronal de impulsos** es un procesador asíncrono neuromórfico dinámico (DYNAP-SE), que comprende *neuronas analógicas/digitales configurables de señal mixta* y circuitos de sinapsis.

Los resultados experimentales validan el enfoque, demostrando cómo los principios de computación neuronal pueden conducir a sistemas electrónicos de procesamiento sensorial robustos, incluso cuando se ven afectados por un alto grado de heterogeneidad, por ejemplo, debido al uso de circuitos analógicos o dispositivos memristivos.

Una de las partes más desafiantes del diseño de un sistema neuromórfico a gran escala es diseñar una red de comunicación de impulsos escalable, que sea capaz de mantenerse al día con

los requisitos de conectividad masivos que se encuentran en estos sistemas. La Tabla 1 resume los diferentes sistemas de comunicación que se encuentran en el hardware neuromórfico. Estos sistemas neuromórficos pueden escalar de manera eficiente a tamaños mayores que los de las computadoras von Neumann, ya que almacenan información con el elemento computacional, lo que elimina el cuello de botella de von Neumann. Los sistemas neuromórficos pueden escalarse horizontalmente con un acoplamiento flexible de placas. (*Young, Dean, Plank, & Rose, 2019)*

Chip neuromórfico	Conexionado	Máximo ancho de banda de la comunicación	Conexión con el host	Tipo de paquete y tamaño	Pérdida de paquetes	Tiempo de simulación
Human brain	Conexiones de larga distancia entre clústeres	No disponible		Pulsos eléctricos Hendidura (cleft) sináptica	Sí	Tiempo real
Neurogrid	Árbol de encaminamiento multicast	Transmisor: 43.4 Mspike/s Receptor: 62.5 Mspike/s Router: 1,17 Gpalabras/s	USB via FX2	Paquetes de longitud variable con identificador cola (tailword). Secuencia de palabra de 12 bits (ruta y dirección)	No	Tiempo real
SpiNNaker	Mesh triangular 2D envuelto en forma toroidal	5 billones paquetes/s Cada nodo puede enviar a un máximo de 7.4 Gbit/s	Ethernet 100 Mbps o 1Gbps	Cabecera de 8 bits Contenido de 32 bits Datos (opcional) de 32 bits	Sí (reconfigurable)	Tiempo real
BrainScaleS	L1: intra wafer Rejilla asíncrona L2: inter wafer Árbol jerárquico	Routing en wafer: 32 Gb/s Routing entre wafers: 2.8 Geventos/s	Ethernet	L1: 6 bits nº de neurona (sin sello temporal) L2: 24 bits (eventos de impulso con sello temporal)	Sí. Pueden perderse paquetes si llegan al mismo tiempo o hay congestión de red	10^2x a 10^2x
TrueNorth	Crossbar sináptica dentro del core Rejilla asíncrona entre cores y chips	160 millones de impulsos/segundo (5.44 Gbits/s)	AXI bus a SoC PCI 2.0 (un carril)	9 bits para dx y 9 bits para dy 4 bits tick de entrega 8 bits para índice de axón 2 bits de corrección/depuración	Sí. Señal de error global cuando se pierde un paquete	Más rápido que tiempo real (1x a 21x) Sync cada ms para el siguiente paso
Loihi	Asíncrono NoC en una red mesh 2D. NoC se extiende en 4 direcciones para otros chips	3.44 Geventos/s de ancho de banda de impulso cross-section por fichero.	Ethernet USB	Escritura. Petición de lectura. Respuesta de lectura Mensaje de impulso Mensaje de barrera.	No. Utiliza barrera síncrona en un tiempo de ciclo de reloj de longitud variable	Tiempo de ciclo variable en función de la carga (más rápido que tiempo real)
Darwin	Las conexiones y pesos se almacenan en chip DRAM externo - Se usa la topología y retardos para actualizar los pesos.	No especificado	UART a USB	Longitud fija de cada paquete con el identificador de la neurona Fuente. Sello temporal de cuando se generó el paquete	No. El tiempo progresa una vez se han enviado todos los paquetes del paso anterior	Reloj de 70 MHz
Dynap-SEL	Esquema de direccionamiento mixto con etiquetas (1ª punto a punto jerárquico en chip y 2ª con difusión 2D)	Tiempo de difusión 27 ns Latencia entre chips 15,4 ns Entrada: 30 M eventos/s Salida: 21 M eventos/s	Vía FGPA	Etiqueta de 10 bits Cabecera de 6 bits (con dx, dy) Destino de 4 bits	No	Tiempo real

Figura 49. Comparación de chips neuromórficos. Fuente: (Young, Dean, Plank, & Rose, 2019)

AIMC (Analog-In-Memory-Computing)

Un paradigma de computación prometedor para abordar estos desafíos es la **computación analógica en memoria (AIMC)**, que elimina la necesidad de un movimiento constante de datos al ubicar el almacenamiento y el procesamiento en un mismo lugar, lo que **reduce significativamente el consumo de energía y acelera los cálculos**. Aprovechando las tecnologías de memoria no volátil programable, como la memoria de acceso aleatorio resistiva (RRAM), la memoria de cambio de fase (PCM) y la memoria flash, la AIMC permite un almacenamiento de alta densidad y ejecuta operaciones de acumulación múltiple (MAC) directamente dentro de las celdas de memoria, siguiendo la Ley de Ohm y la Ley de Corriente de Kirchhoff. (Li, Lammie, Le Gallo, & Rajendran, Dec 2024)

Al integrar los dispositivos emergentes en matrices de barras cruzadas resistivas (RCA), se puede ejecutar la multiplicación de matriz-vector aproximada (MVM) en el dominio analógico. Esto es prometedor porque **el cálculo es significativamente (órdenes de magnitud) más eficiente energéticamente que en el dominio digital** .El movimiento de datos también se reduce sustancialmente al almacenar la matriz en la memoria y realizar el cálculo in situ. Además, MVM es el cálculo dominante en muchas aplicaciones de IA, como el aprendizaje profundo, el procesamiento de imágenes y el análisis de gráficos.

Sin embargo, el principal desafío de aprovechar la computación analógica en memoria es que la precisión computacional puede degradarse por varias fuentes de errores y variaciones. Esto incluye errores de escritura del dispositivo, parásitos de matriz distintos de cero, rendimiento limitado del dispositivo, deriva de resistencia, variaciones de temperatura, ruido aleatorio y aguante limitado del dispositivo. (Channamadhavuni, Thijssen, Jha, & Ewetz, June 22-25, 2021)

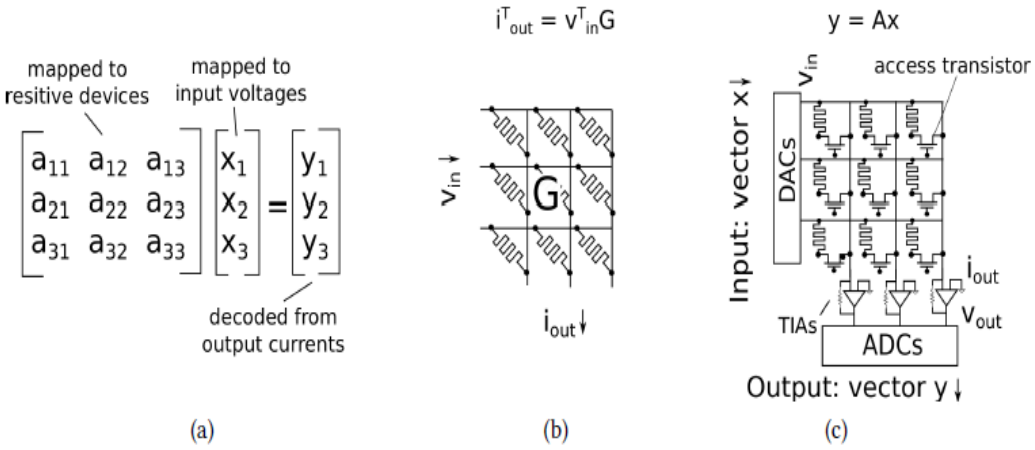


Figura 50. MVM digital (a); MVM analógica (b) y circuito RCA para MVM analógica (c). Fuente: (Channamadhavuni, Thijssen,Jha & Ewetz, 2021)

El circuito de un RCA utilizado para MVM analógico se muestra en la Figura 50(c). Hay convertidores digitales a analógicos (DAC) conectados a las líneas de palabras que se utilizan para convertir un vector de entrada digital x en voltajes de entrada analógicos vin. Los amplificadores de transimpedancia (TIA) conectados a las líneas de bits amplifican las corrientes de salida ($iout$) en voltajes de salida ($vout$), donde $vout$ = $ioutRs$ y Rs es la resistencia de retroalimentación de los TIA. En consecuencia, los voltajes de salida son iguales a $vT\ out$ = $vTi\ nGRs$. A continuación, los voltajes de salida $vout$ se convierten en un vector digital y, utilizando convertidores analógicos a digitales (ADC).

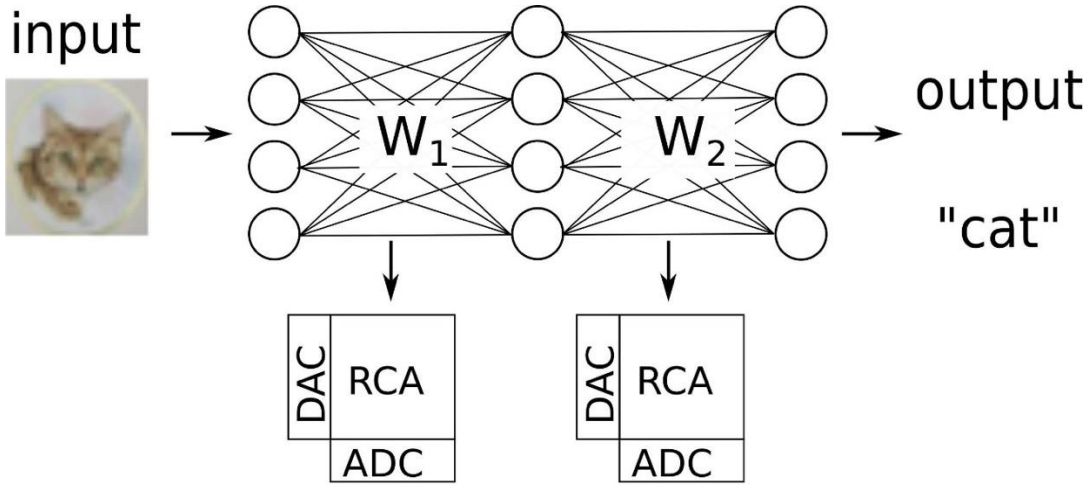


Figura 52. Implementación de inferencia de DNN utilizando RCAs. Fuente: (Channamadhavuni, Thijssen,Jha & Ewetz, 2021)

Las DNN consisten en múltiples capas de neuronas conectadas entre sí por sinapsis ponderadas (pesos), que se muestran en la Figura 52. Las redes operan utilizando una fase de entrenamiento

y una de inferencia. En la fase de entrenamiento, se aprenden los pesos de sinapsis para resolver una tarea de clasificación. En la fase de inferencia, las imágenes/objetos/videos de entrada se clasifican en una de múltiples categorías de salida al pasar una entrada a la primera capa y registrar la salida de la última capa. La evaluación de una capa de una red neuronal implica multiplicar las salidas de la capa anterior con los pesos de sinapsis (una operación MVM) y pasar el resultado a través de funciones de activación no lineales. Es atractivo ***acelerar la fase de inferencia de las DNN al mapear cada matriz de pesos a un RCA*** ya que la operación MVM y el acceso a la memoria asociada a la MVM, es el cuello de botella que limita el rendimiento del sistema. Utilizando este enfoque de alto nivel, los estudios a nivel arquitectónico han demostrado mejoras significativas en potencia, área, latencia y rendimiento.

Los inconvenientes residen en las fuentes de error asociadas a este proceso (errores de cuantización de los ADC y DAC, variaciones de temperatura, derivas de las resistencias, defectos en el componente, precisión de la escritura, etc.). Las líneas de I+D para hacer más robustas las operaciones asociadas, se centran en minimizar estas fuentes de errores, mediante compensaciones, nuevos materiales y técnicas robustas de diseño.

Aceleradores IA (ASICs) : CEREBRAS, SambaNova, TPU GOOGLE, CAMBRICON (China)

Como se ha indicado anteriormente en este estudio, los ASICs, dentro del contexto de los aceleradores de IA, son chips diseñados específicamente para ejecutar ciertas cargas de trabajo de IA de forma más eficiente que las GPUs y, por supuesto, las CPUs, a costa de una pérdida de flexibilidad, por lo que su campo de actuación se enfoca a la resolución de problemas específicos, donde tienen un mayor rendimiento que las GPUs, como por ejemplo, en: a) **Inferencia en dispositivos móviles y wearables**, donde es necesario un bajo consumo, una respuesta inmediata y sin dependencia de la nube; b) **Automoción y vehículos autónomos**, donde se necesita una latencia muy baja y una gran confiabilidad; c) **Dispositivos de Computación en el borde Industrial** (Robótica) donde es necesaria una independencia de la nube por tiempos de respuesta; d) **Centros de datos optimizados para inferencia masiva**, cuando es necesario un coste de inferencia mucho menor que con GPUs con mayor eficiencia energética; e) **Equipos médicos portátiles o de diagnóstico rápido,** con procesamiento de IA en tiempo real, tamaño reducido, muy baja disipación de calor. Estos son algunos de los campos específicos en los que se usan cada vez más los ASICs , dentro del contexto de los aceleradores de IA, que las CPUS por su mejor rendimiento, consumo energético y costes.

Si nos fijamos en su enfoque y arquitectura podemos distinguir varios tipos de ASICs especializados como aceleradores de IA. Entre los tipos más destacados podemos citar:

A. **ASICS para entrenamiento masivo** (Data center AI ASICs):

Su diseño se basa en arquitecturas masivamente paralelas con gran ancho de banda de memoria. Son muy utilizados para entrenamiento de LLMs y modelos multimodales a gran escala. Tienen ventajas importantes en HPC y Deep Learning. En este tipo de ASICs pueden considerarse, entre otros, “Cerebras Wafer-scale engine” y “Sambanova Reconfigurable Dataflow Unit”

CEREBRAS

Entre los ASICs dedicados podemos destacar como importante innovación la integración a nivel de oblea que ha propuesto la compañía CEREBRAS. (Product - Chip - Cerebras).

La integración a escala de oblea (WSI) es una técnica de fabricación de semiconductores donde se utiliza una oblea de silicio completa como un solo circuito integrado (CI) en lugar de dividirse en chips individuales.

En un enfoque tradicional y adoptado universalmente las obleas se dividen en chips más pequeños, que se encapsulan por separado.

Enfoque WSI: toda la oblea funciona como un gran circuito electrónico, lo que reduce la necesidad de interconexiones entre chips separados.

Ventajas de la integración a escala de oblea (WSI):

2. Mayor rendimiento

1. Elimina los retrasos de interconexión de chip a chip, ya que los componentes están directamente integrados en la propia oblea.
2. Permite la transferencia de datos a muy alta velocidad debido a rutas eléctricas más cortas.
3. **Ideal para IA**, computación de alto rendimiento (HPC) y **procesamiento paralelo** a gran escala.

b. **Mayor densidad de integración.**

1. Más transistores en una sola oblea permiten circuitos más complejos en menos espacio.
2. Permite el diseño de procesadores masivamente paralelos (MPP)

c. Eficiencia energética.

1. Menor consumo de energía porque los datos se mueven en distancias más cortas, lo que reduce las pérdidas resistivas
2. Reduce la necesidad de comunicación entre chips que consume mucha energía.

d. Menores costos de encapsulado e interconexión

1. No es necesario un costoso encapsulado de chips ni interconexiones externas entre circuitos integrados separados.
2. Reduce la cantidad de cables de unión y complejos diseños de PCB.

e. Arquitecturas tolerantes a fallos.

1. Puede implementar circuitos redundantes para evitar regiones defectuosas, lo que aumenta el rendimiento de la fabricación.
2. Técnicas como los circuitos de autorreparación pueden mejorar la confiabilidad.

f. Permite arquitecturas únicas.

1. Se utiliza en aplicaciones especializadas como aceleradores de IA, supercomputación y computación neuromórfica.
2. Ejemplo: Cerebras WS-3 (900.000 núcleos).

Inconvenientes de la integración a escala de oblea (WSI):

g. Alta complejidad de fabricación.

a) Requiere litografía avanzada y gestión del rendimiento**.**

b) Es difícil garantizar una calidad uniforme en una oblea grande.

2. Problemas de bajo rendimiento.

a) Incluso un pequeño defecto puede arruinar una oblea entera, lo que genera altas tasas de rechazo.

b) Requiere técnicas complejas de corrección de errores para evitar las regiones defectuosas.

3. **Desafíos de enfriamiento**.

a) Los chips a gran escala generan un calor significativo, lo que dificulta la gestión térmica.

b) Las soluciones de enfriamiento de chips estándar (disipadores de calor, ventiladores, enfriamiento líquido) pueden no ser efectivos para una oblea completa.

4. Reparabilidad y escalabilidad limitada

a) Si un chip estándar falla, se puede reemplazar. Si un dispositivo WSI falla, toda la oblea puede quedar inutilizable.

5. Costos de fabricación muy elevados.

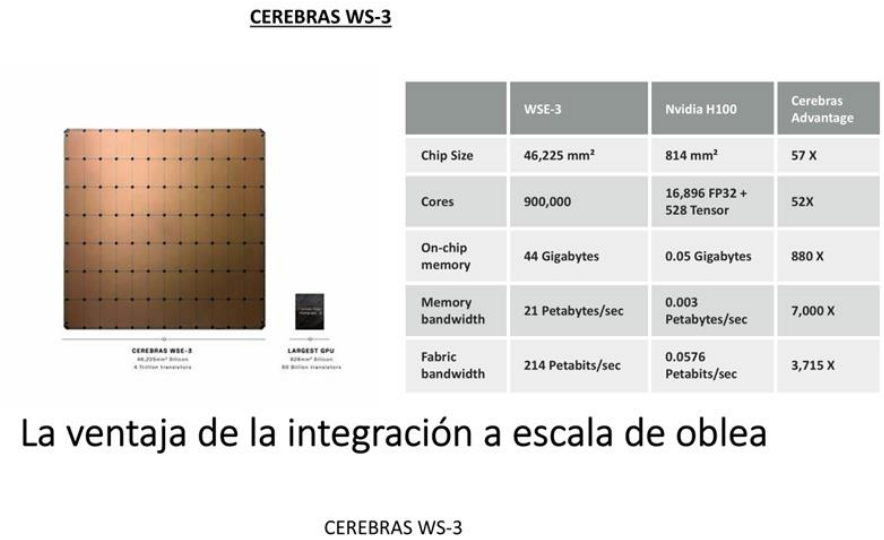
a) Requiere equipos de fabricación de alta precisión y muy costosos.

b) Solo unas pocas empresas (Como Cerebras y Tesla Dojo) pueden permitirse una fabricación de nivel WSI.

Aplicaciones de la integración a escala de oblea (WSI)

- **Inteligencia Artificial (IA) y Aprendizaje Automático (ML**): se utiliza en aceleradores de IA como Cerebras WS-2 y WS-3 para aprendizaje profundo).
- **Computación de Alto Rendimiento (HPC):** Los supercomputadores se benefician de interconexiones de alta velocidad y baja latencia.
- **Computación neuromórfica:** Se utiliza para computadores inspirados en el cerebro con millones de núcleos interconectados.
- **Vehículos autónomos:** La supercomputadora Dojo de Tesla utiliza chips de IA a escala de oblea para algoritmos de conducción autónoma.
- **Aeroespacial y defensa:** Se utiliza para el procesamiento de señales de radar y simulaciones en tiempo real.

La integración a escala de oblea (WSI) es un enfoque revolucionario para el diseño de semiconductores, que permite obtener enormes ganancias de rendimiento para IA, HPC y computación avanzada. Sin embargo, los altos costos de fabricación, los problemas de rendimiento y los desafíos de enfriamiento dificultan su adopción generalizada.



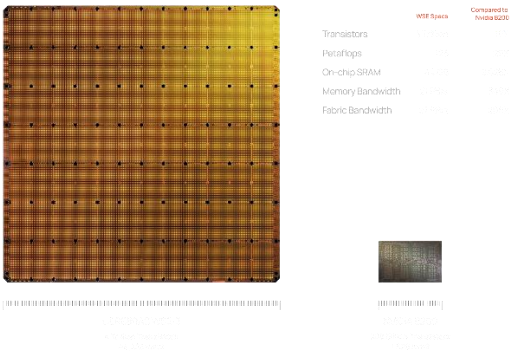


Figura 53. Cerebras WS-3. Integración a escala de oblea. Fuente: Product - Chip - Cerebras

Computación diseñada para IA

El WSE-3 incluye 900.000 núcleos de IA en un solo procesador. Cada núcleo del WSE es programable de forma independiente y está optimizado para las operaciones de álgebra lineal dispersa basadas en tensores que sustentan el entrenamiento y la inferencia de redes neuronales para el aprendizaje profundo, lo que le permite ofrecer el máximo rendimiento, eficiencia y flexibilidad.

Más memoria en chip 880x

Capacidad de memoria y ancho de banda: ¿Por qué elegir?

A diferencia de los dispositivos tradicionales, en los que la memoria caché de trabajo es diminuta, el WSE-3 toma 44 GB de SRAM en chip superrápida y los distribuye de manera uniforme por toda la superficie del chip. Esto le da a cada núcleo acceso en un solo ciclo de reloj a una memoria rápida con un ancho de banda extremadamente alto: 21 PB/s. Esto es 880 veces más capacidad y 7000 veces más ancho de banda que la GPU líder.

3715x

Más ancho de banda de estructura

Gran ancho de banda. Baja latencia.

La interconexión en oblea WSE-3 elimina la ralentización de la comunicación y las ineficiencias que supone conectar cientos de dispositivos pequeños mediante cables. Ofrece un increíble ancho de banda de interconexión entre procesadores de 214 Pb/s. Eso es más de 3715 veces el ancho de banda que se ofrece entre procesadores gráficos.

Sambanova

SambaNova, fundada en 2017, es otra empresa de diseño de chips que aborda el entrenamiento de IA con una arquitectura de chip poco convencional. Su producto estrella, el SN40L, tiene lo que la empresa llama una "*arquitectura de flujo de datos reconfigurable*" compuesta por mosaicos de memoria y recursos informáticos. Los vínculos entre estos mosaicos se pueden alterar sobre la marcha para facilitar el movimiento rápido de datos para grandes redes neuronales. En este sentido, puede afirmarse que SN40L , al tener cierto hardware reconfigurable, comparte, en parte, una filosofía de diseño parecida a las FPGAs, pero con mucha más eficiencia que éstas en el entrenamiento de los modelos IA.

Según afirma Prendki, una experta en IA de Google DeepMind, ese silicio personalizable podría resultar útil para entrenar grandes modelos de lenguaje, porque los desarrolladores de IA pueden optimizar el hardware para diferentes modelos. Ninguna otra empresa ofrece esa capacidad, afirma. SambaNova también está logrando éxitos con SambaFlow, la pila de software que se utiliza junto con el SN40L. "A nivel de infraestructura, SambaNova está haciendo un buen trabajo con la plataforma", afirma Moorhead. SambaFlow puede analizar modelos de aprendizaje automático y ayudar a los desarrolladores a reconfigurar el SN40L para acelerar el rendimiento del modelo. SambaNova todavía tiene mucho que demostrar, pero entre sus clientes se incluyen SoftBank y Analog Devices. ". (Smith, October 2024)

La nueva arquitectura SambaNova SN40L "Cerulean", se implementa en un chip de 5 nm de TSMC con tres niveles de memoria, lo que es una importante ventaja. También es una arquitectura de flujo de datos que está diseñada para ser **un chip de entrenamiento e inferencia.**

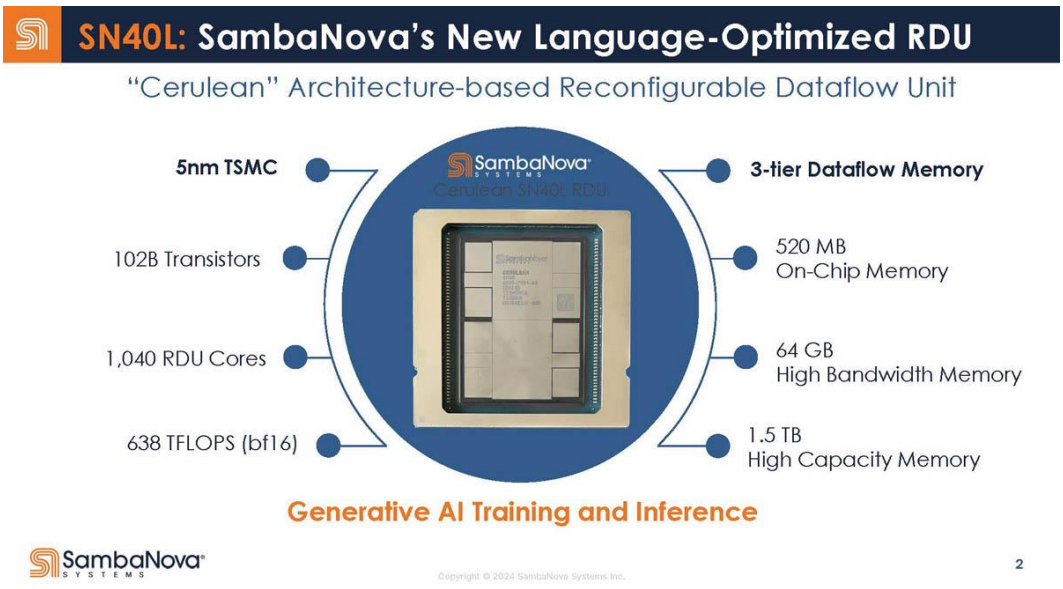


Figura 54. Arquitectura "Cerulean" de la Unidad de Flujo de Datos Reconfigurable. Fuente: : Why SambaNova's SN40L Chip is The Best for Inference

Los tres niveles de memoria son: SRAM en el chip de 520 MB. Luego hay 64 GB de HBM. Luego hay memoria DDR adicional como nivel de capacidad. SambaNova muestra aquí un sistema de 16 sockets para obtener características como 8 GB de SRAM en chip y 1 TB de HBM.

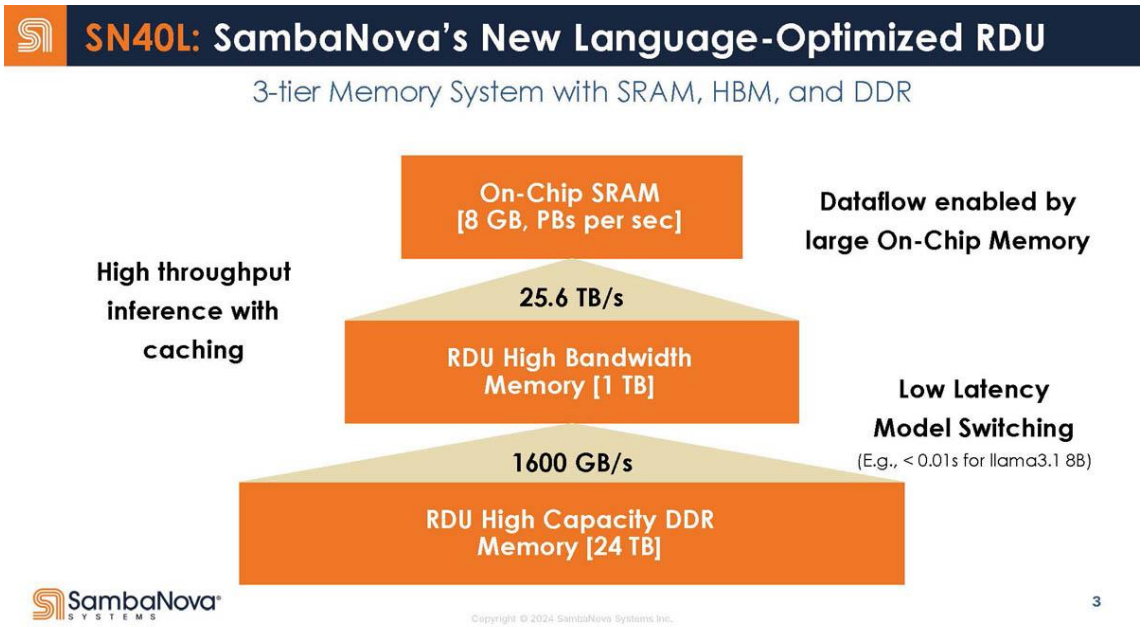


Figura 55. Sistema de la jerarquía de memorias del RDU. Fuente: Why SambaNova's SN40L Chip is The Best for Inference

Aquí se puede observar las 1040 unidades de cómputo y memoria con sus conmutadores de malla en el mosaico de SambaNova.

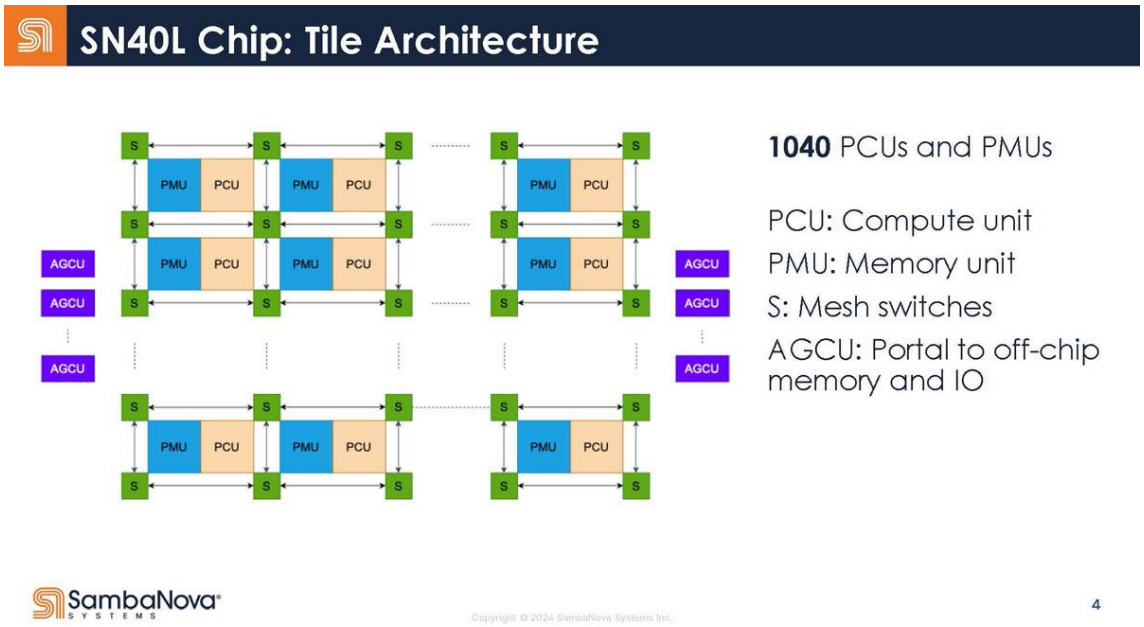


Figura 56. Arquitectura de las unidades (PCU+PMU) junto a los conmutadores de malla (S). Fuente: Why SambaNova's SN40L Chip is The Best for Inference

Aquí se muestra la unidad de cómputo. En lugar de tener una unidad de ejecución de búsqueda/decodificación tradicional, etc., tiene una secuencia de etapas estáticas. La PCU puede funcionar como una unidad de transmisión (datos de izquierda a derecha), el azul es un árbol de reducción de carriles cruzados. En una operación de cómputo matricial, se puede utilizar como una matriz sistólica.

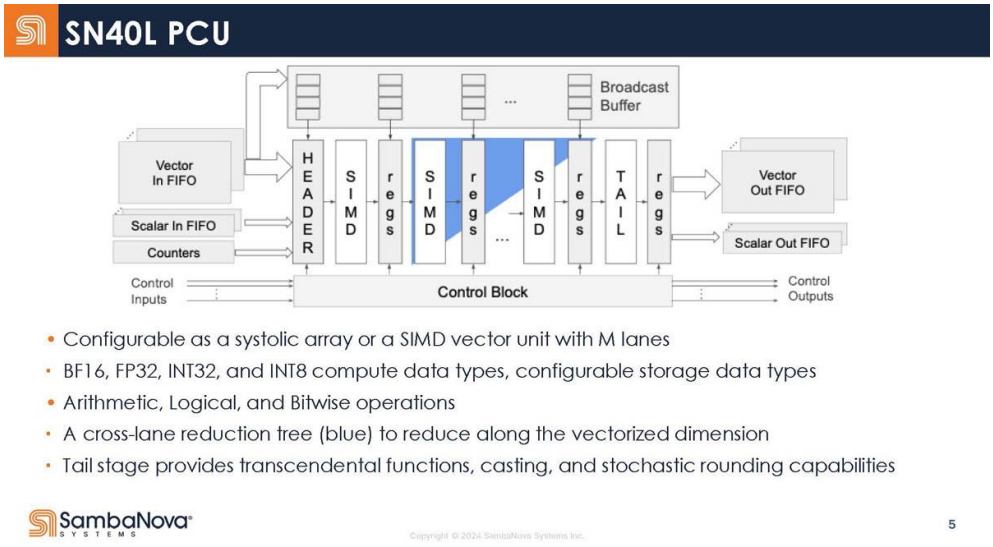


Figura 57. Arquitectura de la PCU del chipSN40L de SambaNova. Fuente: Why SambaNova's SN40L Chip is The Best for Inference

ASICs híbridos o reconfigurables para IA.

Son ASICs especializados, pero con cierta flexibilidad en su programación. Es recomendable su aplicación en los casos donde el modelo o arquitectura IA cambia con frecuencia. Su principal ventaja consiste en una mejor adaptación a nuevos modelos sin sacrificar demasiado rendimiento. Un ejemplo típico es Groq.

GROQ

Otra empresa con un giro único en el hardware de IA es Groq. El enfoque de Groq se centra en emparejar estrechamente la memoria y los recursos informáticos para acelerar la velocidad con la que un modelo de lenguaje grande puede responder a las indicaciones. “Su arquitectura se basa mucho en la memoria. La memoria está estrechamente acoplada al procesador. Se necesitan más nodos, pero el precio por token y el rendimiento son una locura”, dice Moorhead. El “token” es la unidad básica de datos que procesa un modelo; en un LLM, normalmente es una palabra o parte de una palabra.

El rendimiento de Groq es aún más impresionante, dice, dado que su chip, llamado Language Processing Unit Inference Engine, está fabricado con la tecnología de 14 nanómetros de GlobalFoundries, varias generaciones por detrás de la tecnología de TSMC que fabrica el Nvidia H100. En julio, Groq publicó una demostración de la velocidad de inferencia de su chip, que puede superar los 1.250 tokens por segundo ejecutando el LLM de 8 mil millones de parámetros Llama 3 de Meta. Eso superó incluso la demostración de SambaNova, que puede superar los 1.000 tokens por segundo.”. (Smith, October 2024).

TPUs (Tensor Processing Units).

Su arquitectura está diseñada para la optimización de operaciones de álgebra lineal (matrices, tensores). Su aplicación principal es para entrenamiento e inferencia de redes neuronales profundas. Tienen muy alta eficiencia energética en en modelos de Deep learning. Ejemplos principales con : Google TPU v6e y TPU 4e.

GOOGLE TPU

Las unidades de procesamiento tensorial (TPU) son circuitos integrados de aplicación específica (ASIC) de Google que se utilizan para acelerar las cargas de trabajo de aprendizaje automático. Las TPU están diseñadas para realizar operaciones de matrices con rapidez, lo que las hace ideales para las cargas de trabajo de aprendizaje automático. Se pueden ejecutar cargas de trabajo de aprendizaje automático en TPU con frameworks como TensorFlow, Pytorch y JAX

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.

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.

Trillium es el acelerador de IA de nueva generación de Cloud TPU. (v6e)

Con una huella de 256 chips por pod, la v6e comparte muchas similitudes con la v5e. Este sistema está optimizado para ser el producto de mayor valor para el entrenamiento, la optimización y la publicación de transformers, texto a imagen y redes neuronales convolucionales (CNN).

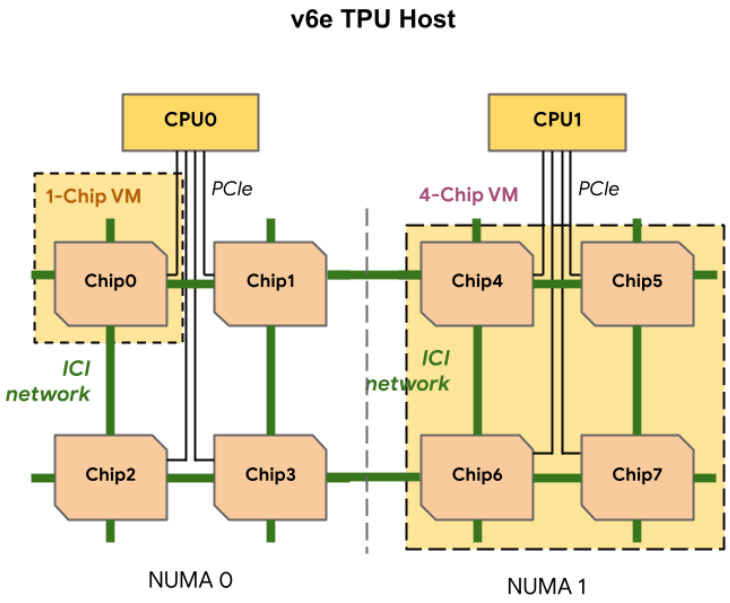


Figura 58. Arquitectura Tensor Processor Unit v6e. Fuente:Google

Cada chip v6e contiene un TensorCore. Cada TensorCore tiene 2 unidades de multiplicación de matrices (MXU), una unidad vectorial y una unidad escalar. En la siguiente tabla, se muestran las especificaciones clave y sus valores para la TPU v6e en comparación con la TPU v5e.

Especificación	v5e	v6e
Rendimiento/costo total de propiedad (TCO) (esperado)	0.65x	1
Procesamiento máximo por chip (bf16)	197 TFLOPS	918 TFLOPS
Procesamiento máximo por chip (Int8)	393 TOPS	1836 TOPS
Capacidad de HBM por chip	16 GB	32 GB
Ancho de banda de HBM por chip	819 GBps	1,640 GBps
Ancho de banda de interconexión entre chips (ICI)	1,600 Gbps	3584 Gbps
Puertos ICI por chip	4	4
DRAM por host	512 GiB	1536 GiB
Chips por host	8	8
Tamaño del pod de TPU	256 chips	256 chips
Topología de interconexión	Torón 2D	Torón 2D
Procesamiento máximo de BF16 por pod	50.63 PFLOP	234.9 PFLOP

Ancho de banda de reducción total por pod	51.2 TB/s	102.4 TB/s
Ancho de banda de bisección por pod	1.6 TB/s	3.2 TB/s
Configuración de la NIC por host	2 NIC de 100 Gbps	NIC de 4 x 200 Gbps
Ancho de banda de red del centro de datos por Pod	6.4 Tbps	25.6 Tbps
Funciones especiales	-	SparseCore

Tabla 8. Especificaciones TPU v6e , comparadas con TPU v5e. Fuente: Google.

Además de los tipos nombrados por su enfoque y arquitectura, están **los chips de IA de procesador en el borde y los chips neuromórficos**, como ASICS en el contexto de los aceleradores de IA, que se expondrán en los siguientes apartados. Como ASICs en el contexto de los aceleradores de IA, describimos, por su interés el chip Siyuan 370 para la nube desarrollados por la **empresa china Cambrian**:

CAMBRICON (China)

La tarjeta aceleradora MLU370-S4 / S8 utiliza el chip Siyuan 370, **el proceso TSMC de 7 nm**, la arquitectura de chip de inteligencia artificial de nueva generación Cámbrica MLUarch03, admite PCIe Gen4 y el consumo de energía de la placa es de solo 75 W, lo que puede proporcionar 3 veces la capacidad de decodificación y 1,5 veces la capacidad de codificación en comparación con las GPU del mismo tamaño. Las tarjetas aceleradoras MLU370-S4/S8 son energéticamente eficientes y compactas para la implementación de alta densidad en servidores.

El **Siyuan 370** es la tercera generación de chips de inteligencia artificial (IA) para la nube desarrollados por **Cambrian**. A continuación, se detallan sus principales características:

- **Proceso de fabricación**: Basado en tecnología de 7 nanómetros (nm), lo que permite una mayor densidad de transistores y eficiencia energética.
- **Tecnología Chiplet**: Es el primer chip de IA de Cambrian que utiliza tecnología **chiplet**, permitiendo una integración más flexible y escalable de componentes.
- **Transistores integrados**: Cuenta con 39.000 millones de transistores, lo que duplica la capacidad de su predecesor, el Siyuan 270.
- **Potencia de cómputo**: Alcanza hasta 256 TOPS (Tera Operaciones de Punto Flotante por Segundo) en precisión INT8, ofreciendo el doble de rendimiento en comparación con la generación anterior.
- **Arquitectura**: Implementa la última arquitectura de chip inteligente de Cambrian, **MLUarch03**, optimizando el rendimiento y la eficiencia energética.
- **Soporte de memoria**: Es el primer chip de IA en la nube en China que soporta memoria **LPDDR5**, triplicando el ancho de banda de memoria en comparación con la generación anterior y mejorando la eficiencia de acceso en un 50% respecto a GDDR6.

Tarjetas aceleradoras: Basado en el Siyuan 370, Cambrian ha lanzado dos tarjetas aceleradoras:

- **MLU370-S4**: Diseño compacto de media altura y longitud, con un consumo de energía de 75W, ideal para implementaciones de alta densidad.

- **MLU370-X4**: Formato completo con un consumo de 150W, enfocada en aplicaciones de alto rendimiento.

Estas características posicionan al Siyuan 370 como una solución avanzada para aplicaciones de IA en la nube, ofreciendo mejoras significativas en rendimiento, eficiencia y flexibilidad en comparación con sus predecesores.

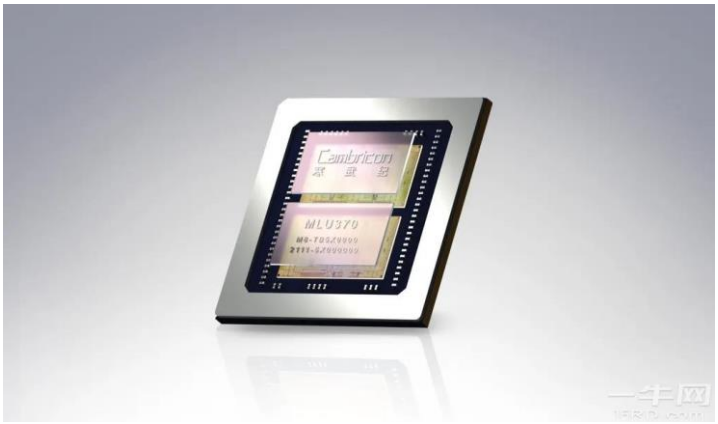


Figura 59. fotografía del procesador de CAMBRIAN MLU 370. Fuente: Cambrian launches third-generation cloud computingAIchip——Siyuan 370,7nmworkmanship,Up to 256 computing powerTOPS ...-16RD

Principales Aplicaciones de chips IA en el borde. ASICs aceleradores de IA en el borde.

En muchísimas aplicaciones militares de IA, la toma de decisiones tiene que ser en tiempo real, por lo que la mayoría de las aplicaciones que utilizan IA en la nube no pueden utilizarse por los tiempos de retardo que introducen las conexiones con los servidores de IA en la nube. En estos casos, **el procesamiento tiene que ser local**, cerca de donde se captan la mayoría de los datos. Este procesamiento local de IA es lo que se conoce como computación de IA en el borde (IA Edge Computing).

Así pues, podemos definir **Computación IA en el borde (Edge AI)** como el proceso de ejecutar modelos de **inteligencia artificial (IA) directamente en dispositivos locales** en lugar de depender de centros de datos o la nube. La Computación IA en el borde se caracteriza por,

- **Procesamiento local** → No depende de servidores remotos.
- **Baja latencia** → Respuestas más rápidas (ideal para tiempo real).
- **Menor consumo de ancho de banda** → No necesita enviar datos a la nube.
- **Mayor privacidad** → Datos sensibles no salen del dispositivo.

Para que la **computación en el borde (Edge AI)** sea eficiente, los modelos deben ser **rápidos, livianos y optimizados** para dispositivos con recursos limitados. Algunos modelos que se utilizan son:

1. **Redes Neuronales Convolucionales** (CNNs) – Visión por Computadora
   - Usadas en detección de objetos, reconocimiento facial, cámaras inteligentes y autos autónomos.

Se optimizan para correr en hardware como **Google Edge TPU, NVIDIA Jetson o Apple Neural Engine**.

Ejemplos de modelos CNN en Edge AI:

- **MobileNet** → Rápido y eficiente en teléfonos y IoT.
- **EfficientNet** → Optimizado para menor consumo de energía.
- **YOLO (You Only Look Once)** → Detección de objetos en tiempo real.

2. **Modelos de Lenguaje Natural (NLP) – Asistentes Virtuales**
3. **Modelos de Redes Neuronales Recurrentes (RNNs) – IA en IoT**
4. **Modelos de Compresión para Edge AI**

Uno de los mayores desafíos a los que se enfrentan los chips IA en el borde es el del **consumo de energía**. La mayoría de chips para IoT no pueden conectarse a fuentes de alimentación de cierta potencia, sino que se suelen alimentar con baterías muy pequeñas (pilas de botón) y además deben tener una duración importante pues no es fácil su recambio por estar situados en ambientes con accesos difíciles. Así pues, estos chips deben operar con un consumo de energía muy reducido. Un ejemplo puede encontrarse en la creciente utilización de Redes neuronales profundas (DNN) en el borde. Como se ha comentado, esto aumenta la seguridad, reduce significativamente la latencia y, fundamentalmente, habilita el procesamiento de datos en tiempo real.

Para ensayar y conciliar la necesidad de recursos limitados en consumo y las necesidades computacionales de las DNN en el borde, se ha propuesto la utilización de un chip GAP8, basado en RISC-V como motor de computación en el borde, de ultrabajo consumo. El GAP8, contiene 8 núcleos de RISC-V. Lo más importante es que GAP8 admite la implementación y aceleración de DNN. Sin embargo, existen limitaciones en el tamaño de la memoria. Los investigadores, (Zhenling, Qi, KANEKO, Li, & Meng, June 2024) proponen un algoritmo de optimización adaptativo al hardware para garantizar el funcionamiento a alta velocidad de las DNN en la IA de borde.

Los resultados experimentales muestran que después de implementar las DNN, el conteo de parámetros y la tasa de poda de operaciones de punto flotante del modelo optimizado con el algoritmo EN-HSPG alcanzan hasta el 98,82 % y el 91,28 % respectivamente, utilizando solo el 54 % de la memoria L2. Además, el tiempo de inferencia del modelo optimizado EN-HSPG es de solo 115,26 ms en GAP8. Por lo tanto, la investigación permite el funcionamiento de redes de aprendizaje profundo complejas en dispositivos de borde con capacidades informáticas limitadas

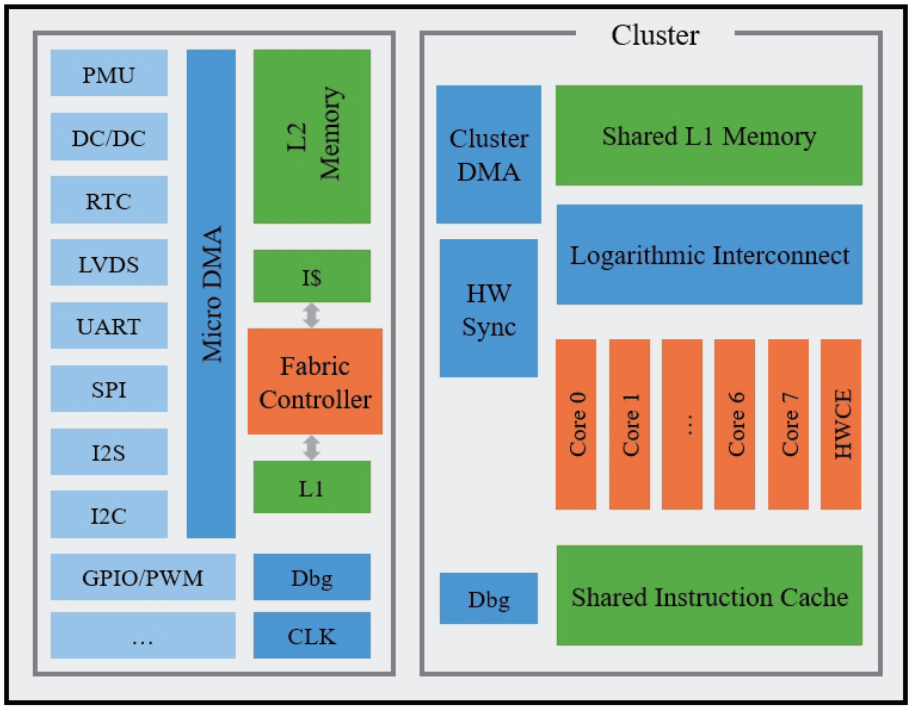


Figura 60. Arquitectura GAP 8, que incluye 8 núcleos de RISC-V (Core0,..,Core7). Fuente: (Zhenling, Qi, KANEKO, Li, & Meng, June 2024)

Otro ejemplo que muestra el método de diseño de sistemas de consumo de energía ultrabajo de la computación de borde de IA es cada vez más importante es el desarrollado por (Meng, et al., 2023): Con un suministro de energía limitado, los dispositivos de computación de borde de IA no solo necesitan completar el algoritmo de razonamiento de borde de IA en tiempo real, sino que también deben entrar en el estado de "sueño profundo" (Deep sleep State) a tiempo para reducir la pérdida innecesaria de energía. Basado en el chip de computación de borde de IA Hisilicon hi3559av100, el trabajo estudia los métodos de diseño de sistemas de consumo de energía ultrabajo como reloj, red de suministro de energía y lógica, de modo que todo el sistema pueda minimizar la pérdida de energía del sistema y cumplir con los requisitos del sistema de aplicación mientras calcula el algoritmo de IA en tiempo real (Meng, et al., 2023).

Uno de los problemas que existen actualmente con los chips IA en el borde es que se utilizan casi exclusivamente en modo de inferencia y se entrenan en la nube. Sin embargo, esto puede plantear problemas de seguridad.

En efecto, un problema de seguridad cada vez más considerable es la "pérdida" o "fuga" de los pesos y los modelos de red neuronal (NN) entrenados desde la nube y luego almacenados en dispositivos de borde, considerando los costosos recursos invertidos en el proceso de entrenamiento y los datos privados del usuario. Algunos chips de IA ya han cifrado algunos pesos e información personal en el chip, como el cifrado XNOR integrado para el cifrado "ligero". Sin embargo, los datos de los usuarios aún deben transmitirse a la nube, lo que genera preocupaciones de seguridad para los consumidores. Además, los dispositivos IoT apenas se adaptan a los cambios ambientales sin entrenamiento local. Por lo tanto, **el entrenamiento en chip** se ha convertido en una función indispensable en los futuros dispositivos IoT.

En este sentido se ha propuesto un sistema de entrenamiento on chip basado en CIM (Computer in Memory). La computación en memoria (CIM) se considera una de las principales arquitecturas de hardware de borde de IA, que puede romper el cuello de botella de Von Neumann logrando

una mayor eficiencia energética y una menor latencia. Los investigadores están explorando actualmente el CIM de punto flotante (FP) para lograr aplicaciones de IA con mayor precisión, especialmente el entrenamiento en el propio chip. Si bien las redes neuronales típicas de hoy que usan FP generalmente exceden la capacidad de memoria en chip, además, los dispositivos IoT están dirigidos a escenarios de aplicaciones livianas y portátiles. Por lo tanto, es urgente que los dispositivos de borde con recursos limitados logren un aprendizaje en chip basado en CIM (Guo, Xue, Zhou, & Zeng, 2024)

Una revisión del hardware para computación IA en el borde, así como las librerías y el software para cargar los modelos IA específicos, como Tensor Flow, se puede encontrar en (Sipola, Alatalo, Kokkonen, & Rantonen, April 2022)

Chips IA para su uso en Defensa: Proyectos de chips IA para Defensa financiados por Agencias (DARPA, EDF, ESA, NASA).

Agencias como DARPA, la Comisión Europea (a través de programas como el Fondo Europeo de Defensa - EDF), la Agencia Espacial Europea (ESA) y la NASA han financiado diversos proyectos para el diseño y desarrollo de nuevos chips semiconductores.
Proyectos más destacados:

1. **DARPA** (Defense Advanced Research Projects Agency):**Programa CHIPS** (Common Heterogeneous Integration and IP Reuse Strategies): Este programa busca establecer un nuevo paradigma en la reutilización de propiedad intelectual (IP) mediante la creación de un ecosistema de bloques IP modulares y reutilizables. Estos bloques pueden integrarse en sistemas utilizando tecnologías de integración existentes y emergentes, facilitando diseños de circuitos integrados más flexibles y reduciendo los costos y tiempos de desarrollo.
2. **DARPA.MIL Programa IDEA** (Intelligent Design of Electronic Assets): El objetivo de IDEA es desarrollar un compilador de hardware de propósito general que permita la traducción automática, sin intervención humana, de código fuente o esquemas a diseños físicos de circuitos integrados en menos de 24 horas. Esto busca acelerar el desarrollo de sistemas electrónicos de próxima generación y reducir la dependencia de grandes equipos de diseño especializados.
3. **DARPA.MIL Programa SAHARA** (Structured Array Hardware for Automatically Realized Applications): Lanzado en 2021, SAHARA tiene como objetivo ampliar el acceso a capacidades de fabricación nacional para abordar los desafíos en el desarrollo seguro de chips personalizados para sistemas de defensa. En colaboración con Intel y varias universidades, el programa busca automatizar la conversión de diseños FPGA a ASICs estructurados, mejorando el rendimiento y reduciendo el consumo energético en aplicaciones militares.
4. **NASA** (National Aeronautics and Space Administration):**Método de Evaluación de Defectos en Semiconductores**: NASA ha desarrollado un método utilizando difracción de rayos X para evaluar la concentración de defectos en la estructura cristalina de obleas semiconductoras. Este enfoque permite determinar la calidad de las obleas de manera no destructiva y en menos tiempo, mejorando la fabricación de chips para aplicaciones aeronáuticas.
5. **Comisión Europea y Agencias Europeas**: Iniciativas del Fondo Europeo de Defensa (EDF) y la Agencia Europea de Defensa (EDA): Estas agencias han lanzado convocatorias para proyectos de investigación y desarrollo en microelectrónica y semiconductores, buscando fortalecer la autonomía tecnológica europea en sectores clave como la defensa y el espacio. Aunque no se detallan proyectos

específicos en las fuentes consultadas, estas iniciativas reflejan el compromiso europeo con el avance en tecnologías de semiconductores.

6. **ESA** (Agencia Espacial Europea):**Proyectos de Desarrollo de Semiconductores para Aplicaciones Espaciales**: La ESA financia proyectos orientados al desarrollo de componentes semiconductores capaces de operar en entornos espaciales extremos, incluyendo la resistencia a la radiación y temperaturas extremas. Estos proyectos son fundamentales para garantizar la fiabilidad de los sistemas electrónicos en misiones espaciales. Estos proyectos reflejan el esfuerzo continuo de diversas agencias para impulsar la innovación en el diseño y fabricación de semiconductores, atendiendo a necesidades específicas en defensa, espacio y otras aplicaciones críticas.

Evolución y tendencias de Chips IA:

En el estudio de los chips de IA ha quedado claro que la gran aportación de las tecnologías de semiconductores al avance de la IA es el aumento constante de la potencia computacional. Sin embargo, también se ha tenido evidencia que este aumento iba parejo a un gran desafío, que es el aumento considerable del consumo de energía. Por este motivo, se han estado buscando técnicas de reducción del consumo de energía, mediante el desarrollo de nuevos algoritmos y arquitecturas hardware más eficientes energéticamente, como los **chips neuromórficos**.

Para inferencia, se ha demostrado que las GPUs son más ineficientes que los chips dedicados (ASICs) , quedando las GPUs para entrenamiento con ingentes cantidades de datos.

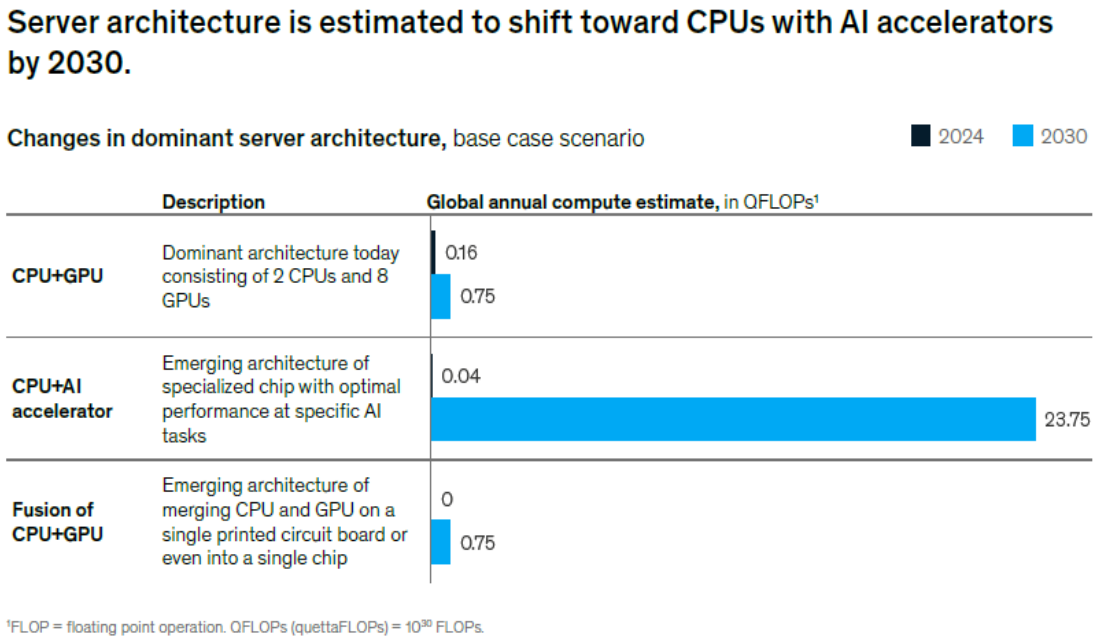


Figura 61. Arquitecturas predominantes de los servidores de IA en 2024 y la previsión en 2030. Se estima que la arquitectura dominante en 2030 será CPU+ Aceleradores IA. Fuente: (Burkacky, et al., March 2024)

En la Fig. 61, se aprecia que la tendencia que se estima prevalecerá hacia 2030 es la de "CPU+ acelerador IA", es decir, una arquitectura de chip especializada (ASIC) y FPGA (si se desea más flexibilidad, a costa de más consumo de energía) con rendimiento óptimo para tareas de IA específicas.

Tendencias futuras para chips IA para defensa y seguridad

Los avances en semiconductores impulsados por la IA están transformando los sistemas de seguridad de defensa y guerra modernos. Los chips de IA de próxima generación se centrarán en el análisis del campo de batalla en tiempo real, la resiliencia cibernética, los sistemas autónomos y la guerra electrónica.

Principales tendencias de chips de IA en defensa y seguridad:

1. **Inteligencia artificial de borde para la toma de decisiones en tiempo real**
   - Los chips de IA permitirán el procesamiento en tiempo real en el campo de batalla sin depender de la nube.

   Casos de uso:

   - Drones y vehículos aéreos no tripulados impulsados por IA
   - Sistemas de radar impulsados por IA
   - Dispositivos portátiles para soldados asistidos por IA

Ejemplo: NVIDIA Jetson Orin → IA de bajo consumo para inteligencia en el campo de batalla.

2. **IA para ciberseguridad y detección de amenazas**
   - Los chips de IA detectarán, evitarán y responderán a los ciberataques en tiempo real.

   Casos de uso:
   - Defensa de guerra cibernética impulsada por IA
   - Detección de intrusiones y cifrado de IA
   - Detección de anomalías basada en IA en redes militares

Ejemplo: Chip de ciberseguridad con IA NVIDIA Morpheus para detección de amenazas cibernéticas.

3. **Chips de IA neuromórficos para sistemas militares autónomos**
   - Los procesadores neuromórficos de próxima generación (IA inspirada en el cerebro) mejorarán la toma de decisiones, el reconocimiento de objetivos y la adaptación en tiempo real.

   Casos de uso:
   - Drones de combate autónomos impulsados por IA
   - Sistemas de defensa antimisiles asistidos por IA
   - Sensores de IA de vigilancia inteligente

Ejemplo: Chip de IA neuromórfico Intel Loihi para inteligencia autónoma de bajo consumo.

4. **Guerra electrónica (EW) y procesamiento de señales impulsados por IA**
   - Los chips de IA bloquearán las señales enemigas, detectarán aviones furtivos y mejorarán las contramedidas electrónicas (ECM).

   Casos de uso:
   - Procesamiento de señales de radar y sonar asistido por IA
   - Interferencias y suplantación de identidad electrónicas impulsadas por IA
   - Sistemas antidrones y antisatélites impulsados por IA

Ejemplo: FPGA AMD Versal AI Edge para inteligencia de señales impulsada por IA en tiempo real (SIGINT).

5. **Defensa contra misiles hipersónicos impulsada por IA**
   - Los chips de IA rastrearán e interceptarán armas hipersónicas que viajen a velocidades de Mach 5+.

   Casos de uso:
   - Sistemas de alerta temprana de misiles mejorados por IA
   - Simulaciones de interceptación en tiempo real impulsadas por IA
   - Predicción de trayectoria hipersónica impulsada por IA

Ejemplo: Sistemas de seguimiento de misiles impulsados por IA de Raytheon asistidos por IA.

6. **IA cuántica para criptografía militar y comunicaciones seguras**
   - La IA + la computación cuántica mejorará el cifrado militar y la transmisión segura de datos.

   Casos de uso:
   - Cifrado cuántico impulsado por IA para comunicaciones seguras en el campo de batalla IA cuántica en ciberseguridad para descifrar códigos criptográficos clásicos
   - Seguridad de satélites espía mejorada por IA

Ejemplo: Satélite chino habilitado con IA cuántica (Micius) para comunicaciones inhackeables

7. **Guerra de enjambre impulsada por IA (enjambres de drones y unidades de combate autónomas)**
   - Los chips de IA controlarán enjambres de drones autónomos para ataques coordinados.

   Casos de uso:
   - Enjambres de UAV impulsados por IA para ataques militares
   - Defensa antidrones basada en IA
   - Unidades robóticas de campo de batalla mejoradas por IA

   Ejemplo: Programa OFFSET de DARPA → Combate de enjambre de drones impulsado por IA.

En cuanto a las tendencias futuras sobre tecnologías de los chips IA para defensa, podemos considerar como más destacables las siguientes:

Tecnologías futuras de chips de IA para la defensa

- **Chips de IA apilados en 3D** → Chips de IA de alta velocidad y bajo consumo para el procesamiento de IA en el campo de batalla.
- **Procesadores de IA militares RISC-V** → Chips de IA de código abierto para aplicaciones de defensa seguras.
- **Chips de IA fotónicos** → Chips de IA ultrarrápidos que utilizan luz en lugar de electricidad para análisis de datos en tiempo real.
- **Seguridad de IA + Blockchain** → Blockchain impulsado por IA para proteger la IoT y las comunicaciones militares.
- **Detección neuromórfica**

Conclusión: chips de IA y el futuro de la defensa militar

Los chips de IA redefinirán la guerra moderna, haciendo que los sistemas de defensa sean más inteligentes, rápidos y autónomos. La ciberdefensa impulsada por IA, la guerra electrónica, el

seguimiento de misiles hipersónicos y la inteligencia autónoma en el campo de batalla se convertirán en estándar.

Detección neuromórfica

La detección neuromórfica es un enfoque avanzado impulsado por IA que imita las capacidades de procesamiento sensorial del cerebro, lo que permite la percepción en tiempo real, la toma de decisiones y la inteligencia adaptativa en entornos militares. Estos sensores utilizan el procesamiento basado en eventos, lo que **reduce el consumo de energía y aumenta la eficiencia en comparación con los sensores tradicionales**.

La información codificada en serie de impulsos es fundamental para el entrenamiento y el funcionamiento de las SNN (Redes Neuronales de Impulsos). Sin embargo, los datos sensoriales recopilados del entorno por varios sensores son normalmente de naturaleza analógica y, por lo tanto, no se pueden introducir directamente en la SNN para su procesamiento. En consecuencia, existe una necesidad crítica de dispositivos especializados capaces de convertir señales analógicas en picos. Un método convencional para abordar este desafío es el uso de **convertidores analógico-digitales (ADC)**. Los ADC funcionan muestreando la señal analógica a intervalos regulares y luego cuantificando cada muestra para producir un valor digital. Los atributos de inmunidad al ruido, facilidad de integración y flexibilidad han contribuido a su adopción generalizada en los sistemas electrónicos modernos. Además de los ADC, también se han empleado **osciladores en anillo** para traducir señales analógicas en trenes de picos.

Sin embargo, el consumo de energía tanto de los ADC como de los osciladores en anillo sigue siendo una preocupación importante, en particular en escenarios que requieren alta resolución y velocidad. Además, sus complejos diseños de circuitos dan como resultados tamaños físicos que pueden ser demasiado grandes en diseños compactos. En este sentido, el desarrollo de dispositivos que puedan convertir de manera eficiente y compacta las señales del dominio analógico al dominio de impulsos es de suma importancia.

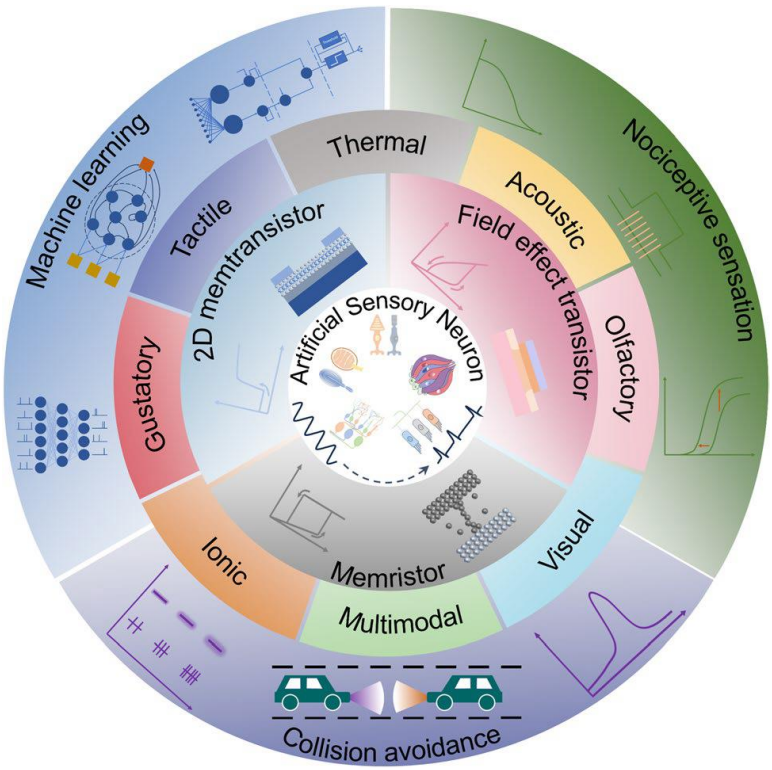


Figura 62. Descripción general de las ASN (Neuronas Sensoriales Artificiales) desde los fundamentos biológicos, los dispositivos hasta la implementación y las aplicaciones de sensaciones. Fuente: (Zhong, et al., 2025)

El sistema sensorial biológico procesa estímulos externos, como el tacto físico, la luz, el sonido y los productos químicos, en paralelo con un consumo de energía ultrabajo que es detectado por los receptores, se genera un pico eléctrico y se transmite al sistema nervioso central (SNC) para descodificarlo.

En los últimos años, inspirándose en el mecanismo operativo del sistema sensorial biológico, se han explorado ampliamente dispositivos neuromórficos emergentes basados en neuronas sensoriales artificiales (ASN) que pueden convertir de manera eficiente la información ambiental en impulsos eléctricos, con el objetivo de superar las limitaciones de las contrapartes convencionales basadas en CMOS. Los dispositivos neuromórficos emergentes generalmente ofrecen alta escalabilidad, bajo consumo de energía y alta velocidad, lo que puede reducir los costos de energía y hardware asociados con la transducción sensorial (Zhong, et al., 2025). Un conocimiento profundo del mecanismo de funcionamiento de los receptores biológicos puede ofrecer algunas perspectivas para el desarrollo futuro de las ASN.

Mediante una serie de dispositivos electrónicos como memristor, STLFET, 2D memtransistor se ha construido neuronas sensoriales artificiales (ASN) que simulan los sentidos humanos como gusto, oído, tacto, vista, olfato. Estas ASN son : Neuronas Tactiles Artificiales (ATTN), Neuronas Térmicas Artificiales (ATMN), Neuronas Acústicas Artificiales (AAN), Neuronas Gustativas Artificiales (AGN), Neuronas Olfativas Artificiales (AON) , Neuronas Bioquímicas Artificiales (ABCN) y Neuronas Visuales Artificiales (AVN)

Las AON pueden ser de gran utilidad para reconocer gases. Estos dispositivos pueden detectar compuestos orgánicos volátiles y otras sustancias químicas en el aire, y se aplican ampliamente en una variedad de aplicaciones, como la calidad de los alimentos, el monitoreo ambiental y el diagnóstico.

Para abordar este problema, Wang y sus colegas desarrollaron un sistema olfativo artificial que integra detección de gases, almacenamiento de datos y funciones de procesamiento (Fig. 63a) Las unidades de codificación del sistema se realizan conectando los sensores de gas comerciales a un memristor difusivo. Las sinapsis, basadas en dispositivos memristivos no volátiles, transmiten señales desde AON a neuronas de relevo según pesos sinápticos que se entrenan a través de plasticidad dependiente de la tasa de picos supervisada (SRDP). Las neuronas de relevo procesan las señales de las sinapsis y clasifican los gases. Una unidad de procesamiento, organizada con una matriz de puertas programables en campo (FPGA) y un circuito integrado de aplicación específica (ASIC) para el procesamiento y la generación de señales, recibe las salidas de las neuronas y distribuye las señales de control necesarias a los respectivos componentes. Este sistema olfativo puede identificar claramente cuatro muestras de gas (formaldehído, etanol, acetona y tolueno) con diferentes patrones de picos. Sin embargo, el sistema no codifica la concentración de gas, que es una dimensión esencial de la información sobre olores. (Ver Fig. (a)) Han et al. informaron sobre un AON que puede convertir tanto el tipo de gas como la concentración en picos, lo que amplía significativamente su aplicabilidad en áreas como la calidad del aire y el monitoreo toxicológico (Fig. 63b). El AON contiene un sensor de óxido metálico semiconductor (SMO) y un STLFET.

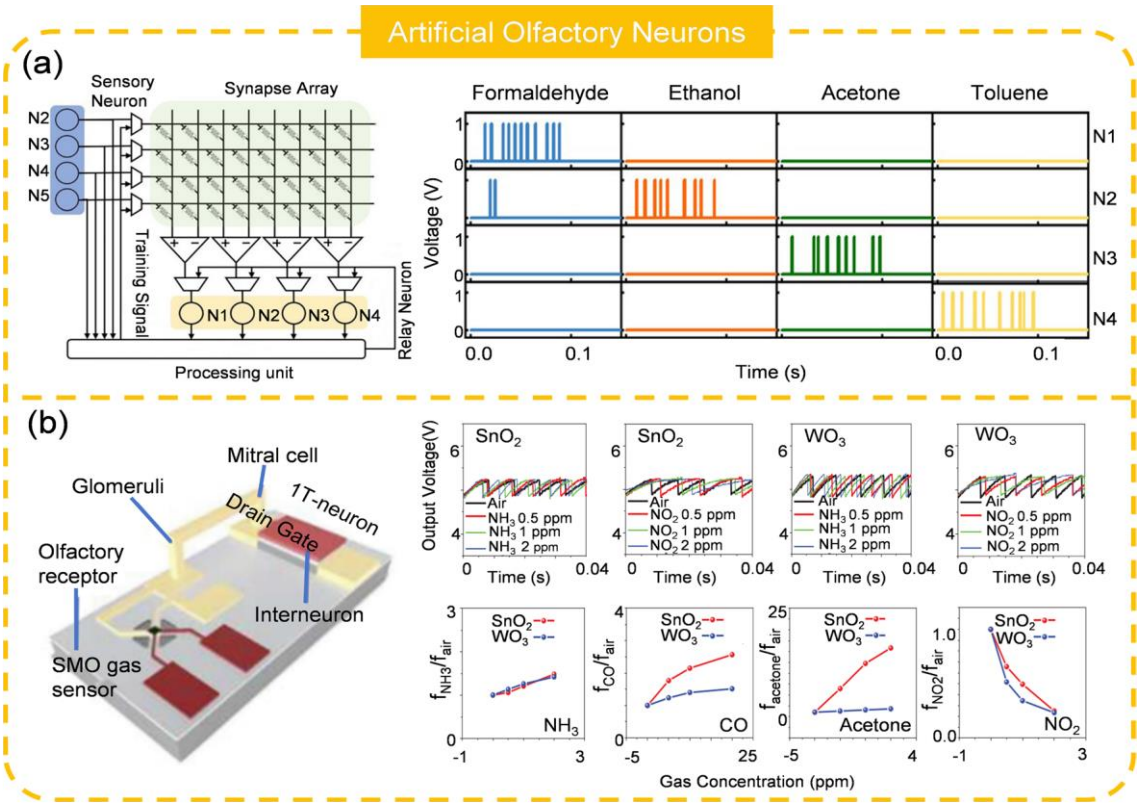


Figura 63. Neuronas olfativas Artificiales. (a) y (b). Fuente: (Zhong, et al., 2025)

Las ASN al consumir menos energía que los ADCs y además poder entregar una serie de impulsos a las SNN para la inferencia **son ideales para aplicaciones en tiempo real en el borde.** Sin embargo, para aprovechar al máximo los indudables beneficios de las ASN, todavía necesitan pulirse muchos aspectos y deben examinarse cuidadosamente a varios niveles.

Ventajas clave de la detección neuromórfica en aplicaciones militares:

- **Consumo de energía ultrabajo** → La detección basada en eventos minimiza el procesamiento redundante de datos.
- **Toma de decisiones a alta velocidad** → Tiempos de respuesta casi instantáneos para la detección de amenazas.
- **Resiliencia en entornos hostiles** → Funciona en condiciones extremas de iluminación, ruido o guerra electrónica.
- **Aprendizaje y adaptación en tiempo real** → Aprendizaje continuo para amenazas en evolución.

Tendencias futuras en detección neuromórfica para uso militar:

- **Sensores bioinspirados (ASN)** → Imitando la visión de los insectos y la audición de los animales para una detección avanzada.
- **IA cuántica neuromórfica** → IA cuántica + neuromórfica híbrida para una criptografía superior.
- **Interfaz cerebro-computadora (BCI) para soldados** → Comunicación directa mente-máquina impulsada por IA.

Fotónica Integrada

Es bien sabido, como se ha expuesto repetidamente en este informe, que uno de los problemas en los centros de datos de IA es el excesivo consumo de energía. Uno de los factores que influyen, además del movimiento de datos procesador/memorias son las **interconexiones electrónicas entre chips y equipos**. La fotónica que utiliza fotones (luz) en lugar de electrones para la transmisión de la información, puede ser parte de la solución.

En efecto, la fotónica, que aprovecha el poder de la luz para la comunicación y la computación, presenta una alternativa prometedora para superar las limitaciones que plantean las interconexiones electrónicas. A diferencia de los electrones, los fotones proporcionan diversas dimensiones para controlar, como la longitud de onda, la polarización y el modo espacial. Estas ventajas de las propiedades fotónicas permiten un procesamiento de datos más rápido, un **menor consumo de energía** y un **paralelismo inherente** en todos los niveles de los circuitos fotónicos integrados, lo que la posiciona como una solución transformadora para la IA. Los modelos informáticos no digitales, en particular los que se posibilitan mediante la fotónica, prometen soluciones a desafíos como la baja latencia, el alto ancho de banda y las bajas energías, lo que da lugar a un campo de **fotónica neuromórfica** en la intersección de la fotónica y el marco de la ingeniería neuronal. (Gupta & Jolly, 2024)

Las **redes fotónicas neuronales** (NPN) son arquitecturas de IA que utilizan luz (**fotones) en lugar de electrones** para realizar cálculos de redes neuronales. Aprovechan componentes ópticos como láseres, guías de ondas y circuitos fotónicos para acelerar las tareas de aprendizaje profundo con una velocidad y una eficiencia energética ultra altas.

Las principales ventajas de las redes fotónicas neuronales

1. **Velocidad y paralelismo**
   - Los fotones viajan más rápido que los electrones, lo que permite un procesamiento de IA casi instantáneo.
   - Computación masivamente paralela → Las interconexiones ópticas procesan múltiples flujos de datos a la vez.

   Caso de uso: análisis de amenazas en tiempo real impulsado por IA y toma de decisiones militares.

2. **Consumo de energía ultra bajo**
   - Los chips fotónicos consumen significativamente menos energía en comparación con los chips de IA electrónicos.
   - Sin pérdidas resistivas → Reduce la disipación de calor en centros de datos de IA y dispositivos de IA de borde.

   Caso de uso: cifrado de ciberseguridad impulsado por IA con costos de energía ultrabajos.

3. **Sin interferencias electromagnéticas (EMI)**
   - Inmune a interferencias de radiofrecuencia (RF) y guerra electrónica (EW).
   - Más adecuado para aplicaciones militares y aeroespaciales donde la EMI es un desafío importante.

   Caso de uso: Comunicaciones militares seguras asistidas por IA y procesamiento de IA por satélite.

4. **Interconexiones ópticas escalables**
    - Sin necesidad de interconexiones eléctricas de gran ancho de banda (cuello de botella en chips electrónicos).
    - Admite computación en la nube de IA de alta densidad y larga distancia e inferencia de IA de borde.

Caso de uso: Seguimiento de misiles hipersónicos impulsado por IA e inteligencia de IA basada en el espacio.

Desventajas de las redes neuronales fotónicas

1. **Desafíos de fabricación**
    - Los circuitos fotónicos son difíciles de fabricar a escala.
    - Falta de un proceso de fabricación fotónica estandarizado.
    - La fotónica de silicio está avanzando, pero aún está por detrás de la producción de chips CMOS.
2. **Limitaciones de memoria y almacenamiento**
    - Los procesadores fotónicos de IA tienen dificultades para integrar elementos de memoria.
    - Los chips electrónicos almacenan y recuperan datos de manera más eficiente (por ejemplo, DRAM, SRAM).
3. **Se requieren soluciones híbridas fotónicas-electrónicas para sistemas de IA prácticos.**
    - Complejidad de programación
    - Falta de marcos de software de IA optimizados para redes neuronales fotónicas.
    - La mayoría de los modelos de IA están diseñados para hardware electrónico (chips de IA de NVIDIA, AMD e Intel).

El futuro de las redes fotónicas neuronales

- Los aceleradores de IA fotónicos-electrónicos híbridos combinarán las ventajas de la óptica ultrarrápida y la memoria electrónica eficiente.
- Se espera que la IA fotónica revolucione la inteligencia militar, la ciberseguridad y la computación de alto rendimiento (HPC).

La **fotónica de silicio** es una tecnología de vanguardia que integra circuitos fotónicos (basados en luz) con circuitos electrónicos de silicio tradicionales para mejorar el procesamiento de IA neuromórfica. Al combinar la velocidad y la eficiencia energética de los fotones con la versatilidad computacional de los chips de silicio, este enfoque híbrido puede revolucionar las aplicaciones de IA en defensa, ciberseguridad e inteligencia en tiempo real.

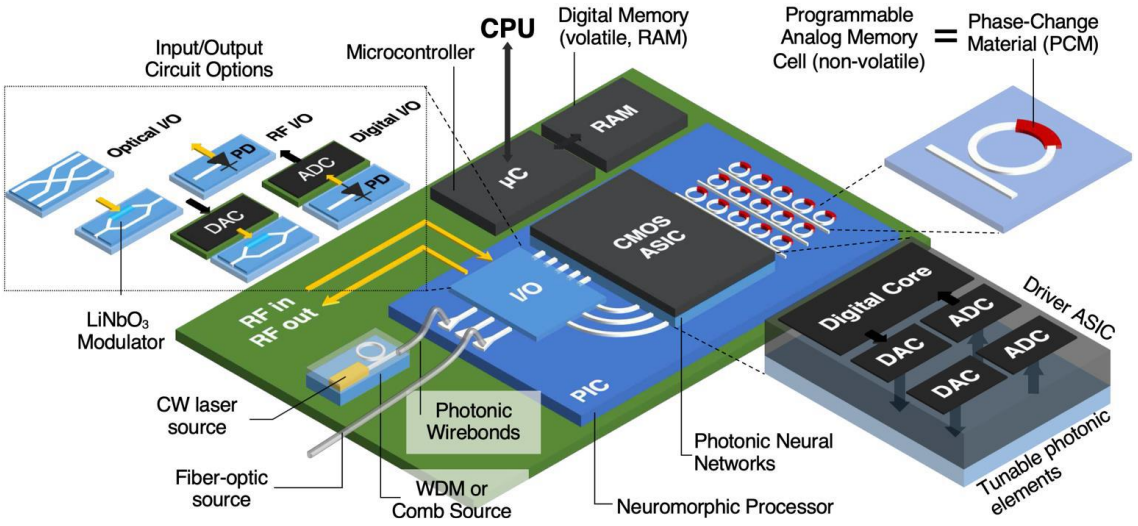


Figura 64. Arquitectura de procesador fotónico neuromórfico. Fuente: (Shastri, et al., February 2021)

La red neuronal fotónica con elementos ópticos configurables reside en una matriz de circuito integrado fotónico (PIC) de silicio. Algunos elementos se pueden configurar de forma analógica y no volátil mediante materiales de cambio de fase (PCM). Las líneas blancas representan el enrutamiento de la guía de ondas.

Un desafío clave es hacer llegar la energía óptica a la matriz de silicio. La energía óptica puede ser proporcionada por una matriz ópticamente activa, es decir, capaz de generar luz, o, alternativamente, una interfaz de fibra externa. Para la E/S de señal (recuadro superior izquierdo), las conversiones eléctricas a ópticas y viceversa se realizan mediante moduladores fotónicos de silicio (estructuras hexagonales) y fotodetectores (triángulos con cruces). Esto significa que todas las E/S a nivel de paquete pueden ser eléctricas u ópticas, digitales o analógicas, según la aplicación del usuario.

El otro desafío clave es **controlar la red neuronal fotónica**. Los recuadros negros representan matrices CMOS que satisfacen varias funciones de control, interfaz y programación. La interfaz de programación digital consta de un microcontrolador (µC) con memoria digital (RAM) coubicada, ambos componentes estándar. (Shastri, et al., February 2021)

Microelectrónica para computación cuántica.

Para mejorar el rendimiento de la IA, se está trabajando en aplicar los conocimientos que ha proporcionado la computación cuántica a resolver los desafíos de la IA. Esto implica un cambio importante de paradigma. En lugar de redes neuronales que funcionen digitalmente (operaciones GEMM) se trabajará con redes neuronales cuánticas (QNN).

Las redes neuronales cuánticas (QNN) son una fusión de computación cuántica y redes neuronales que tienen como objetivo aprovechar el poder de la mecánica cuántica para mejorar el rendimiento de la IA. A diferencia de las redes neuronales clásicas, las QNN utilizan cúbits y operaciones cuánticas para procesar y aprender de los datos de manera más eficiente.

¿Qué es una red neuronal cuántica (QNN)? Una red neuronal cuántica (QNN) es un tipo de modelo de aprendizaje automático que aplica principios de computación cuántica para mejorar la eficiencia y el rendimiento de las redes neuronales. Utiliza **superposición cuántica, entrelazamiento e interferencia** para procesar y aprender de los datos de formas que las redes neuronales clásicas no pueden.

- Redes neuronales clásicas (CNN, RNN, etc.): utilizan bits tradicionales (0 y 1).
- Redes neuronales cuánticas (QNN): utilizan cúbits, que pueden ser 0, 1 o una superposición de ambos.

En lugar de las multiplicaciones de matrices tradicionales como en el aprendizaje profundo clásico, las QNN utilizan circuitos y puertas cuánticas para procesar la información de manera más eficiente.

¿Cómo funcionan las redes neuronales cuánticas? Las QNN utilizan circuitos cuánticos en lugar de las capas tradicionales de neuronas. Así es como funcionan:

1. **Codificación de datos cuánticos** (mapas de características cuánticas) Los datos clásicos (por ejemplo, imágenes, texto) primero deben codificarse en un estado cuántico. Esto se hace utilizando mapas de características cuánticas, que transforman los datos de entrada en estados cuánticos que pueden procesarse mediante puertas cuánticas.

2. **Procesamiento cuántico con cúbits** Las QNN utilizan la superposición y el entrelazamiento cuánticos para crear una representación rica y de alta dimensión de los datos. Las puertas cuánticas (por ejemplo, Hadamard, CNOT, Pauli-X/Y/Z) manipulan los estados cuánticos para extraer patrones de los datos.

3. **Medición y resultados cuánticos**. Después del procesamiento cuántico, se mide el sistema cuántico para extraer resultados clásicos. El resultado se utiliza luego para actualizar los parámetros de la red neuronal cuántica, de forma similar a la retropropagación en el aprendizaje profundo clásico.

Ventajas de las redes neuronales cuánticas (QNN) sobre la IA clásica

- **Aceleración en el entrenamiento y la inferencia**
  - Los ordenadores cuánticos pueden realizar cálculos exponencialmente más rápidos que las GPU clásicas para ciertas tareas (por ejemplo, resolver operaciones con matrices grandes).
  - Aceleración cuántica: las QNN podrían acelerar tareas como GEMM (multiplicación general de matrices), que es un cuello de botella en el aprendizaje profundo.
- **Uso más eficiente de la memoria**
  - Los modelos clásicos de aprendizaje profundo requieren una gran cantidad de memoria para pesos y activaciones. Las QNN almacenan y procesan información en estados cuánticos, lo que permite una representación compacta de grandes conjuntos de datos.
- **Solución de problemas complejos**
  - Se espera que las QNN se destaquen en problemas altamente complejos como:
    - Plegamiento de proteínas (biofísica)
    - Descubrimiento de fármacos (farmacéuticos)
    - Problemas de optimización (logística, finanzas, criptografía)
    - Ciberseguridad y criptografía cuántica
- **Aprendizaje a partir de conjuntos de datos pequeños (ventaja cuántica**)
  - El aprendizaje profundo clásico requiere conjuntos de datos etiquetados masivos. Las QNN explotan las correlaciones cuánticas para aprender patrones con menos muestras de entrenamiento.

Hoy en día hay circuitos microelectrónicos que emulan el comportamiento de los cúbits

La computación cuántica requiere un conjunto altamente especializado de tecnologías microelectrónicas para respaldar el control, la lectura, la corrección de errores y la integración del sistema de cúbits. A diferencia de los ordenadores clásicos, que dependen de transistores y puertas lógicas, los procesadores cuánticos utilizan cúbits que requieren microelectrónica criogénica, superconductora, fotónica o de iones atrapados para su funcionamiento.

Microelectrónica clave para la computación cuántica:

- Electrónica superconductora (uniones Josephson y control de cúbits)
- Espintrónica y cúbits basados en semiconductores
- Microelectrónica fotónica para computación cuántica óptica (puede trabajar a temperatura ambiente, no es necesaria la criogenia).
- Microelectrónica de trampa de iones

Tendencias futuras en microelectrónica cuántica

- Chips cuánticos compatibles con CMOS → Integración híbrida de procesadores clásicos y cuánticos.
- Inteligencia artificial neuromórfica y cuántica → Combinación de computación inspirada en el cerebro con mecánica cuántica.
- Sistemas híbridos fotónicos y superconductores → Fusión de cúbits superconductores y basados en luz para un procesamiento cuántico ultraeficiente.

Recientemente Microsoft ha presentado el circuito cuántico Majorana-1 (Ver Fig.65) Interferometric single-shot parity measurement in InAs–Al hybrid devices | Nature,.

Microsoft ha afirmado que "ha construido el primer topoconductor del mundo. Esta revolucionaria clase de materiales permite crear *la superconductividad topológica,* un nuevo estado de la materia que antes solo existía en teoría. El avance se deriva de las innovaciones de Microsoft en el diseño y **fabricación de dispositivos** definidos por puerta que combinan **arseniuro de indio** (un semiconductor) y aluminio (un superconductor). Cuando se enfrían hasta casi el cero absoluto y se sintonizan con campos magnéticos, estos dispositivos forman nanocables superconductores topológicos con modos cero de Majorana (MZM, por sus siglas en inglés) en los extremos de los cables". Microsoft presenta Majorana 1, el primer procesador cuántico del mundo impulsado por qubits topológicos - Source LATAM

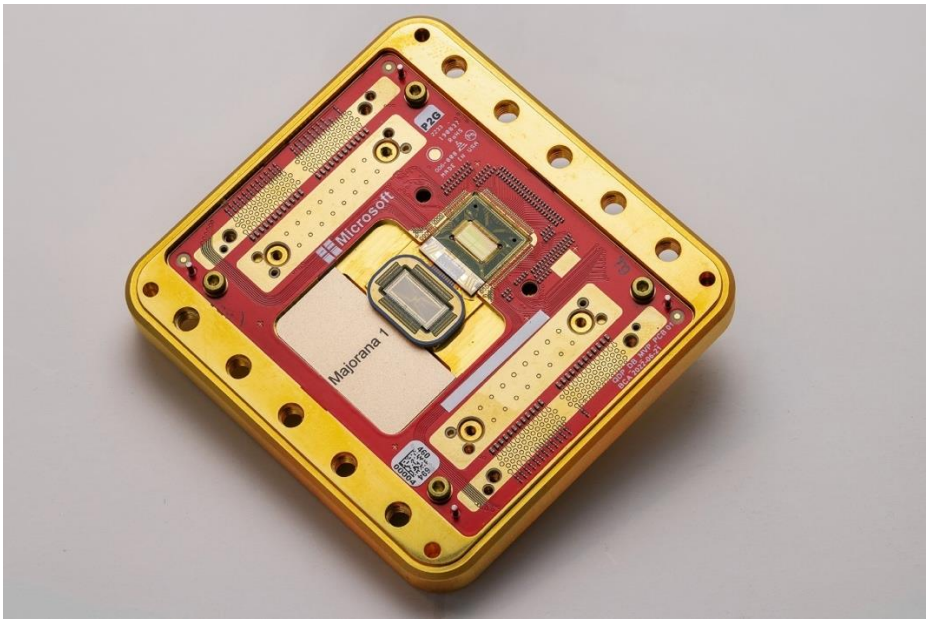


Figura 65. Circuito cuántico Majorana-1 de Microsoft. Fuente: Interferometric single-shot parity measurement in InAs–Al hybrid devices | Nature,.

Ejemplos de uso avanzado de chips para IA en Defensa

Chips para cifrado Inteligente

Las operaciones militares modernas dependen de la **ciberseguridad** para proteger los **activos digitales y físicos esenciales** como las redes militares, la infraestructura y las personas (civiles y militares). La ciberseguridad protege la **estrategia militar crítica** y **clasificada** para las operaciones actuales, pasadas y futuras. Asegura la infraestructura nacional vital, como la energía y el agua, y protege contra los ataques que podrían apoderarse de los drones militares críticos o los sistemas de comunicación.

La **guerra cibernética** es cuando un tercero manipula, roba o accede a datos y sistemas con la intención maliciosa de interrumpir, destruir o negar su uso.

En la ciberguerra militar se conocen siete tipos específicos de ciberataques Exploring the vital role of cybersecurity in the military: applications, technologies, & training - Lighthouse Labs :

- **Espionaje:** robo de secretos de otros países (generalmente mediante botnets o phishing esférico para comprometer los sistemas de información)
- **Sabotaje:** robo o aprovechamiento de información o uso de amenazas internas a través de empleados insatisfechos o empleados con afiliaciones externas
- **Ataques de denegación de servicio (DoS)**: inundar sitios web y sistemas con solicitudes falsas para evitar que los usuarios legítimos accedan al sistema
- **Red eléctrica:** ataque mediante la desactivación de redes eléctricas para interrumpir la infraestructura o detener las comunicaciones

- **Ataques de propaganda:** difusión de verdades o mentiras embarazosas para controlar las mentes y las acciones de otros
- **Perturbación económica:** ataques cibernéticos para interrumpir los mercados de valores, los sistemas de pago o los bancos
- **Ataques sorpresa:** ataques estratégicos a gran escala diseñados para debilitar al enemigo, a veces como precursores de un ataque físico terrestre.

Actualmente existen muchos nuevos desafíos para la ciberseguridad como:

- Las **amenazas persistentes avanzadas** (ATP) son ataques cibernéticos diseñados para extraer datos críticos de un sistema militar o gubernamental durante semanas, meses o años. A menudo pasan desapercibidos y pueden expandir su penetración en una red con el tiempo.
- **Amenazas internas**. Se utilizan ataques de sabotaje e ingeniería social para hacer que la gente esté "dentro" de las organizaciones gubernamentales y militares. Las naciones en guerra y los piratas informáticos se están aprovechando de los empleados gubernamentales o militares a través del chantaje o ataques de ingeniería social para acceder a los sistemas y datos desde el interior.

Al utilizar las capacidades de la IA, las organizaciones militares la están utilizando para:

- **Detectar amenazas y respuestas:** la IA analiza de manera eficiente grandes cantidades de datos para detectar amenazas cibernéticas en tiempo real, una tarea prácticamente imposible para el análisis manual realizado por humanos.
- **Prevención de ataques cibernéticos:** la IA monitorea de manera proactiva los sistemas militares y de infraestructura. Busca vulnerabilidades para que su equipo de ciberseguridad pueda identificar y mitigar los riesgos de manera más rápida y precisa.
- **Crear estrategia y capacitación**: la IA puede crear escenarios de amenazas cibernéticas para que los profesionales mantengan sus habilidades actualizadas y aprendan sobre nuevas amenazas. El aprendizaje automático es la forma en que la IA utiliza modelos matemáticos y datos para aprender sin recibir instrucciones directas. Permite que los sistemas informáticos “aprendan” por sí solos.

Las organizaciones militares pueden utilizar el aprendizaje automático para detectar de forma independiente patrones y desviaciones de la actividad habitual de la red (por ejemplo, un ataque). El aprendizaje automático identifica de forma eficiente anomalías en grandes cantidades de datos y presenta estos informes urgentes en tiempo real.

Los **chips impulsados por IA** desempeñan un papel fundamental en la ciberseguridad defensiva y ofensiva para aplicaciones militares. Estos chips permiten la detección de amenazas en tiempo real, el cifrado, la guerra cibernética y la respuesta automatizada en entornos de alta seguridad.

Aplicaciones clave de los chips de IA en la ciberseguridad militar

- **Inteligencia de amenazas y detección de anomalías** → Identifica ciberataques en tiempo real utilizando modelos de IA.
- **Sistemas de ciberdefensa autónomos** → Detección y respuesta a intrusiones impulsadas por IA (IDPS).
- **Cifrado resistente a los datos cuánticos** → Protege los datos clasificados mediante criptografía mejorada con IA.

- **Guerra cibernética impulsada por IA** → Capacidades cibernéticas ofensivas como defensa y piratería mejoradas con IA.
- **Redes de comunicación seguras** → Protege los sistemas de mando y control (C2) militares.

Algunos Gobiernos han establecido programas sobre ciberseguridad en la que se aplica la IA con chips especializados. Algunos de los programas más notables son:

Programas de ciberseguridad de chips de IA militares notables:

- **DARPA HALO** (aprendizaje jerárquico con adaptación en chip):
  Chip de IA neuromórfico para detección de amenazas cibernéticas en tiempo real con consumo ultrabajo. Se utiliza en sistemas de defensa autónomos seguros.
- **Iniciativa de criptografía post-cuántica de la NSA e IBM**
  Utiliza procesadores de IA + cuánticos para cifrado de grado militar. Diseñado para resistir futuras amenazas cibernéticas cuánticas.
- **Proyecto MAVEN + NVIDIA AI para ciberdefensa**
  Inteligencia cibernética impulsada por IA para el análisis de amenazas en redes militares. Utiliza GPU NVIDIA AI para la monitorización de la ciberseguridad basada en aprendizaje profundo.
- **Sistemas militares mejorados con IA de Intel**
  Se utilizan procesadores Intel TDT + Xeon para la detección de malware basada en IA en defensa. Protege las comunicaciones clasificadas y los puntos finales militares.

Respecto a los **chips de IA para ciberseguridad** podemos destacar:

- **NVIDIA Jetson & Morpheus (AI for Military Cyber Defense)**
  - Para aplicaciones de ciberseguridad, especialmente para el entorno de la defensa, NVIDIA propone una solución que utiliza sus GPUs. Esta solución es conocida como Morpheus-IA
  - **NVIDIA Morpheus: marco de ciberseguridad impulsado por IA**
    - NVIDIA Morpheus es un marco de ciberseguridad impulsado por IA que utiliza aprendizaje profundo y **detección de anomalías en tiempo real para proteger las redes de las amenazas cibernéticas**. Aprovecha las GPU NVIDIA para acelerar la detección de amenazas, la privacidad de los datos y las operaciones de seguridad impulsadas por IA (SecOps)

  - **Características clave de NVIDIA Morpheus**
    - **Detección de amenazas impulsada por IA →** Utiliza el aprendizaje profundo para detectar ataques de día cero, malware y amenazas internas.
    - **Seguridad de red en tiempo real →** Monitorea y analiza el tráfico de red a alta velocidad.
    - **Inspección profunda de paquetes (DPI) →** Analiza los paquetes en busca de anomalías sin descifrarlos.
    - **Análisis de registros y telemetría →** Procesa registros de sistemas de seguridad y detecta comportamiento sospechoso**.**
    - **Inteligencia artificial que preserva la privacidad →** Utiliza computación confidencial para analizar datos cifrados de forma segura.

- **Ciberseguridad con inteligencia artificial escalable →** Se ejecuta en GPU NVIDIA AI (A100, H100) para realizar análisis de seguridad de alto rendimiento.

Para equipos de ciberseguridad se han utilizado los productos de NVIDIA Jetson Xavier y Orin nvidia-jetson-agx-orin-technical-brief.pdf

Característica	NVIDIA Jetson & Morpheus (IA para ciber defensa militar)	Intel Xeon + TDT (Detección de ciber amenazas mejorada por IA)	AMD EPYC + AI Security (IA cifrada para defensa)	DARPA HALO (IA neuromórfica para seguridad militar)	IBM AI + Quantum Cybersecurity (Cifrado a prueba de futuro)
Caso de uso principal	Detección y defensa contra amenazas cibernéticas impulsadas por IA en redes militares	Detección de amenazas basada en hardware IA en sistemas militares seguros	Cifrado y procesamiento de datos seguro-basados en IA	Procesadores neuromórficos basados en IA para ciberdefensa de consumo ultrabajo	Cifrado cuántico seguro basado en IA e infraestructura de nube segura
Tipo de Hardware	GPUs IA (Jetson Xavier, Orin)	CPUs x86 Aceleración IA (Xeon)	Servidores CPUs Seguro, Optimizado IA	Chips IA Neuromórficos	Chips Híbridos de IA Cuántica.
Fortalezas Clave	Detección de ciberataques en tiempo real, inspección profunda de paquetes, defensa de red	Protege los sistemas seguros contra malware, APT y ciber espionaje	Cifrado y virtualización impulsados por IA para infraestructura de nube de defensa	IA inspirada en el cerebro para detección de anomalías en tiempo real y con consumo ultrabajo	Cifrado poscuántico basado en IA para ciberseguridad militar a prueba de futuro
Beneficios de Ciberseguridad	Previene ataques de guerra cibernética y amenazas de día cero	Detección a nivel de hardware de amenazas cibernéticas de clase militar	Computación de IA segura y procesamiento de datos cifrados	Monitoreo de red en tiempo real basado en IA con un consumo mínimo de energía	Protege contra amenazas cibernéticas cuánticas de próxima generación
Consumo de energía	**Moderado** (Optimizado para análisis de seguridad impulsados por IA)	**Bajo** (Integrado en CPUs seguras)	**Moderado** (Eficiencia de seguridad de IA de nivel de servidor)	**Consumo ultra bajo** (Eficiencia neuromórfica)	**Alto** (Cifrado híbrido de inteligencia artificial y cuántica)
Ideal para	Agencias de inteligencia militar, redes de defensa, comunicación segura	Ciberseguridad táctica con IA para entornos militares seguros	Sistemas de defensa cifrados basados en IA y protección de datos clasificados	Operaciones cibernéticas mejoradas con IA que requieren un consumo mínimo de energía	Ciberseguridad poscuántica para sistemas criptográficos militares

Chips para IoT militar

El término ***IoT militar*** (*MIoT*) se refiere a sensores, dispositivos y sistemas interconectados que funcionan con inteligencia artificial y que se utilizan en aplicaciones de defensa para vigilancia, logística, sistemas autónomos y ciberseguridad. Los chips de inteligencia artificial desempeñan un papel crucial al proporcionar procesamiento, toma de decisiones y seguridad de datos **en tiempo real** en entornos de combate y defensa.

Entre las principales características que poseen los chips y dispositivos que constituyen el IoT militar, podemos destacar las siguientes:

- **Procesamiento de IA de borde** → Procesamiento en el dispositivo para la toma de decisiones en tiempo real sin depender de la computación en la nube.
- **Bajo consumo de energía** → Chips de IA optimizados para la eficiencia energética en dispositivos IoT militares remotos y operados por batería, generalmente de poca capacidad, ya que tiene que ser ligeras de peso y tamaño.
- **Seguridad y cifrado** → Soluciones criptográficas impulsadas por IA para proteger las redes militares.
- **IA para fusión de sensores** → Integra datos de múltiples sensores (radar, cámaras, LiDAR, etc.) para el conocimiento del campo de batalla.
- **Operaciones autónomas** → Admite drones, vehículos robóticos y sistemas no tripulados con navegación basada en IA.

Un ejemplo de aplicación de IA IoT militar lo encontramos en las **Redes de campo de batalla inteligentes.** En efecto, la IA mejora las comunicaciones seguras de IoT militares (redes de sensores a soldados y a comandantes).

La inteligencia artificial se está integrando cada vez más en los sistemas de comunicación para mejorar su eficiencia, seguridad y adaptabilidad. La IA puede gestionar el tráfico de la red, detectar y responder a las amenazas cibernéticas en tiempo real y optimizar las vías de comunicación para lograr velocidad y confiabilidad. Algunas de las funciones que se logran al añadir IA a los dispositivos IoT militares son:

- **Gestión de la red:** la IA puede gestionar y optimizar de forma autónoma las redes de comunicación, lo que garantiza que el ancho de banda se asigne donde más se necesita.
- **Ciberseguridad:** los sistemas impulsados por IA pueden detectar y mitigar las amenazas cibernéticas más rápido que los operadores humanos, lo que proporciona una capa adicional de defensa.
- **Comunicación adaptativa**: la IA podría permitir que los sistemas de comunicación se ajusten dinámicamente a las condiciones cambiantes del campo de batalla, lo que garantiza una comunicación ininterrumpida en entornos complejos.

Chips para guiado de misiles hipersónicos

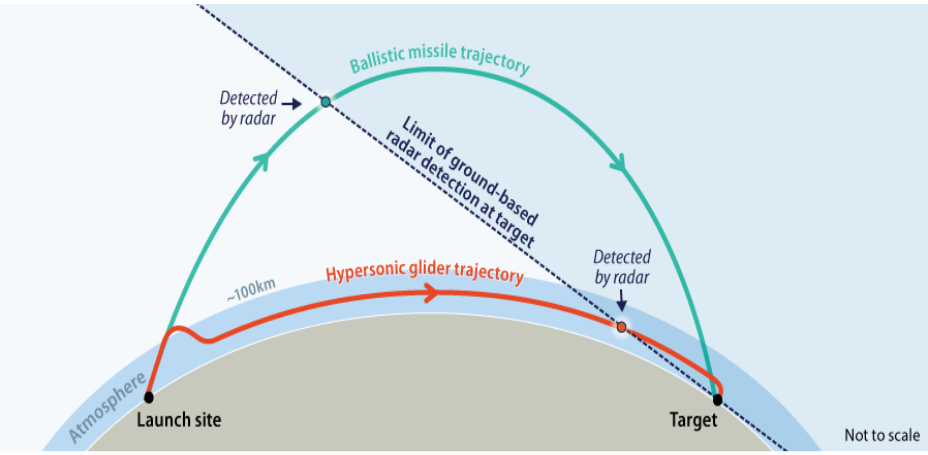


Figura 66. Trayectoria de un misil hipersónico comparada con la de un misil balístico. Fuente: Hypersonic Weapons: Background and Issues for Congress, February 11, 25

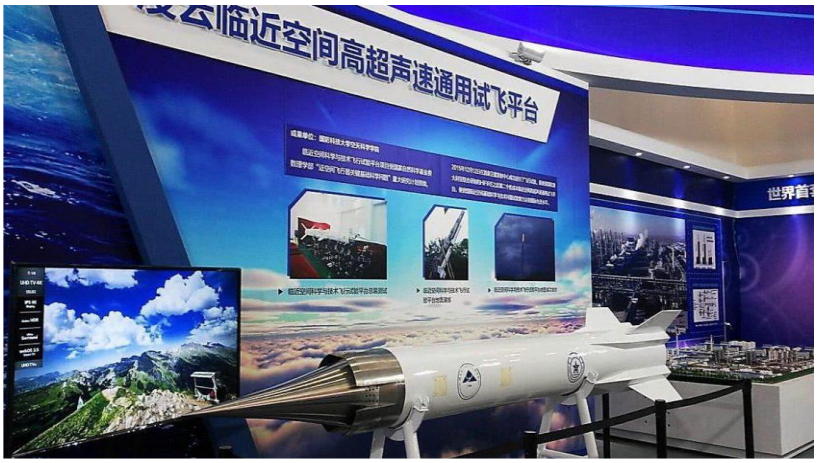


Figura 67. Misil hipersónico chino GDF-600.Fuente: Hypersonic Weapons: Background and Issues for Congress, February 11, 25 Photo accompanying Drake Long, "China Reveals Lingyun-1 Hypersonic Missile at National Science and Technology Expo," *The Defense Post*, May 21, 2018

La inteligencia artificial (IA) está transformando los sistemas de guía de misiles hipersónicos al proporcionar capacidades avanzadas para el procesamiento de datos, la toma de decisiones y el ataque a objetivos.

La Navegación y selección de objetivos impulsadas por IA: Los algoritmos de IA analizan los datos entrantes de varios sensores, lo que permite que los misiles adapten sus rutas de vuelo en tiempo real. Esta capacidad mejora la capacidad de respuesta del misil a las condiciones dinámicas del campo de batalla. Al emplear el aprendizaje automático, estos sistemas pueden mejorar el reconocimiento y la clasificación de objetivos, lo que permite a los misiles diferenciar entre objetivos legítimos y señuelos u otras señales falsas. La IA permite que los misiles ejecuten maniobras evasivas basadas en rutas previstas de interceptores enemigos. Al anticipar las amenazas, el misil puede ajustar su trayectoria para evitar la interceptación.

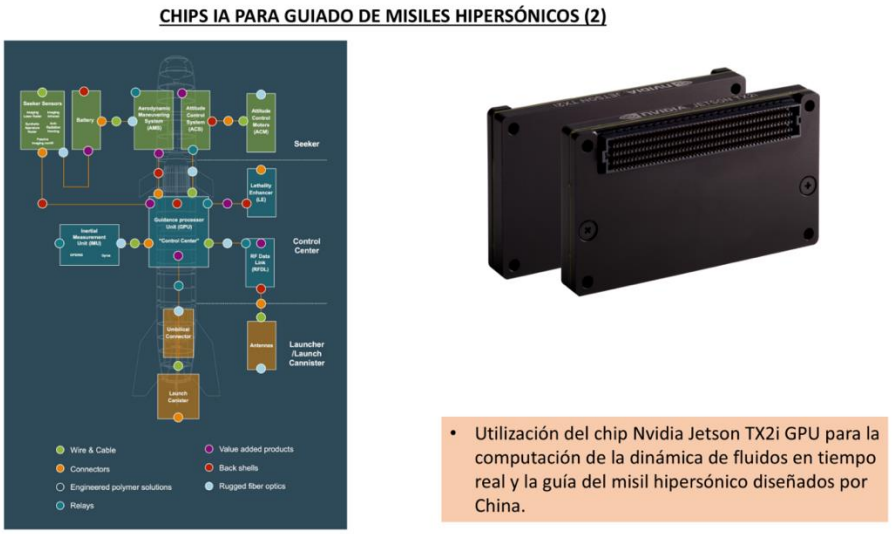


Figura 68. El chip Jetson TX2i GPU de Nvidia se utiliza en el guiado de misiles hipersónicos. Fuente: Chinese researchers use low-cost Nvidia chip for hypersonic weapon —unrestricted Nvidia Jetson TX2i powers guidance system | Tom's Hardware

Chips para enjambre de drones

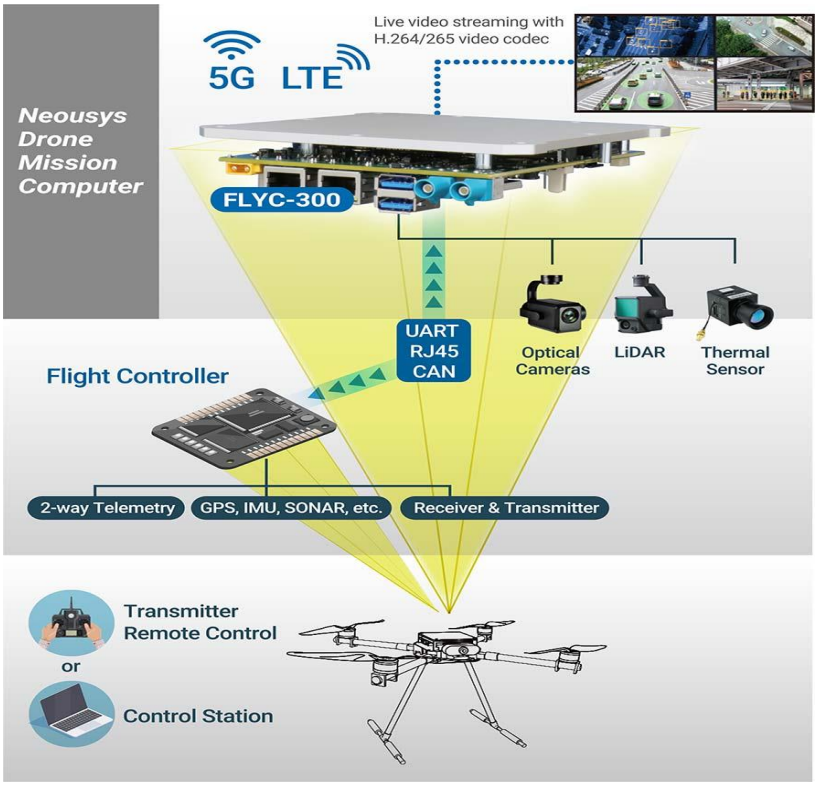


Figura 69. Esquema de FLYC-300 de Neousys. Chip IA en el borde a bordeo de drones. Fuente: Low-SWaP AI Mission Computer | NVIDIA Orin NX | FLYC-300 - Neousys Technology

La implementación de una computadora de IA en un dron presenta varios desafíos:
- **Entorno Operativo:** Condiciones de humedad, temperatura, presión, fuerza G ascendente/descendente varían a media que ascienden los drones puede afectar capacidades operativas.
- **Duración batería**: Es un tema crítico que exige diseño muy eficiente en consumo energía.
- **Seguridad y confiabilidad:** Comunicaciones robustas, integración de la IA con otros elementos (sensores,etc.) es compleja.

Figura 70. sistema FLYC-300 de Neousys. Fuente: Low-SWaP AI Mission Computer | NVIDIA Orin NX | FLYC-300 - Neousys Technology

Computadora ligera para misiones con drones con tecnología NVIDIA® Orin™ NX.

124mm x 123mm x 30.5mm
- Pesa solo 297 g para instalación a bordo
- Hasta 100 TOPS GPU de Jetson Orin NX
- Admite múltiples interfaces de cámara y sensor
· 2x GbE y 2x USB3 para cámaras RGB/infrarrojas/hiperespectrales y lidar/radar
· 2x GMSL2 para cámaras HDR/3D
- UART y CAN integrados para interactuar con el controlador de vuelo
- 1x M.2 2230 para almacenamiento y comunicación 4G/5G disponible
- Admite paquete de baterías para drones 4S-14S

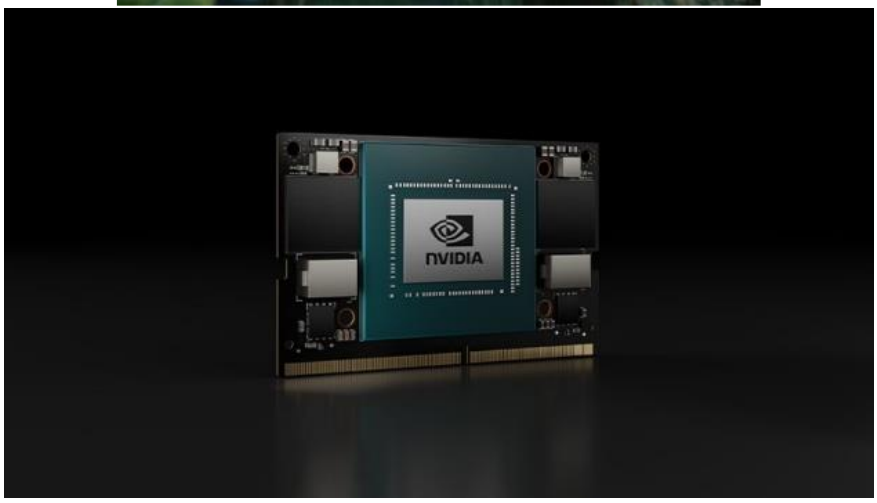


Figura 71 y Figura 72. Dron con FLY-300 y Chip IA Jetson Orin NX de Nvidia. Fuente: Neousys y Nvidia

Los sistemas de IA miniaturizados pueden mejorar significativamente la eficiencia de los drones al integrar algoritmos avanzados y capacidades de mapeo 3D. Al analizar datos en tiempo real e información topográfica preexistente, la IA puede optimizar las rutas de vuelo, minimizando los desvíos y apuntando directamente a las áreas de interés.

Esta navegación inteligente permite que los drones se adapten a entornos dinámicos, evitando obstáculos y ajustando rutas sobre la marcha.

NVIDIA® Jetson Orin™ NX lleva el rendimiento de la supercomputadora de IA al límite, en un sistema en módulo (SOM) compacto, para “Edge Computing”, que es más pequeño que una tarjeta de crédito.

Jetson Orin NX está construido en torno a una **versión de bajo consumo de energía del SoC NVIDIA Orin**, que combina la arquitectura de **GPU NVIDIA Ampere™** con capacidad operativa de 64 bits, procesamiento de imágenes y video multifunción avanzado integrado y **aceleradores de aprendizaje profundo** (LDA) NVIDIA.

El rendimiento de cómputo de hasta 100 TOP INT8 (dispersos) y 50 TOP INT8 (densos) en Jetson Orin NX de 16 GB y hasta 70 TOP INT8 (S) y 35 TOP INT8 (D) en Jetson Orin NX de 8 GB permite que Jetson Orin NX ejecute múltiples redes neuronales en paralelo y procese datos de múltiples sensores de alta resolución simultáneamente.

También ofrece una combinación única de ventajas de rendimiento y energía con un amplio conjunto de E/S, desde CSI y PCIe de alta velocidad hasta I2C y GPIO de baja velocidad, lo que permite dispositivos informáticos integrados y de borde que exigen un mayor rendimiento, pero están limitados por el tamaño, el peso y el consumo de energía.

CHPS IA PARA ENJAMBRE DE DRONES

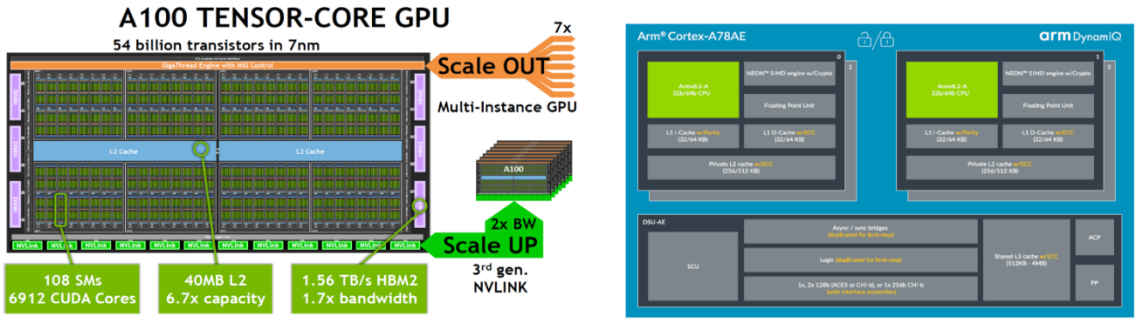


Figura 73. Izda. A100 Tensor-Core GPU y Dcha. ARM Cortex A78AE . Fuente: Nvidia y ARM

Modelo de negocio de los semiconductores

Cadena de Valor de los Semiconductores

Es conocido el creciente coste que deben asumir las industrias de semiconductores para afrontar las innovaciones tecnológicas que se derivan del seguimiento de la Ley de Moore. Este hecho característico de la industria de los semiconductores ha conducido a la fragmentación y globalización de la cadena de valor de dicha industria. Antes de 1.980 las industrias que nacían con nuevas innovaciones de producto llevaban también aparejadas innovaciones de proceso tecnológico y operaban toda la cadena de valor, incluyendo el diseño, la fabricación, el encapsulado y la prueba. Así pues, prácticamente todas las empresas de este sector industrial estaban integradas verticalmente. Estas compañías se conocen como IDM (Integrated Device

Manufacturing). Sin embargo, las compañías IDM, incluso mucho antes de 1.980, efectuaban las operaciones de encapsulado, que era una actividad relativamente simple, de mano de obra intensiva y baja tecnología, en países asiáticos como Malasia y Filipinas, con objeto de beneficiarse de los bajos costes de producción, básicamente de bajos salarios, en dichos países. Muy pronto esta actividad fue subcontratada en compañías radicadas en el sudeste asiático. De esta forma empezó la internacionalización de la industria de semiconductores mediante ***la fragmentación de la cadena de valor*** por la parte en que aportaba menor valor añadido.

A principio de los años 80 aparecen pequeñas compañías con ingenieros muy experimentados, principalmente en Silicon Valley, que aportaban innovaciones significativas basadas en chips. Estas compañías solían contar con la ayuda de capital riesgo, además de las compras públicas promovidas por el Gobierno de los Estados Unidos en los proyectos de la NASA y del departamento de la Defensa. En esa época los costes de montar una fábrica de chips ya eran prohibitivos, aunque mucho menos que en la actualidad, especialmente para compañías "start-up".

Sin embargo, muchos de los fundadores y empleados clave de estas empresas procedían de compañías IDM y conocían que muchas de éstas, especialmente en Japón, tenían **exceso de capacidad**, por lo que llegaban a acuerdos con dichas compañías, para *utilizar esta capacidad sobrante para fabricar los chips que las nuevas y pequeñas empresas lanzaban al mercado*. En principio eran acuerdos beneficiosos para ambos tipos de compañías, ya que las IDM llenaban su capacidad sobrante, con lo que la fabricación de estos chips diseñados por las "start-up" ayudaba a rentabilizar la inversión efectuada y las "start-up" podían implementar sus diseños sin tener que realizar inversiones costosísimas. Las compañías que hacían el diseño del chip, subcontrataban su fabricación, inicialmente a las IDM con capacidad en exceso, y posteriormente comercializaban los chips producidos, se conocen con el nombre de **fabless**. ¡Había nacido el exitoso **modelo de fabless**, compañías que "fabricaban" chips sin tener fábrica!

Este modelo de subcontratación de la capacidad sobrante de las IDM tenía un serio inconveniente: puesto que las IDM también tenían sus propios centros de diseño "in house", las pequeñas compañías que fabricaban sus diseños en las compañías tipo IDM se exponían a que los fabricantes capturaran conocimientos de sus diseños. No debe olvidarse que, a principios de los años 80, existía un alto grado de interrelación entre el diseño y la fabricación de éste. En definitiva, las compañías tipo IDM, que ofrecían capacidad en exceso a las fabless, se convertían en competidores de éstas en el mercado de los productos finales (circuitos integrados). Por este motivo, las compañías fabless preferían proveedores de los servicios de fabricación independientes de las compañías tipo IDM.

El aumento de compañías tipo fabless y la oportunidad de satisfacer sus necesidades dio origen a otro nuevo modelo de negocio, el de las compañías tipo **foundry**. Una foundry es una compañía de semiconductores que no realiza diseños de chips, pero los fabrica por encargo, principalmente de las compañías tipo fabless. La industria de las compañías tipo foundry empezó con la fundación en 1987 de la compañía Taiwan Semiconductor Manufacturing Company (**TSMC**), cuyo CEO era un ingeniero muy experimentado de Texas Instruments (USA), Morris Chang, nacido en China continental y formado como ingeniero en USA.

A medida que el modelo de negocio de las "foundries" empezó a prosperar, las compañías tipo fabless pudieron subcontratar la fabricación de sus diseños a un precio más competitivo, de esta forma también más compañías tipo fabless pudieron entrar en el mercado. En sentido contrario, la prosperidad de modelo de negocio de las compañías tipo fabless influyó en gran medida para

que las industrias tipo foundry también prosperaran, ya que las foundries trabajaban a plena capacidad, lo que las obligaba a un incremento de la capacidad (número de obleas procesadas por mes).

Esta realimentación positiva influyó en el comportamiento de las compañías tipo IDM. En efecto, dichas compañías sufrían directamente el ritmo de inversiones impuesto por el seguimiento de la ley de Moore por lo que muchas de ellas adoptaron en parte el modelo fabless, es decir, parte de su producción la continuaban realizando y parte de los diseños más avanzados, implementados en tecnologías en el límite del estado del arte, se subcontrataban a foundries, este tipo de comportamiento, en el que parte de la producción es subcontratada, al no haber realizado las costosísimas inversiones exigidas por el seguimiento de la Ley de Moore, es lo que se denomina modelo "fab-lite".

Si se examina la cadena de valor de un dispositivo semiconductor, pueden distinguirse tres grandes actividades: a) **diseño**, b) **fabricación obleas (front-end)** c) **Ensamblado, encapsulado y test (Back-end)** (OSAT[16]) (Ver Fig. 74)

Para este informe sobre chips de IA para Defensa es conveniente que se establezcan las definiciones básicas de la cadena de valor.

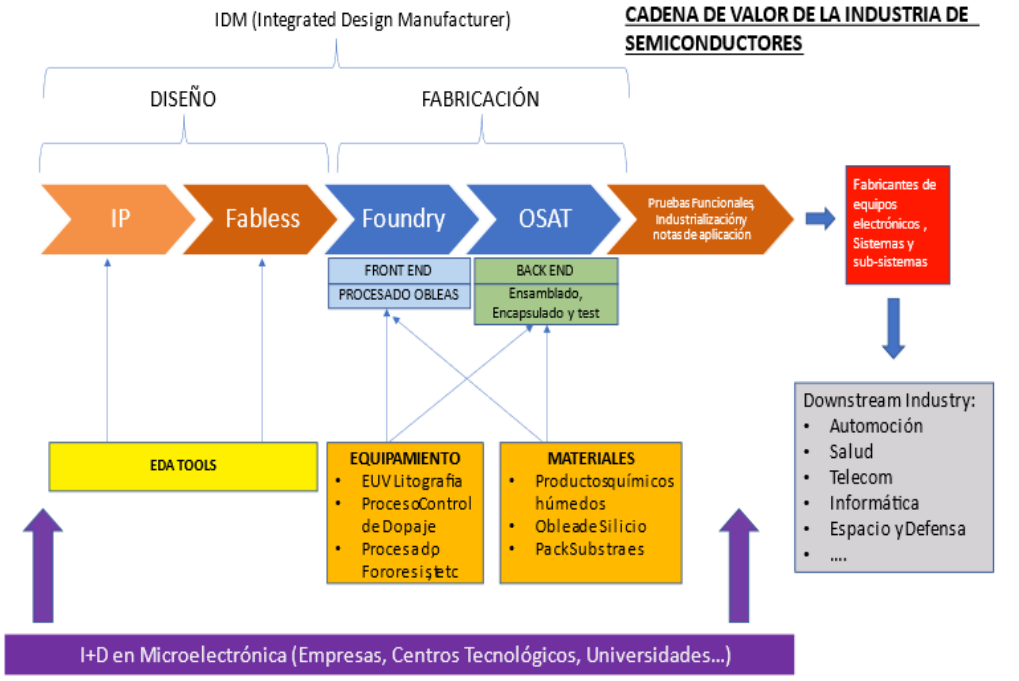


Figura 74. Representación gráfica de la cadena de valor, amplia, de los semiconductores. Fuente: Elaboración propia

Según las funciones que realizan en la cadena de valor y su modelo de negocio, distinguimos:

- Empresas **Proveedores de bloques IP** (Intelectual Property): Son grupos de diseño que producen los bloques de propiedad intelectual (IP) que consisten en partes modulares reutilizables de diseños de chips que se pueden incorporar en diseños de chips completos. Una

[16] **OSAT (Outsourced semiconductor assembly and test):** empresas que brindan capacidad de fabricación de back-end a otros para el ensamblaje y prueba ("encapsulado") de chips.

IP consiste, normalmente, de una descripción hardware en alto nivel simulable, firmware y software, así como un conjunto de vectores de prueba y notas técnicas. Una IP, ya implementada en silicio en un nodo tecnológico compatible con los usuarios que la desean integrar en un chip más complejo, tiene un mayor valor. Las IPs digitales pueden también implementarse en FPGAs. La comercialización de la IP suele hacerse por medio de licencias de uso. También se puede pactar un retorno económico vía royalties por unidad vendida (Chip) en la que se incluye la IP.

- Empresas **Fabless**: Diseñan y venden chips, pero compran servicios de fabricación de chips a las foundries (compañías que ofrecen servicios de fabricación de chips en diversos procesos tecnológicos) y servicios de ensamblaje, prueba y empaquetado a empresas subcontratadas de ensamblaje y prueba de semiconductores (OSAT). Las empresas Fabless, suelen hacer también pruebas funcionales, escriben notas de aplicación y realizan la industrialización (Especificaciones de funcionamiento, pruebas de fiabilidad, calidad, etc.). Estas empresas también acuden a las empresas Proveedoras de bloques IP para acelerar sus diseños y el acceso al mercado de sus productos.

- **Fabricación de chips**: La fabricación convierte los diseños en chips, apoyándose en varios equipos de fabricación de semiconductores (Equipamiento de litografía, implantadores iónicos, hornos de difusión, etc.) y materiales de fabricación (Fotorresist, productos químicos de alta pureza, gases, etc.) en ambientes de extrema pureza del aire (salas blancas).

- Empresas **Foundries**: Son instalaciones de fabricación de semiconductores que fabrican chips para las fabless y clientes de terceros

- Empresas **OSAT (**Outsourced Semiconductor Assembly and Test): Las empresas subcontratadas de ensamblaje y prueba de semiconductores (OSAT) realizan ensamblaje, prueba y encapsulado para clientes externos.

- Empresas **IDM** (Integrated Device Manufacturing): Los fabricantes de dispositivos integrados (IDM) son empresas que realizan los tres pasos de producción: diseño; fabricación; y ensamblaje, prueba y empaquetado. Normalmente el diseño y la fabricación son realizados en instalaciones propias, aunque puntualmente puedan recurrir a centros de diseño externos para integrar en chips partes en las que tienen un menor conocimiento.

- Proveedores de **herramientas EDA** (Electronic Design automation) : Herramientas software que son utilizadas para el diseño de los chips y sistemas electrónicos. Últimamente estas herramientas incluyen, con objeto de acelerar y optimizar los diseños, un uso importante de la IA. Igualmente, los grandes proveedores de herramientas EDA se encargan también de comercializar IPs, por lo que los principales proveedores EDA, son también proveedores de IPs.

- **Equipamiento para fabricación:** El equipamiento de fabricación de semiconductores incluye maquinaria y herramientas utilizadas para fabricar, ensamblar, probar y encapsular chips. (se incluyen los equipos de fotolitografía para generar el conjunto de "máscaras" para un determinado proceso tecnológico, como los EUV[17])

[17] **EUV,** son las siglas de Extreme Ultra Violet, son un tipo de equipos de litografía usados en la fabricación de chips semiconductores para imprimir patrones extremadamente finos (pequeños) en las obleas de silicio para la fabricación del chip. Dichos equipos de litografía utilizan **luz ultravioleta extrema con una longitud de onda cercana a 13,5 nanómetros**, mucho menor que la usada en tecnologías anteriores (como DUV, Deep Ultraviolet, con 193 nm). Esto permite fabricar transistores y circuitos cada vez más pequeños

- Proveedores de **materiales**: Proporcionan los materiales utilizados en la fabricación de los chips: Obleas de silicio, "wet chemicals", gases, materiales para encapsulados, etc.

- **Fabricantes de equipos electrónicos:** son empresas que diseñan y ensamblan sistemas y sub-sistemas, utilizando componentes electrónicos semiconductores proporcionados por compañías Fabless y/o IDM (procesadores, memorias, SoC de comunicaciones, etc) , componentes electrónicos pasivos y componentes mecánicos. Normalmente producen también el software y controlan la fiabilidad y calidad de los equipos que suministran. Estos fabricantes suelen comercializar sus equipos en distintas industrias (Downstream Industry).

- **Downstream Industry**: Son los fabricantes de los sistemas y productos finales en los que se insertan los equipos, sistemas y sub-sistemas electrónicos, que suelen ser proporcionados por los "fabricantes de equipos electrónicos", muchas veces bajo especificación de dichos fabricantes de sistemas finales. Un ejemplo de “downstream Industry" son los fabricantes de automóviles, que incluyen en sus sistemas finales (automóvil), equipos y sistemas electrónicos para control de las diversas partes del vehículo ("powertrain", LIDAR[18], ADAS[19], etc.). Otros ejemplos de "downstream Industry" son los fabricantes de equipamiento para la defensa, aeronaves, aerogeneradores, náutica, la industria "agro-food", robótica, salud, etc.

- **Centro de I+D en microelectrónica**: Aquí se incluyen aquellos Centros Tecnológicos y Universidades (Departamentos o grupos), públicos y privados, que realizan, de forma sistemática, actividades de I+D en uno o varios de los eslabones de la cadena de valor de los semiconductores, aunque no forman parte estricta de dicha cadena. No se incluyen las actividades de I+D que realizan las propias empresas, que transforman los resultados del I+D en innovaciones, y que son la espina dorsal de la Cadena de valor de los semiconductores: Diseño, fabricación...Es por ello por lo que los Centros de I+D contribuyen a la cadena de valor y por tanto forman parte del ecosistema completo de los semiconductores.

Mercado de los Semiconductores

La industria de semiconductores tuvo un sólido 2024, con un crecimiento esperado de dos dígitos (19%) y ventas de **627 mil millones US$** para el año. Pero eso es incluso mejor que el pronóstico anterior de 611 mil millones US$. El año 2025 podría ser incluso mejor, con ventas previstas de **697 mil millones US$,** alcanzando un nuevo máximo histórico. Durante el primer semestre de 2025, el mercado global de semiconductores alcanzó unos **US$ 346.000 millones**, lo que supone un **aumento interanual del 18,9 %,** de acuerdo con las cifras proporcionadas por World Semiconductor Trade Statistics **(** Recent News Release**).** El mayor incremento se ha

y densos. Es esencial en la producción de chips en nodos tecnológicos avanzados (7 nm, 5 nm, 3 nm y menores). Actualmente, hay un solo fabricante a escala global cpaz de producir equipos EUV, la empresa europea ASML.

[18] **LiDAR (Light Detection and Ranging):** Tecnología de detección remota que utiliza pulsos de luz láser para medir distancias y generar representaciones tridimensionales (3D) precisas del entorno. La precisión de la detección es alta, ya que puede detectar objetos con una exactitud de centímetros e incluso milímetros. Tiene múltiples aplicaciones, en Defensa, Automoción (detección de obstáculos), cartografía y topografía, etc.

[19] **ADAS (Advanced Driver-Assistance Systems):** Son **sistemas avanzados de asistencia al conductor** diseñados para **mejorar la seguridad, la comodidad y la eficiencia en la conducción**, mediante sensores, cámaras, radares y software que ayudan a prevenir accidentes o reducir su gravedad.

producido, de acuerdo con WSTS, en procesadores y memorias, impulsados por las demandas de infraestructura para centros de datos y el surgimiento de aplicaciones iniciales de IA en el borde. Parece que las ventas están bien encaminadas para alcanzar la meta aspiracional ampliamente aceptada de **1 billón US$** en ventas de chips para **2030.**

Esto sugiere que la industria solo necesita crecer a una tasa de **crecimiento anual compuesta del 7,5% entre 2025 y 2030** (Ver Fig.76)
Suponiendo que la industria continúe creciendo a ese ritmo, podría alcanzar los US$2 billones en 2040.

Las ventas mundiales de semiconductores aumentaron de 139.000 millones de dólares en 2001 a 555.900 millones de dólares en 2021, una tasa de crecimiento anual compuesta del 7,18 % anual (Ver Fig.75)

Puede apreciarse que el CAGR previsto entre 2025 y 2030 es del 7,5 %, mientras que el CAGR de las dos décadas anteriores (desde 2001 a 2021) es del 7,18 %, es decir, dos CAGR bastante similares, con un ligero incremento de 0,3 puntos para los próximos años.

Si nos preguntamos qué zonas geográficas del mundo son las que contribuyen a un mayor crecimiento, puede constatarse del análisis de WSTS (WSTS Semiconductor Market Forecast Fall 2024) (ver Tabla 11), que, en el año 2024, EEUU crece un 38,9 %, mientras que Europa decrece un 6,7 %. Asia Pacífico crece un 17,5 %, mientras Japón aumenta un 1,4 %.

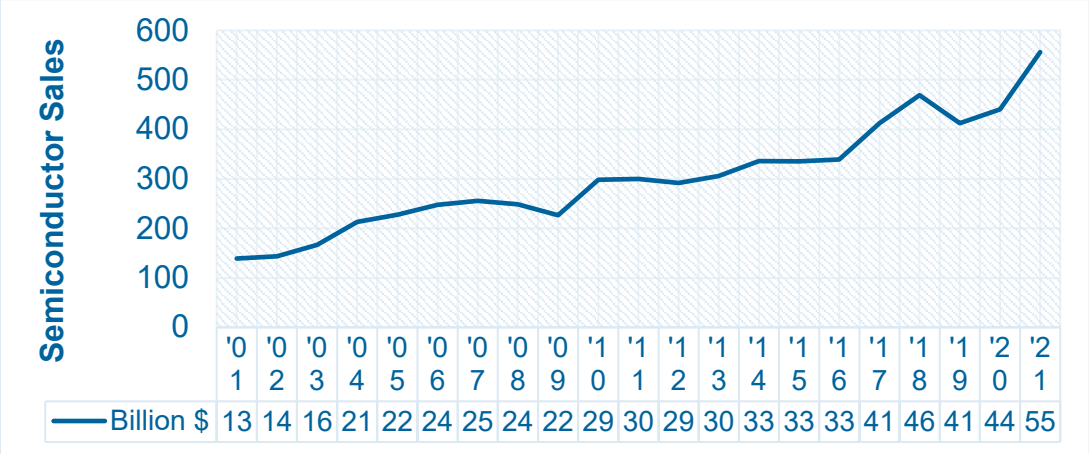


Figura 75. Evolución de las ventas de semiconductores en el mercado global desde 2001 a 2021 en miles de millones de dólares. Fuente: SIA

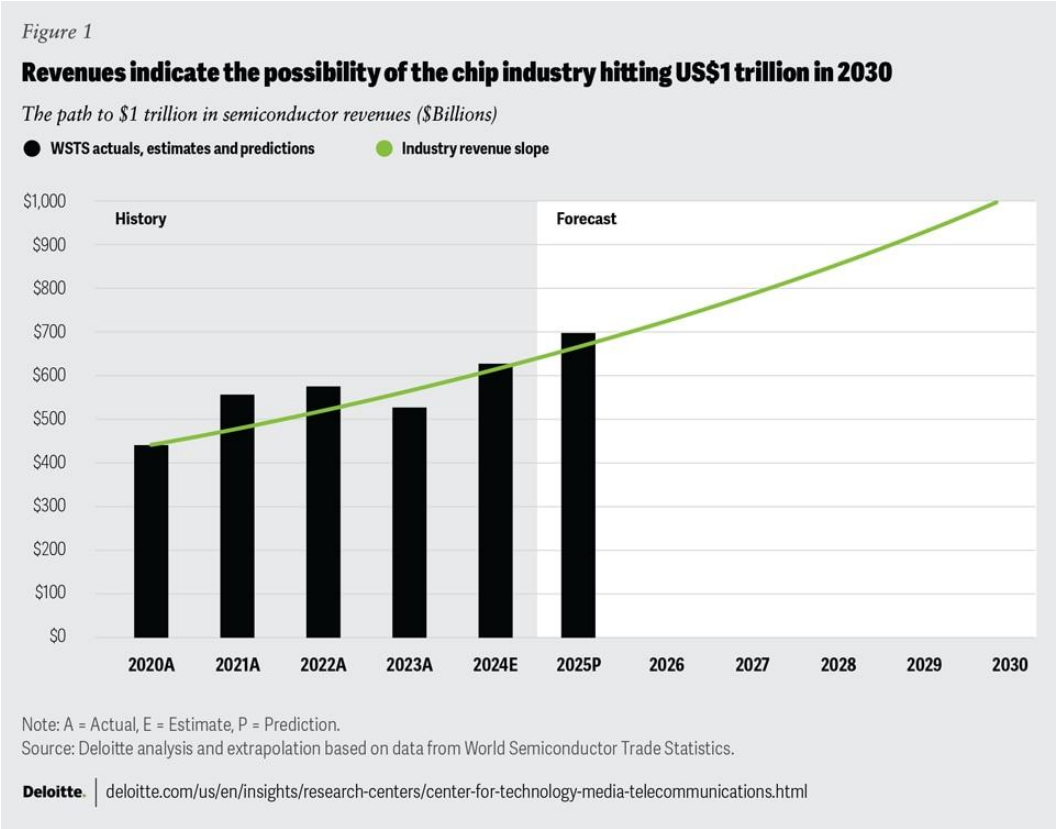


Figura 76. Ventas actuales, estimadas y previsiones hasta 2030, donde espera alcanzarse la cifra de un billón de dólares. Fuente: (Jeroen Kusters, 2025) 2025 semiconductor industry outlook | Deloitte Insights

Con respecto al tipo de semiconductor el crecimiento se da en **Circuitos Integrados** que aumenta un 24,8 %, traccionando fuertemente las memorias con un 81 %, seguido de circuitos lógicos un 16,9 % y microprocesadores un 3.9 %. El fuerte crecimiento de los Centros de Datos, ante la expansión de las aplicaciones de la IA generativa explica la demanda masiva de memorias.

Fall 2024	Amounts in US$M			Year on Year Growth in %		
	2023	2024	2025	2023	2024	2025
Americas	134,377	**186,635**	**215,309**	-4.8	38.9	15.4
Europe	55,763	**52,031**	**53,736**	3.5	-6.7	3.3
Japan	46,751	**47,410**	**51,866**	-2.9	1.4	9.4
Asia Pacific	289,994	**340,792**	**376,273**	-12.4	17.5	10.4
Total World - $M	526,885	**626,869**	**697,184**	-8.2	19.0	11.2
Discrete Semiconductors	35,530	**31,546**	**33,377**	4.5	-11.2	5.8
Optoelectronics	43,184	**42,092**	**43,705**	-1.6	-2.5	3.8
Sensors	19,730	**18,732**	**20,034**	-9.4	-5.1	7.0
Integrated Circuits	428,442	**534,499**	**600,069**	-9.7	24.8	12.3
Analog	81,225	79,433	83,157	-8.7	-2.2	4.7
Micro	76,340	79,291	83,723	-3.5	3.9	5.6
Logic	178,589	208,723	243,782	1.1	16.9	16.8
Memory	92,288	167,053	189,407	-28.9	81.0	13.4
Total Products - $M	526,885	**626,869**	**697,184**	-8.2	19.0	11.2

Tabla 11. Ventas anuales por zonas geográficas y por tipo de semiconductor, para los años 2023, 2024 y previsiones para 2025.

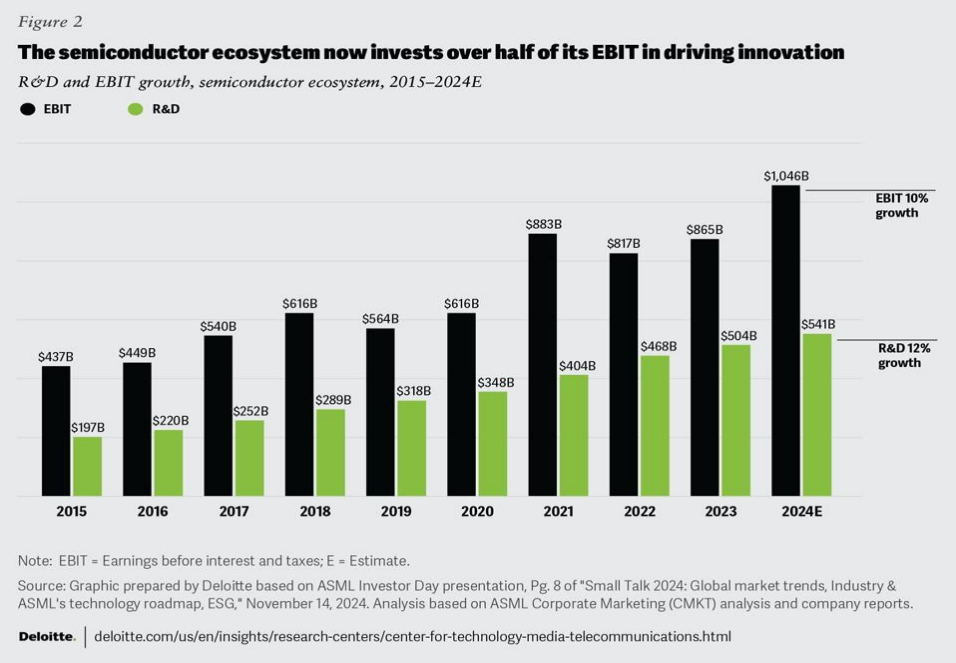


Figura 77. El ecosistema de los semiconductores invierte actualmente la mitad de su EBIT para impulsar la innovación. Fuente: 2025 semiconductor industry outlook | Deloitte Insights

Impulsar la innovación en semiconductores requiere un importante esfuerzo económico. En 2015, el gasto promedio de la industria de chips en investigación y desarrollo fue del 45% de sus ganancias antes de intereses e impuestos (EBIT), pero para 2024, se estima que será el 52% de lo mismo.19 La I+D parece estar creciendo a una CAGR del 12%, mientras que el EBIT solo crece a un 10%. (Ver Fig. 77*)* (Jeroen Kusters, 2025)

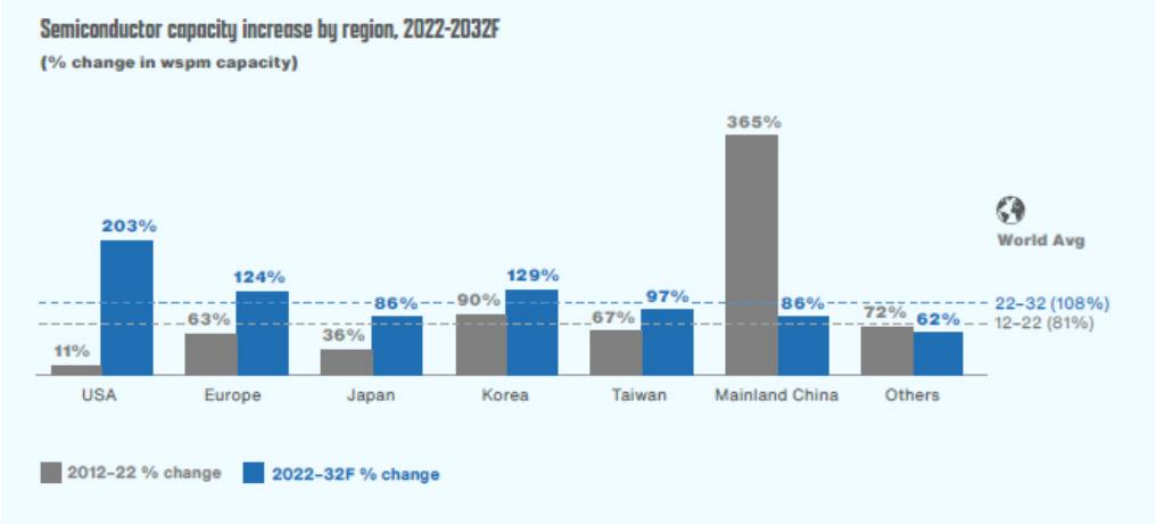


Figura 78. Incremento de la capacidad global de fabricación de semiconductores por zona geográfica. Fuente: SIA Emerging-Resilience-in-the-Semiconductor-Supply-Chain_SIA-Summary.pdf

Vista la demanda creciente de los próximos años, impulsada por el auge de la IA, así como las comunicaciones (5G/6G), el IoT, la HPC (supercomputación), la movilidad autónoma (ADAs, Vehículos autónomos, UAVs, etc.), cabe preguntarse cuál es el incremento de capacidad prevista para los próximos años y compararlo con el crecimiento en capacidad producida durante la década 2012-2022, datos que proporciona (SIA (Semiconductor Industry Association) & BCG (Boston Consulting Group), 2025).

Estos datos pueden observarse en la figura 78. En dicha figura se observa que la Capacidad de producción ha aumentado un 365% en el período 2012-2022 en China y solo un 11 % en EE. UU. Para el período 2022-2032 EE. UU. prevé un aumento del 203%, triplicando la capacidad de los EE. UU. Este cambio de tendencia se corresponde claramente con los problemas en la cadena de suministro provocadas por causas catastróficas, como la pandemia covid-19 y geopolíticas, como el conflicto Taiwan-China. La capacidad de producción se expresa en wpm (Obleas/mes).

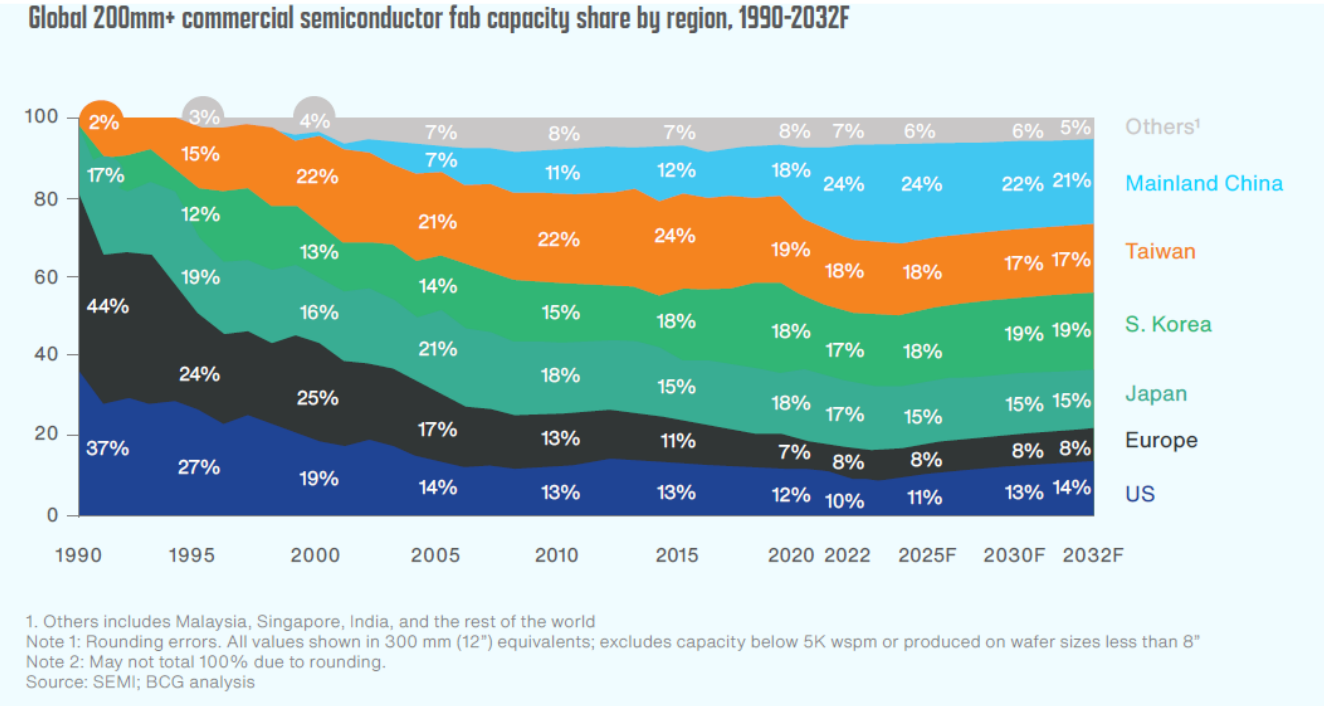


Figura 79. Capacidad de fabricación de semiconductores (obleas de 200 mm y más) por zona geográfica. Fuente: SIA Emerging-Resilience-in-the-Semiconductor-Supply-Chain_SIA-Summary.pdf

Si se examina la distribución en porcentaje de la capacidad de fabricación de las obleas mayores de 8 pulgadas (más de 200 mm) por regiones geográficas desde el año 1990 hasta 2022 (Ver Fig. 79), se puede apreciar que, en 1990, un 44 % de la capacidad correspondía a Europa, un 37 % a EE. UU., un 17 % a Japón y solo un 2 % a Taiwan , mientras que en 2022, la distribución era de un 8% a Europa, un 10 % a EE.UU., un 17 % a Japón, un 17 % a Corea del Sur, un 18 % a Taiwan , un 24 % a China Continental y un 6 % a "Otros"[20]. Este importante cambio en la distribución geográfica de la producción de semiconductores es debido a varios factores: a) En lugar destacado, al **cambio de modelo de negocio de los semiconductores**, al pasar de un mayoría de compañías IDM , al modelo fabless-foundry, esto explica en gran parte la disminución del porcentaje de EE.UU. y de Europa; b) en segundo lugar, la apuesta por la industria de semiconductores de Corea del Sur, especialmente en memorias, que han sido una palanca importante para la producción de semiconductores con la proliferación de PCs, teléfonos móviles, videojuegos y los centros de datos y de **China continental en sus planes tecnológicos**; c) La falta de inversión en nodos tecnológicos avanzados por parte de los IDMs de Europa y EE.UU.

[20] "Otros" incluye a Malasia, Singapur, Vietnam, India y el resto del mundo.

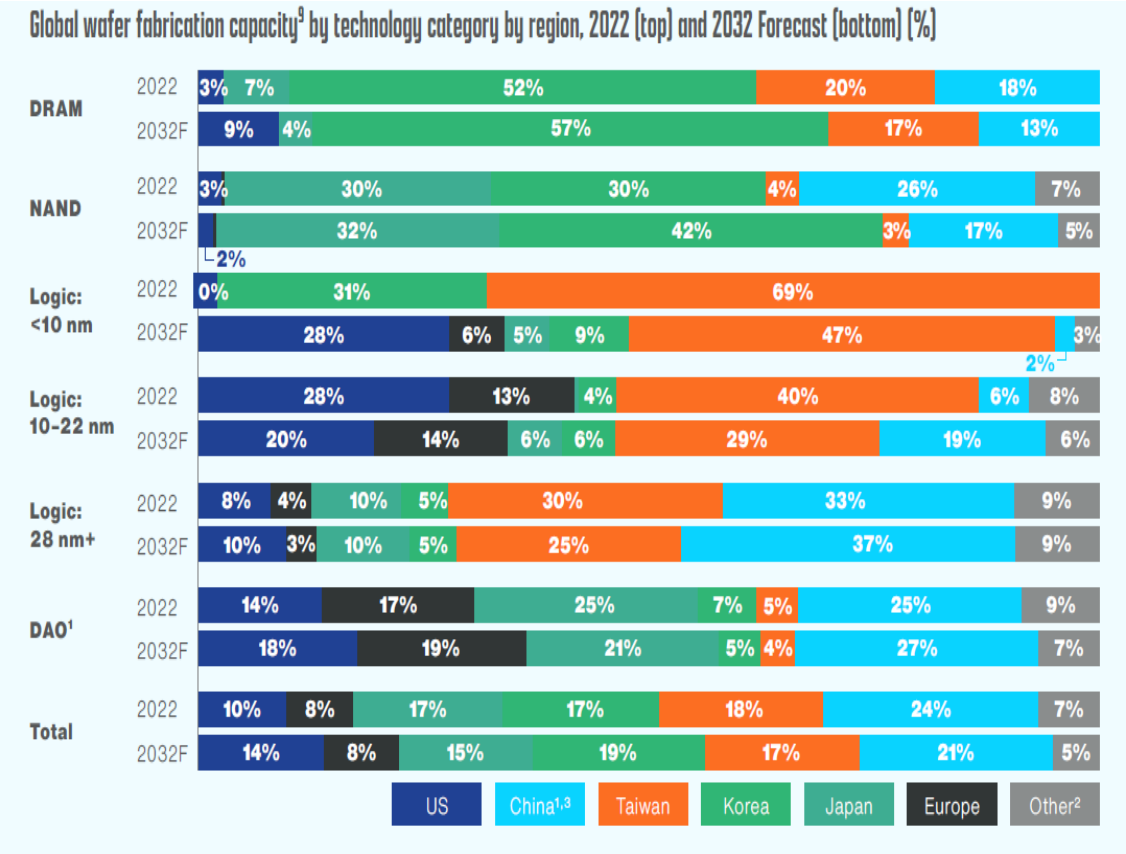


Figura 80. Capacidad de fabricación de semiconductores por tipo de semiconductor y por nodos tecnológicos y por zona geográfica. Fuente: SIA Emerging-Resilience-in-the-Semiconductor-Supply-Chain_SIA-Summary.pdf

En la figura 80 puede observarse en las filas la capacidad para cada tipo de tecnología: Memorias DRAM, memorias no volátiles Flash "NAND", Circuitos lógicos con nodos tecnológicos de <10nm, entre 10 y 22 nm y de >28 nm, DAO (Discretos, Analógicos, Optoelectrónicos) y la última fila el "total", para cada tipo de tecnología hay dos filas: la superior corresponde al año 2022 y la inferior a las previsiones para el año 2032. En cada barra horizontal se indica el porcentaje que corresponde a cada zona geográfica de forma que el total de la barra es el 100 %. Se observa que en los nodos más avanzados (<10nm) en 2022 el 69% corresponde a Taiwan y el 31 % a Corea del Sur, en EE. UU. (y en las demás regiones (Europa, Japón, China, Otros) es 0%.

En la previsión para 2032, EE. UU. salta a un 28 % (principalmente debido a las inversiones de TSMC en EE. UU.), mientras Europa pasa a un 6 %, disminuyendo, lógicamente. Taiwan a un considerable 47 % (sigue siendo líder mundial) y Corea del Sur que pasa a tener un 9 %. Estos cambios se suponen que Estados Unidos aumentará su participación en la capacidad global del 10% al 14%. Sin la Ley CHIPS (CHIPS Act), Estados Unidos habría visto disminuir su participación al 8% de la capacidad global para 2032.

Esto corresponde a la proyección de (SIA (Semiconductor Industry Association) & BCG (Boston Consulting Group), 2025) en que Estados Unidos captará el 28% de los gastos de capital globales, a diferencia del ritmo de inversión previo a la Ley CHIPS, en el que Estados Unidos habría captado solo el 9% de los gastos de capital globales.

Además, de acuerdo con la proyección mencionada, se espera que cada región importante aumente su capacidad en más del 80% durante la próxima década.

Crecimiento previsto del mercado de semiconductores en la década (hasta 20230) por sectores verticales

El crecimiento del 4 al 6 por ciento en el mercado de **computación y almacenamiento de datos** podría ser impulsado por la demanda de servidores para **respaldar aplicaciones como IA** y computación en la nube, según muestra el análisis.

Mientras tanto, en el **segmento inalámbrico**, los teléfonos inteligentes podrían representar la mayor parte de la expansión, en medio de un cambio de segmentos de nivel inferior a segmentos de nivel medio en los mercados emergentes y respaldados por el **crecimiento de 5G**.

Es probable que el segmento de mayor crecimiento sea **el automotriz**, donde podríamos ver una triplicación de la demanda, impulsada por aplicaciones como la **conducción autónoma y la movilidad eléctrica**. (Ver Fig. 81)

Representando solo el 8 por ciento de la demanda de semiconductores en 2021, la industria automotriz podría representar entre el 13 y el 15 por ciento de la demanda para fines de la década. Sobre esa base, el segmento sería responsable de hasta el 20 por ciento de la expansión de la industria en los próximos años.

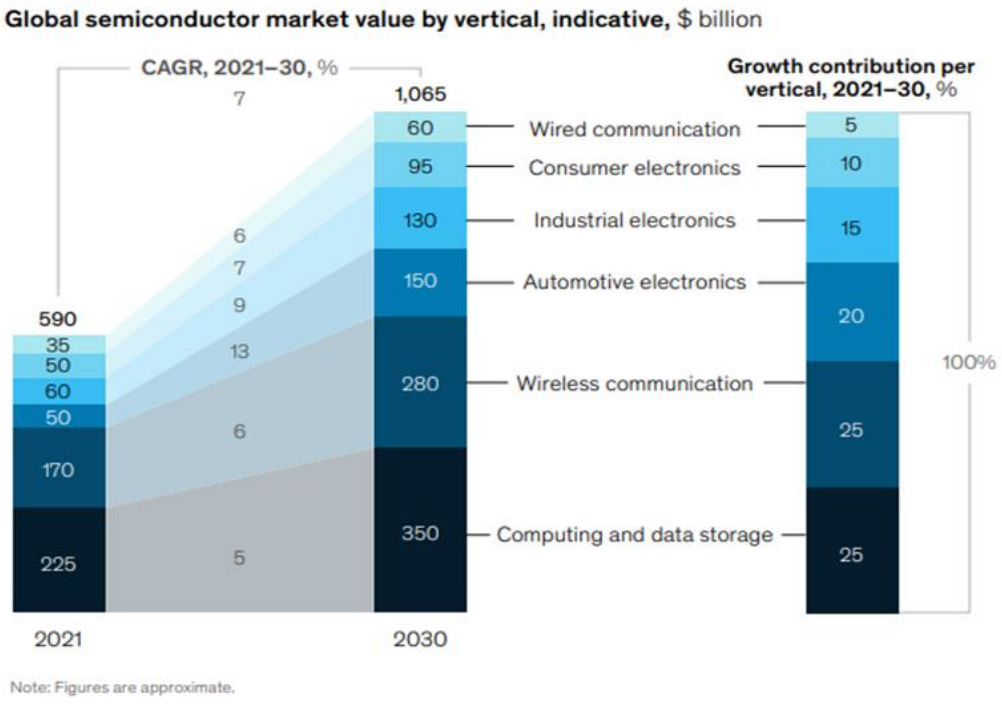


Figura 81. Mercado global de semiconductores por sectores verticales y su crecimiento entre 2021 y el previsto en 2030. Fuente: (Ondrej Burbacky, 2022)

Mercado de los chips para IA

Si se centra la mirada en los chips para aplicaciones de Inteligencia Artificial, se espera un notable crecimiento para la próxima década.

Se prevé que el tamaño del mercado mundial de chips de IA crezca de **123.160 millones de dólares en 2024 a 311.580 millones de dólares en 2029**, creciendo a una CAGR del **20,4 %** durante el período de pronóstico de 2024 a 2029. (Ver Fig. 82)

El mercado de chips de IA está impulsado por la creciente adopción de servidores de IA por parte de los hiper-escalares y el uso creciente de tecnologías y aplicaciones de IA generativa, como GenAI y AIoT, en varias industrias, incluidas BFSI, atención médica, comercio minorista y comercio electrónico, y medios y entretenimiento.

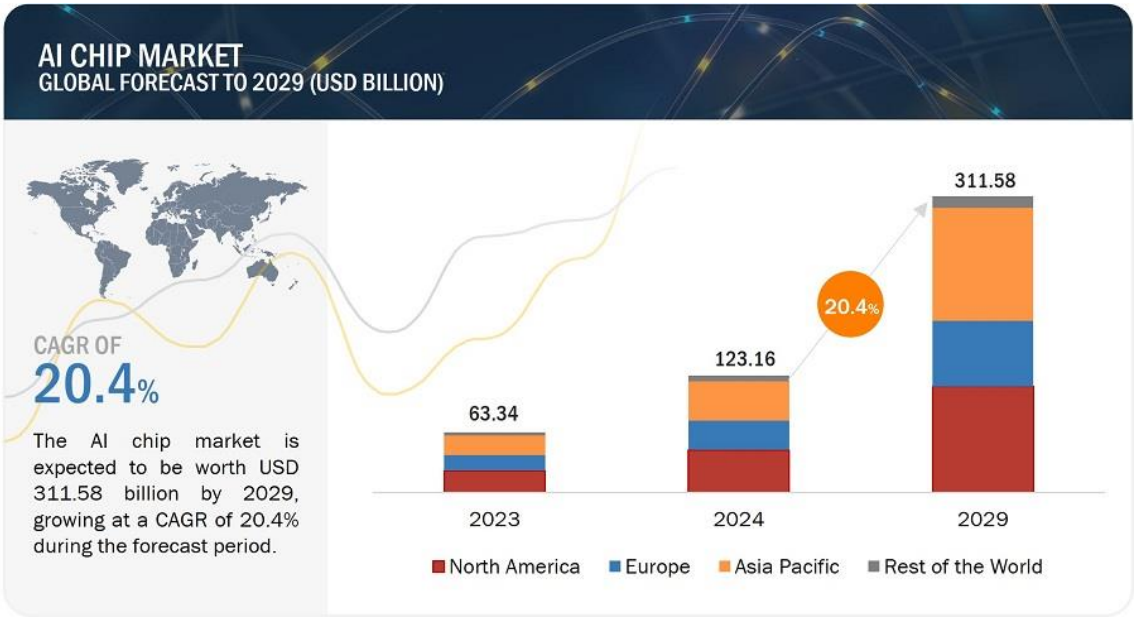


Figura 82. Mercado global de chips para IA por zonas geográficas y su crecimiento entre 2024 y el previsto para 2029. Se prevé un CAGR del 20,4 %. Fuente: AI Chip Market Size, Share, Industry Report 2032

Los chips de IA ayudan a lograr un procesamiento paralelo de alta velocidad en servidores de IA, ofreciendo un alto rendimiento y manejando de manera eficiente las cargas de trabajo de IA en el ecosistema del centro de datos en la nube. Además, la creciente adopción de la computación de IA en el borde y el creciente enfoque en el procesamiento de datos en tiempo real, junto con sólidas inversiones lideradas por las AAPP en el desarrollo de infraestructura de IA, especialmente en las economías de la región de Asia Pacífico, contribuyen aún más al crecimiento de la industria de chips de IA. AI Chip Market Size, Share, Industry Report 2032

Mercado de los Semiconductores para la Defensa

El mercado global de semiconductores para la Defensa ha sido valorado en **26.600 millones de $** en 2021 y la previsión es **superar los 55.000 millones de $ en 2032** con un CAGR de más del 8 % entre 2024 y 2032. El segmento de comunicaciones dominaba el mercado en 2023 con una cuota del 25 %.

Los mayores impulsores en el mercado de semiconductores para la defensa de acuerdo con los análisis de mercados son:

- Aumento de las inversiones gubernamentales en el ámbito militar
- Utilización de componentes semiconductores tolerantes a la radiación
- Aumento de los programas de modernización y actualización de aeronaves
- Creciente demanda de tecnologías avanzadas de semiconductores

- **Aumento del uso de la inteligencia artificial (IA) y el aprendizaje automático (AA) en operaciones militares**

Por el contra, las mayores dificultades/desafíos son:

- Alto costo y largos ciclos de desarrollo
- Restricciones a la exportación y tensiones geopolíticas

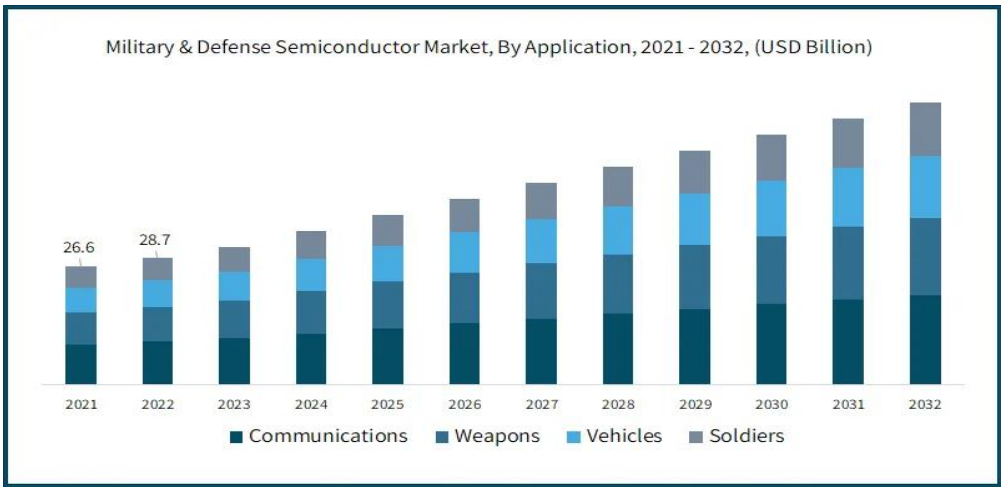


Figura 83. Mercado de los semiconductores para aplicaciones militares y Defensa, 2021-2032 (en USD miles de millones). Fuente: Global Market Insights Military & Defense Semiconductor Market Size & Share - 2032

Centrándonos en el sector aeroespacial, los mayores impulsores del mercado de la defensa:

- Mayor eficiencia energética en aplicaciones de semiconductores aeroespaciales
- Aumento de la demanda de comunicaciones por satélite
- Mayor adopción de IoT en la industria aeroespacial
- Crecimiento de aeronaves eléctricas e híbridas
- Avances en IA para sistemas de aviónica

Figura 84. Mercado de los semiconductores para aplicaciones militares y Defensa, donde puede verse que el CAGR (2024-2032) es > 8%. Fuente: Global Market Insights Military & Defense Semiconductor Market Size & Share - 2032

Ventas de semiconductores por compañías y segmentos cadena de valor.

Los ingresos globales por semiconductores alcanzaron los **626 mil millones de dólares en 2024,** lo que marca un aumento del **18,1 %** con respecto al año anterior, según los resultados preliminares de Gartner, Inc. Se proyecta que el mercado seguirá creciendo, y se espera que los ingresos totales por semiconductores alcancen los 705 mil millones de dólares en 2025.

Los principales impulsores de este crecimiento fueron **las unidades de procesamiento gráfico (GPU) y los procesadores de IA** utilizados en aplicaciones de centros de datos, incluidos servidores y tarjetas aceleradoras. A medida que aumentó la demanda de cargas de trabajo de IA e IA generativa (GenAI), los centros de datos se convirtieron en el segundo mercado más grande de semiconductores, detrás de los teléfonos inteligentes. En 2024, los ingresos por semiconductores de los centros de datos alcanzaron los 112 mil millones de dólares, frente a los 64,8 mil millones de dólares de 2023.

El desempeño positivo general del mercado de semiconductores generó un fuerte crecimiento para muchos proveedores, y once de las principales empresas de semiconductores experimentaron un crecimiento de dos dígitos. Solo ocho de los 25 principales proveedores registraron caídas de ingresos en 2024.

Top Semiconductor Companies' Revenue

Change versus prior quarter in local currency

	Company	US$B 2Q24	Reported 2Q24	Guidance 3Q24	Comments on 3Q24
1	Nvidia	28.0	7.5%	n/a	2Q24 is guidance from 1Q24
2	Samsung SC	20.7	23%	n/a	strong demand from server AI
3	Broadcom	13.0	4.1%	n/a	2Q is estimate
4	Intel	12.8	0.9%	1.3%	excess inventory
5	SK Hynix	11.9	32%	n/a	strong demand from server AI
6	Qualcomm (IC)	8.1	0.5%	4.1%	handset normal seasonality
7	Micron	6.8	17%	12%	supply below demand
8	AMD	5.8	6.6%	15%	growth in data center & client
9	Infineon	4.0	1.9%	8.0%	growth across all divisions
10	MediaTek	3.9	-4.6%	0.5%	flat across all groups
11	TI	3.8	4.4%	7.3%	strength in personal electronics
12	STMicro	3.2	-6.7%	0.6%	auto & personal electronics up
13	NXP	3.1	0.0%	3.9%	auto, IoT & mobile up
14	Kioxia	2.7	33%	n/a	supply-demand in balance
15	Analog Devices	2.3	7.1%	3.8%	inventory levels improving
	Total of above		8%		
	Memory Cos. (US$)		22%		Samsung, SK Hynix, Micron, Kioxia
	Non-Memory Cos.		3%	5%	companies with guidance

Tabla 12. Ventas de las principales compañías de semiconductores en el segundo trimestre de 2024

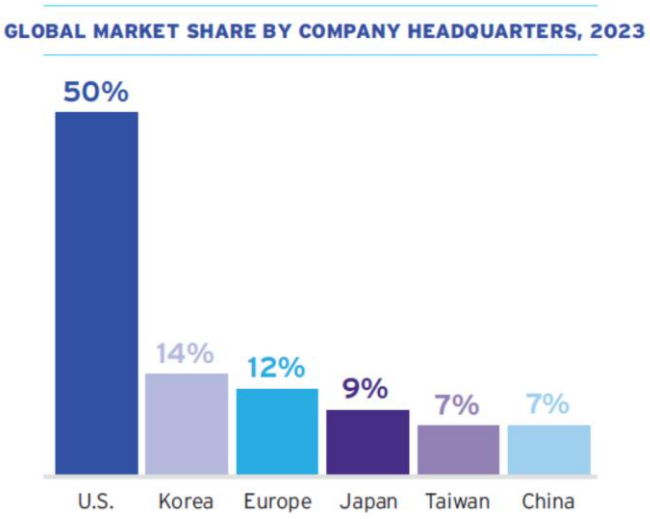


Figura 85. Cuota de mercado de las compañías según localización de sus sedes centrales (HQ). Fuente: (Ondrej Burbacky, 2022)

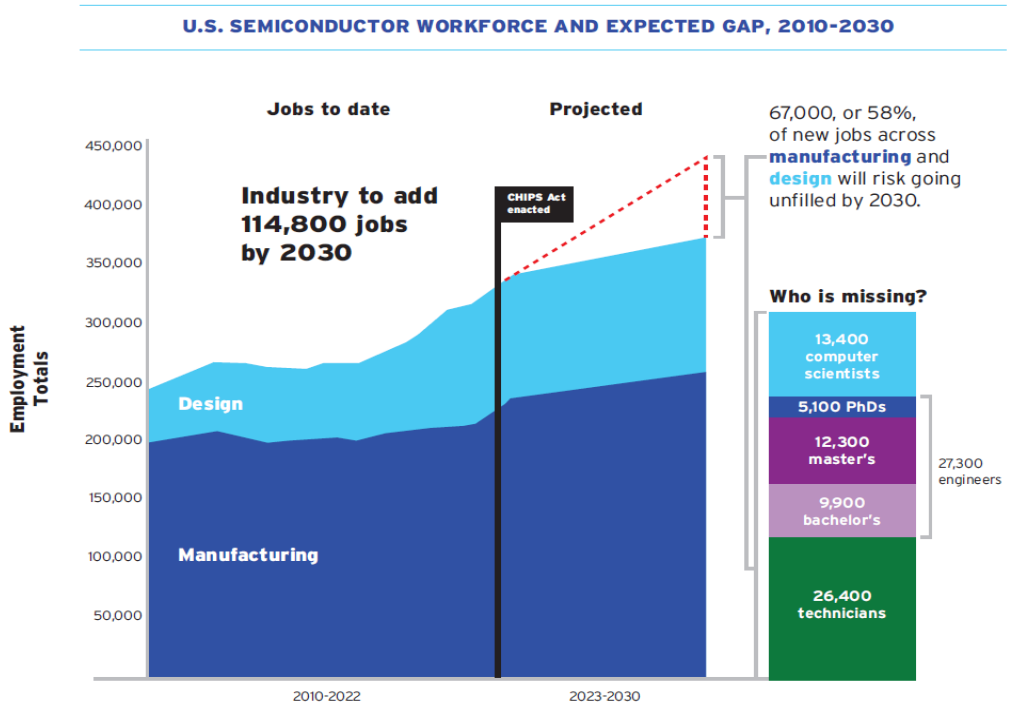


Figura 86. Déficit de puestos de trabajo para diseño y fabricación de semiconductoras previsto en EE. UU. para 2030. Fuente: (Ondrej Burbacky, 2022)

Foundries

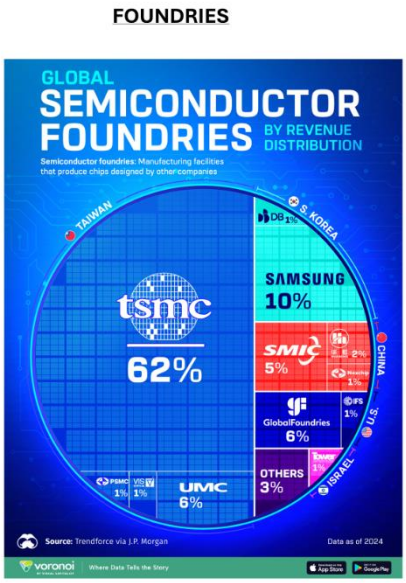
A partir de enero de 2025, varios países asiáticos, incluidos Taiwán, China y Corea del Sur, serán actores clave en el mercado de "foundries" de semiconductores.

Taiwán domina específicamente, con TSMC por sí sola con una considerable participación del 62% de los ingresos totales de fundición.

Si se incluyen las contribuciones de otras empresas **con sede en Taiwán, como UMC, VIS y PSMC, la participación combinada del país supera el 70%.**

Le sigue Corea del Sur, con Samsung representando el 10% del mercado. Mientras tanto, Estados Unidos y China hacen contribuciones notables, lideradas por GlobalFoundries (6%) y SMIC (5%), respectivamente.
China, Estados Unidos y Corea del Sur están intensificando los esfuerzos para expandir sus industrias de semiconductores a través de mayores inversiones y apoyo gubernamental para reducir la dependencia de Taiwán.

Si bien las empresas estadounidenses son los líderes mundiales en la industria general de semiconductores, representando el 71,5% de la capitalización de mercado global, este dominio está impulsado principalmente por sus fortalezas en diseño de chips, propiedad intelectual (PI) y herramientas de software (EDA). La fabricación de los chips más avanzados se lleva a cabo en el extranjero, en fundiciones como TSMC. Estas fábricas utilizan equipos de última generación y altamente especializados, como máquinas de litografía ultravioleta extrema (EUV), para fabricar chips avanzados con precisión a escala nanométrica, lo que las hace excepcionalmente capaces de producir los semiconductores más complejos y potentes.



Compañía	País	Cuota sobre el total mundial de foundries
Taiwan Semiconductor Manufacturing Company (TSMC)	Taiwan	62%
Samsung	Corea del Sur	10 %
United Microelectronics Corporation (UMC)	Taiwan	6%
GlobalFoundries (GF)	EEUU	6%
Semiconductor Manufacturing International Corporation (SMIC)	China	5%
Hua Hong Semiconductor	China	2%
Tower	EEUU	1%
Vanguard International Semiconductor Corporation (VIS)	Taiwan	1%
Powerchip Semiconductor Manufacturing Corporation (PSMC)	Taiwan	1%
Nexchip	China	1%
DB HiTek	Corea del Sur	1%
Tower Semiconductor	Israel	1%
IFS	EEUU	1%
Otros	Otros	3%

Fuente: Trendforce via J.P.Morgan

Figura 87. Principales Foundries a nivel global con tabla indicando países en los que están establecidas sus sedes (HQs) y porcentaje de mercado global. Fuente: Trendforce

Compañías Fabless

La Tabla 13 muestra un listado de las 10 mayores compañías de semiconductores Fabless de acuerdo con su cifra de ventas anual en el año 2024. Como puede apreciarse la compañía líder es Nvidia que produce el hardware (fundamentalmente diseña los chips (principalmente GPUs) que se utilizan en aplicaciones de IA. Como puede apreciarse, solo Nvidia factura del orden de más de 60 mil millones de dólares.

Tabla 13. Listado de las 10 mayores compañías de semiconductores fabless de acuerdo con su cifra de ventas anual (2024)

NOMBRE COMPAÑÍA	VENTAS ANUALES (USD) 2024	HEADQUARTERS	AÑO FUNDACIÓN
Nvidia Corporation	60.9 billion	Santa Clara, California, U.S	1993
Qualcomm Inc.	36.99 billion	San Diego, California, U.S.	1985
Broadcom Inc.	35.82 billion	Stanford Research Park, Palo Alto, California, U.S.	1961
Advanced Micro Devices Inc. (AMD)	23.3 billion	Santa Clara, California, United States	1969
MediaTek Inc.	15.58 billion	Hsinchu, Taiwan	1997
Marvell Technology, Inc.	5.50 billion	Wilmington, Delaware, U.S	1995
Realtek Semiconductor Corporation	3.02 billion	Hsinchu Science Park, Hsinchu, Taiwan	1987
Altera Corporation	1.99 billion	San Jose, California, U.S.	1983
Cirrus Logic	1.79 billion	Austin, Texas, US	1984
LSI Logic Corporation	469.64 million	Milpitas, California, U.S	1980

Geopolítica y semiconductores

Situación de dependencia de Europa con EEUU en chips para IA.

(Bourzac, January 2025)

En general, la dependencia en chips semiconductores avanzados de Europa con respecto a EE. UU. se va agrandando, con una brecha importante, no tanto por la fabricación avanzada en la que EE. UU. también tiene déficit, sino por las empresas de diseño fabless (8 de las top 10, son

de EE. UU. y de ellas las 4 primeras son americanas), herramientas EDA e IPs. Igualmente, en el campo de los chips IA, especialmente GPUs y FPGAs, son todas compañías americanas.

Sin embargo, estas compañías dependen de las foundries en Taiwan, especialmente las que fabrican con nodos tecnológicos de 7nm o inferiores.

Con la situación geopolítica actual, en la que Taiwan, está en un escenario con un riesgo no despreciable de conflicto bélico, así como de riesgos de catástrofes naturales, tampoco despreciables, al que puede añadirse riesgos derivados de posibles pandemias, como ocurrió con el covid-19, tanto EE. UU., como Europa, intentan atraer, con importantes ayudas, a fabricantes asiáticos para que instalen foundries con tecnologías inferiores a 7nm en sus territorios.

Las únicas atracciones que han tenido cierto éxito, por el momento, son las de EE. UU., como la fábrica de TSMC en Arizona.

La dependencia de fabricación de Taiwán y Corea.

TSMC opera:

- cuatro fábricas de obleas GIGAFAB® de 12 pulgadas, en Taiwan
- cuatro fábricas de obleas de 8 pulgadas, en Taiwan
- una fábrica de obleas de 6 pulgadas, en Taiwan
- una fábrica de obleas de 12 pulgadas en una subsidiaria de propiedad absoluta, TSMC Nanjing Company Limited,
- Una fábrica de obleas de 12 pulgadas en Japón (Japan Advanced Semiconductor Manufacturing, Inc.)
- dos fábricas de obleas de 8 pulgadas en subsidiarias de propiedad absoluta, TSMC Washington en los Estados Unidos y TSMC China Company Limited.
- **Nueva planta en Arizona:** TSMC está construyendo una nueva fábrica en Arizona como parte de la iniciativa estadounidense para fortalecer la producción nacional de semiconductores. Esta planta forma parte de los esfuerzos para reducir la dependencia de la producción asiática y reforzar la cadena de suministro de chips en Estados Unidos. La planta de TSMC en Arizona ha iniciado recientemente la producción de semiconductores utilizando el nodo tecnológico N4, que pertenece a la familia de 5 nanómetros (nm). Este nodo es una versión mejorada **del proceso de 5 nm**, ofreciendo mayor eficiencia y rendimiento. Los primeros productos fabricados en esta instalación incluyen los chips A16 y S9 para Apple
- Cinco fábricas de “Back-end”, todas en Taiwan
- El Gobierno de Taiwán, ha declarado que, aunque TSMC puede construir fábricas en EE. UU., Europa y Japón, como está haciendo, no puede transferir sus tecnologías de integración más avanzadas **a instalaciones de producción en el extranjero** debido a que están protegidas por la ley taiwanesa. En la práctica esta regulación implica que TSMC no podrá fabricar chips de 2 nm fuera de Taiwán mientras esta sea su mejor litografía.

Estado de las foundries en Europa.

1. **Global Foundries en Dresde (Alemania) :** GlobalFoundries opera una de las plantas de semiconductores más grandes de Europa en Dresde. Esta instalación es fundamental para la producción de chips utilizados en sectores como el automotriz, industrial y de consumo. gf.com . **Es la única foundry en funcionamiento en Europa.**
2. **TSMC en Dresde, Alemania:** Taiwan Semiconductor Manufacturing Company (TSMC) ha iniciado la construcción de su primera planta en Europa, ubicada también en Dresde. Este proyecto, en colaboración con Bosch, Infineon y NXP, representa una inversión de aproximadamente 10.000 millones de euros y se espera que la producción comience en 2027.
3. **Intel en Magdeburgo, Alemania:** Intel ha anunciado planes para construir una planta de fabricación de semiconductores en Magdeburgo. Esta instalación está destinada a ser una de las más avanzadas del mundo, produciendo chips de 1,5 nanómetros y reforzando la capacidad de producción de semiconductores en Europa. micro-electronics.eu En junio de 2025 Intel anunció que cancelaba definitivamente la construcción de la planta de fabricación de chips en Magdeburgo (***Está cancelada por problemas internos de Intel y el Gobierno alemán se replantea reasignar los 10.000 millones de euros a otro proyecto***).
4. **Infineon en Dresde, Alemania:** Infineon Technologies está construyendo una nueva planta de fabricación de semiconductores en Dresde, conocida como MEGAFAB-DD. Este proyecto, con una inversión de 3.500 millones de euros, ha recibido 920 millones de euros en ayudas estatales aprobadas por la Comisión Europea y se espera que alcance su plena capacidad para 2031. reuters.com
5. **Proyecto Sparc en Vigo, España:** En Vigo, se está planificando la instalación de una fábrica de microchips fotónicos denominada Sparc. Este proyecto, que cuenta con el apoyo de la empresa Indra y diversas instituciones, será la tercera fábrica de su tipo en Europa y busca impulsar la industria tecnológica en la región. cadenaser.com

Estas iniciativas forman parte de la estrategia de la Unión Europea para fortalecer su autonomía tecnológica y asegurar el suministro de semiconductores, especialmente en sectores clave como el automotriz e industrial. **El European Chips Act** es una de las medidas implementadas para atraer inversiones y mejorar la capacidad de producción en el continente. digital-strategy.ec.europa.eu

Además de estas plantas, Europa alberga otras instalaciones de fabricación de semiconductores **(IDM)** operadas por empresas como **STMicroelectronics**, **NXP Semiconductors** y **Bosch**, que contribuyen significativamente al ecosistema tecnológico europeo

TSMC en Europa

Taiwan Semiconductor Manufacturing Company (TSMC) ha iniciado la construcción de su primera planta en Europa, ubicada en Dresde, Alemania. Esta instalación, que representa una inversión de aproximadamente 10.000 millones de euros, se está desarrollando en colaboración con Bosch, Infineon y NXP, formando la empresa conjunta European Semiconductor Manufacturing Company (ESMC). La Comisión Europea ha aprobado una ayuda estatal de 5.000 millones de euros para este proyecto, con el objetivo de fortalecer la resiliencia de la cadena de suministro de semiconductores en Europa y reducir la dependencia de proveedores externos. reuters.com

La planta se centrará en la producción de chips con tecnologías avanzadas de 28/22 nm CMOS y 16/12 nm FinFET, con una capacidad estimada de 40.000 obleas de 300 mm al mes. Se espera que la construcción comience en el cuarto trimestre de 2024 y que la producción inicie en 2027, generando alrededor de 2.000 empleos directos. es.rti.org.tw

Este proyecto es parte de la estrategia de la Unión Europea para incrementar su autonomía tecnológica y asegurar el suministro de semiconductores, especialmente para industrias clave como la automotriz. La ubicación en Dresde, conocida como "Silicon Saxony" por su concentración de empresas

Control de acceso a las tecnologías

Los expertos en China también han prestado mucha atención a la **importancia estratégica de los chips de IA**.

Un informe de 2018 de la Universidad de Tsinghua en Beijing lo expresó en términos claros: "*Ya sea la realización de algoritmos, la adquisición y una base de datos masiva, o la capacidad de computación, **el secreto detrás del rápido desarrollo de la industria de la IA radica en la única base física, es decir, los chips**. Por lo tanto, no es exagerado decir, 'Sin chip, no hay IA' dado el papel irreemplazable del chip de IA como piedra angular para el desarrollo de la IA y su importancia estratégica*". Al igual que sus homólogos de Estados Unidos, los líderes chinos, incluido Xi Jinping, secretario general del Partido Comunista Chino, creen que el liderazgo en IA es fundamental para el futuro del poder económico y militar.

La estrategia nacional de IA de China de 2017 describe a la IA como "**un nuevo foco de competencia internacional**". El ejército chino se embarcó en su campaña de modernización de la IA aproximadamente al mismo tiempo que Estados Unidos en 2017, y se encontró con el mismo problema que descubrió el **Proyecto Maven**: **una necesidad desesperada de chips de IA**. (Allen, 2023)

Pero el ejército chino se enfrentó a un problema adicional: **las empresas chinas no fabricaban chips de IA**, por lo que el gobierno chino se embarcó en una estrategia de dos frentes. Primero, comprar los chips de IA de Estados Unidos (por ahora). Segundo, desarrollar alternativas chinas a los chips estadounidenses (lo antes posible). La realidad de las compras militares chinas de chips de IA no está en debate. Se publicó abiertamente en contratos de adquisiciones militares chinos no clasificados. Investigadores del Centro de Seguridad y Tecnología Emergente de la Universidad de Georgetown revisaron 21.088 contratos de este tipo entre abril y noviembre de 2020 y descubrieron que todas las compras de chips de IA especificaban productos estadounidenses. Ni un solo contrato especificaba la compra de chips de IA a empresas chinas. La misma arquitectura de paralelización masiva que hace que las GPU sean tan atractivas para las aplicaciones de IA también se aplica a otros tipos de trabajo computacional complejo, como el modelado y la simulación aeroespacial de alta fidelidad para misiles hipersónicos y armas nucleares.

Una revisión reciente del Wall Street Journal de los registros de adquisiciones militares chinas desde 2020 descubrió que el principal instituto de investigación de armas nucleares de China también es un comprador habitual de chips GPU estadounidenses. La comunidad de inteligencia estadounidense sin duda está al tanto de otras compras y planes chinos (clasificados). Sin embargo, el objetivo de producir chips de IA chinos que compitan con los productos

estadounidenses ya estaba firmemente establecido en la mente de los líderes de China, dos años antes de que comenzara el Proyecto Maven. En septiembre de 2015, el Consejo de Estado de China publicó "**Made in China 2025"**, la piedra angular de la política industrial china. El principal objetivo para los semiconductores en "Made in China 2025" era "la sustitución de las importaciones por productos fabricados en China, básicamente logrados en industrias clave".Blocking China's Access to AI Chips Matters to U.S. National Security

El objetivo político explícito de China, el que subyace a su anuncio de 2014 de más de **100 mil millones de dólares en subsidios a la industria de semiconductores**, era **reemplazar los chips** y las tecnologías de fabricación de chips estadounidenses y aliados **con alternativas chinas**. Este era el plan de China no solo antes de que el presidente Biden asumiera el cargo, sino dos años antes de que Donald Trump asumiera el cargo, en su primer mandato.

Por supuesto, China redobló la apuesta por el plan a raíz de la guerra comercial que se estableció en el primer mandato de Trump y las **sanciones** a empresas tecnológicas chinas como **ZTE** y **Huawei**. Cuando el presidente Biden asumió el cargo en enero de 2021, era una prioridad de seguridad nacional china.

Este es el contexto vital a través del cual se deben analizar **los controles de exportación del 7 de octubre de 2022**, de la administración Biden, que restringieron las exportaciones estadounidenses de chips avanzados de inteligencia artificial y el equipo avanzado utilizado para fabricarlos.

Los líderes de China ya habían decidido hacer todo lo posible para eliminar su dependencia de la tecnología estadounidense. No había ninguna posibilidad realista de persuadir a China de que abandonara esa política. Por lo tanto, los controles de exportación del 7 de octubre se centran en garantizar que ambos aspectos de la estrategia china fracasen.

Para impedir que China compre chips avanzados estadounidenses, los controles de exportación se aplican a nivel nacional, no solo a las organizaciones militares chinas conocidas. Esa es una respuesta directa al hecho de que la estrategia china de fusión militar-civil ha funcionado para profundizar y oscurecer los vínculos entre las empresas comerciales chinas y el ejército chino. Aunque las empresas estadounidenses han cumplido de manera confiable con las restricciones de control de exportación de "sin usuario final militar" anteriores a 2022**, las redes de revendedores dentro de China han podido hacer llegar los chips al ejército de manera confiable.** (Allen, 2023)

Para evitar que China fabrique alternativas chinas a los chips de IA estadounidenses, los **controles de exportación siguen permitiendo las ventas de chips por debajo de ciertos umbrales de rendimiento tecnológico** y también **restringen la venta de equipos avanzados para la fabricación de chips**. Las empresas chinas de diseño de chips de IA como Biren y Cambricon estaban pisándole los talones a las empresas estadounidenses como Nvidia antes de los controles de exportación del 7 de octubre. Tal vez estaban solo un año detrás de Nvidia en términos de calidad de diseño de chips. Pero ahora, esas empresas chinas de diseño de chips no pueden acceder a los equipos avanzados de fabricación de semiconductores estadounidenses. Las autoridades americanas también han estado negociando para que los proveedores de equipos de fabricación de chips avanzados, como Países Bajos (ASML) y Japón (TEL), que proporcionan equipos clave para la fabricación de chips con nodos tecnológicos por debajo de los 7 mm como los EUV de litografía, en el caso de ASML y los equipos de deposición en fase vapor como TEL en Japón también se sumasen a los controles de exportación.

Para construir tecnologías de alto rendimiento (nodos iguales o inferiores a 7 nm) las empresas chinas tienen que escalar un Everest, con los recursos y esfuerzo que ello comporta, siendo el tiempo, quizás, la principal barrera. Sin embargo, como se ha visto, en este informe, los avances en chips de IA, no son únicamente conseguir chips fabricados con los nodos tecnológicos más avanzados. Influyen otros parámetros, como se ha podido comprobar con DeepSeek, algoritmos y software que aproveche los recursos hardware de forma más eficiente, nuevos diseños que eviten el movimiento de datos entre los procesadores y la memoria, co-diseño hardware-software, nuevos algoritmos, etc. Sin embargo, no hay que olvidar la geopolítica. Los chips más avanzados se producen en Taiwan, que China considera parte indisociable de su territorio y está unida al continente por vínculos idiomáticos y culturales, por lo que no se pueden descartar flujos de información hacia el continente, que faciliten el proceso de que China entre en el club de los chips avanzados, más pronto de lo que pueda parecer debido al freno que suponen los controles de exportación del 7 de octubre, que, por cierto, fueron actualizados poco más de un año después, por el "Bureau of Industry and Security" del Departamento de Comercio de los Estados Unidos, que ha publicado un conjunto actualizado de normas (Ver Fig. 88) que revisan las originales en un intento de cerrar varias lagunas que debilitaban la eficacia de las normas. Estas normas no representan una desviación importante de la intención original de los controles de exportación, pero sí contienen cambios significativos que afectan su funcionamiento y sus objetivos.

FOR IMMEDIATE RELEASE
October 17, 2023
https://bis.doc.gov

BUREAU OF INDUSTRY AND SECURITY
Office of Congressional and Public Affairs
Media Contact: OCPA@bis.doc.gov

Commerce Strengthens Restrictions on Advanced Computing Semiconductors, Semiconductor Manufacturing Equipment, and Supercomputing Items to Countries of Concern

Updates to Modify and Reinforce Restrictions Initially Released on October 7, 2022, to Address National Security Concerns Posed by PRC Military Modernization

Figura 88. Actualizaciones sobre las restricciones a la exportación de chips IA del 7 octubre 2022. Fuente:https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3355-2023-10-17-bis-press-release-acs-and-sme-rules-final-js/file

Las actualizaciones contienen tres reglas principales:

- **Primera: Regla de chips de computación avanzada (AC/S IFR)**
  La IFR AC/S conserva los estrictos requisitos de licencia para toda la República Popular China impuestos en la regla del 7 de octubre de 2022 y realiza dos categorías de actualizaciones:

• a) Ajusta los parámetros[21] que determinan si un chip de computación avanzado está restringido;
• b) Impone nuevas medidas para abordar los riesgos de elusión de los controles[22]

- **Segunda: Se refiere a los equipos de fabricación de semiconductores (SME).**
  Esta norma añade **varias docenas de artículos a la lista de equipos controlados**, la mayoría de los cuales se utilizan para fabricar chips lógicos por debajo de un umbral de 16 nanómetros.
  Las incorporaciones incluyen equipos utilizados en el procesamiento químico húmedo, grabado en seco y varios tipos de deposición, entre otros. Estas normas tienen por objeto alcanzar la paridad con las recientes restricciones de licencias japonesas y holandesas como parte de **un acuerdo trilateral alcanzado con los Estados Unidos** en enero 2023, aunque en ciertos equipos de litografía ultravioleta profunda (DUV), las normas estadounidenses actualizadas superan los umbrales holandeses.

- **Tercera: Amplía el número de empresas de la Lista de entidades (Entity List)**
  La "Entity List" es el equivalente estadounidense de una **lista negra de empresas** a las que se prohíbe la exportación de cierta tecnología. Esta nueva norma añade 13 entidades a esta lista. Estas entidades están involucradas en el desarrollo de inteligencia artificial que la administración americana cree una amenaza a la seguridad nacional y los objetivos de política exterior de Estados Unidos.

Política de sanciones y contra sanciones

Los semiconductores son el insumo estratégico que condiciona cada vez más la capacidad de países y empresas de actuar económica, industrial y militarmente (Duchâtel; 2022; SWD, 2022). Estados Unidos inventó el semiconductor, pero hoy produce alrededor del 12% del suministro mundial, y ninguno de los chips más avanzados. En cambio, depende de Asia Oriental para el 75% de la producción. Es clave comprender que su liderazgo se basa en activos intangibles, es decir, en la propiedad intelectual a través de las patentes. Esta posición explica su capacidad para imponer sanciones formales e influir en las decisiones de terceros países y empresas23. La mayor parte de la fabricación del hardware real se lleva a cabo fuera de su territorio. Es muy dependiente del exterior para la producción de los semiconductores lógicos de vanguardia que alimentan todos los algoritmos de IA fundamentales para los sistemas de defensa (NSCAI, 2021),

[21] La Administración está renunciando a la velocidad de interconexión (la velocidad a la que los chips se comunican entre sí) a cambio de la densidad de rendimiento como parámetro. Exceder los parámetros de rendimiento o el nuevo umbral de densidad de rendimiento claramente estaría dentro de la línea roja.
[22] Un pilar principal de esta parte de la norma es abordar los problemas relacionados con el uso por parte de China de empresas subsidiarias para obtener chips que de otro modo estarían controlados. Según las normas actualizadas, las exportaciones de chips estarán restringidas a las empresas con sede en China, Macao y países con los que Estados Unidos mantiene un embargo de armas. El efecto de esta actualización es que 43 países adicionales, que abarcan las designaciones de países D:1, D:4 y D:5, ahora estarán sujetos a requisitos de licencia ampliados.
[23] En agosto de 2022, el Departamento de Comercio de EE.UU. prohibió a Nvidia vender sus unidades de procesamiento gráfico (GPU por sus siglas en inglés) a clientes de China y desde noviembre 2023 no puede vender sus productos más avanzados.

lo cual explica muchas decisiones recientes de inversión, pero también de “reconstruir” su capacidad doméstica. Además, el país también vende la mayor cantidad de semiconductores específicos para el aprendizaje automático, las denominadas matrices de puertas programables en campo (FPGA en inglés) y las GPU que son esenciales para el entrenamiento de algoritmos.

La Ley de CHIPS y Ciencia de 2022 pretende revitalizar la industria nacional de semiconductores de Estados Unidos e invertir en la construcción de nuevas plantas de fabricación, incluidas las de empresas extranjeras como TSMC24 y Samsung25. La envergadura de la Ley CHIPS y de Ciencia se advierte por los 52.700 millones de dólares en subvenciones asignadas a la fabricación de semiconductores. Sin embargo, diversas iniciativas enfrentan problemas tanto por los costos como por la escasez de mano de obra (Nikkei Asia, 19/03/2024)26. El asunto tiene implicaciones y es interesante porque no se trata únicamente de la instalación de fundiciones (foundries), sino que se trata de un conjunto de firmas que proveen los insumos necesarios para el funcionamiento de dichas fundiciones. Esta situación de falta de recursos humanos debe ponderarse ya que el hecho de que varios proveedores hayan ralentizado sus proyectos demuestra que el problema no está causado por una o dos empresas, sino que es más estructural.

[24] Según ha sido reportado, la taiwanesa TSMC acordó con las autoridades norteamericanas que la subvención federal ascienda a 6.600 millones de dólares, comprometiéndose a aumentar su inversión a más de 65.000 millones de dólares y producir los chips de 2 nanómetros más avanzados del mundo en su futura planta en Arizona. La secretaria de Comercio de Estados Unidos, Gina Raimondo, dijo que TSMC también construirá una tercera planta de chips no anunciada previamente en Phoenix, Arizona, que estará operativa en 2030 (Nikkei Asia, 8/04/2024). Por otro lado, el Departamento de Comercio de Estados Unidos ofrecerá hasta 6.400 millones de dólares en financiación directa a Samsung Electronics para construir instalaciones de fabricación de chips en el centro de Texas, alcanzado la inversión total de Samsung hasta aproximadamente 45.000 millones de dólares (Nikkei Asia, 15/07/2024).

[25]. El 4 de marzo de 2025, TSMC anunció su intención de ampliar su inversión en la fabricación avanzada de semiconductores en Estados Unidos en 100.000 millones de dólares adicionales. Sobre la base de la inversión en curso de 65.000 millones de dólares de la empresa en sus operaciones de fabricación avanzada de semiconductores en Phoenix, Arizona, se espera que la inversión total de TSMC en Estados Unidos alcance los 165 000 millones de dólares. La expansión incluye planes para tres nuevas plantas de fabricación, dos instalaciones de empaquetado avanzado y un importante centro de I+D, lo que consolida este proyecto como la mayor inversión extranjera directa en la historia de EE. UU. TSMC espera crear cientos de miles de millones de dólares en valor de semiconductores para la IA y otras aplicaciones de vanguardia. Esta medida revela la dedicación de TSMC a apoyar a sus clientes, incluidas las principales empresas estadounidenses de inteligencia artificial e innovación tecnológica, como Apple, NVIDIA, AMD, Broadcom y Qualcomm.

[26] Tres ejecutivos de materiales para chips declararon a Nikkei Asia que el coste de construcción de una planta en Arizona era cuatro o cinco veces superior al de Asia y "varias veces" más elevado de lo que habían previsto.

Conclusiones

Del estudio de los chips para IA en defensa se concluye que existen importantes desafíos que se deben afrontar:

1. Los sistemas de IA, cada vez más, se van desplazando hacia el borde (Edge), donde los sensores recogen datos en tiempo real, y los sistemas deben tomar decisiones críticas, es decir, realizar la tarea de "inferencia IA" en el propio hardware en el borde, en tiempos muy cortos, lo que exige una gran capacidad de computación y ser eficientes energéticamente. Los chips, conocidos como "Aceleradores de IA" (ASICs), se utilizan cada vez más para tareas de inferencia específicas.
2. El desafío del **consumo energético** de los chips IA, tanto en la nube (GPUs), como en el borde es un tema capital al que se dedica mucho esfuerzo y que está produciendo avances como los chips neuromórficos (redes neuronales de impulsos) (SNN) , los chips AIMC (Analog in memory computing) que al evitar el trasiego de datos entre procesador y memoria ahorran una gran cantidad de energía. La fotónica integrada también entra para conseguir hardware IA más eficiente.
3. El desarrollo de sensores del tipo de neuronas sensoriales artificiales puede ser de gran interés en el mundo de la seguridad.
4. En los sistemas militares para la detección de objetivos, va a ser de mucho interés los sistemas de encapsulado utilizando integración heterogénea donde puede interconectarse de forma eficiente MEMs, como sensores o antenas, receptores/emisores para radar en chips de GaN y procesadores de IA, tanto para entrenamiento como para inferencia en un único encapsulado. Igualmente es de gran interés para acortar tiempos de diseño y poder hacer cierta reutilización el uso de chiplets en SiP.
5. Los chips IA tienen un gran valor industrial y una posición estratégica importante.
6. Sin chips de IA en el borde no hay posibilidad de aplicar los algoritmos de IA para las aplicaciones de Defensa en tiempo real, por tanto, hay que asegurar el suministro y, mucho más deseable, **tener capacidad de diseñarlos y producirlos**.
7. Europa no tiene FPGAs, una de las soluciones para chips IA, que equilibran flexibilidad y eficiencia energética y que no requieren tiempos largos de diseño.

Recomendaciones

1. Establecer una estrategia nacional de semiconductores, alineada con el Chips Act europeo, en el que participe con un peso específico importante Defensa, la industria de Defensa española y representantes de las asociaciones industriales de electrónica, telecomunicaciones, IA, así como las AAPP y las universidades.
2. Establecer una Política de compra pública innovadora para aplicaciones de IA duales (Defensa y civil), desarrollando proyectos con ambición, en los que participara el ecosistema microelectrónico español.
3. Establecer programas público-privados de desarrollo de la fotónica integrada para IA. Para ello contar con la línea piloto de fotónica integrada que lidera el IFCO, junto al cluster de Valencia y SPARC en Vigo.
4. Revisar todos los programas de Cátedras chip para formar personal en chips de IA para la defensa, especialmente chips neuromórficos y AiMC. Es importante contar con el conocimiento en diseño de señal mixta analógico/digital.

5. Estudiar la posibilidad del Factory Academy para tener una vía de acceso a los procesos tecnológicos de microelectrónica.
6. En la Estrategia nacional de semiconductores plantear la posibilidad de una factoría con fabricación híbrida GaN / Silicio para automoción. (dual). Esta factoría debería contar con capital público y capital privado, pero de mayoría nacional por intereses estratégicos. Puede ser una factoría "llave en mano" y controlada por una comisión formada por defensa, la industria de defensa, la industria de automoción y los inversores privados. Atraer una factoría asiática como Samsung, TSMC, etc. no funcionará aunque se le entreguen importantes fondos como transferencia, puesto que los costes operativos son importantes y si no está asegurado el funcionamiento de ésta a un mínimo del 80 % de su capacidad.

Bibliografía

Acemoglou, D. (May 2024). THE SIMPLE MACROECONOMICS OF AI. *National Bureau of Economic Research Working Paper 32487*.

Allen, G. C. (2023, July 31). *Blocking China's Access to AI Chips Matters to U.S. National Security.* Retrieved from Center for Strategic & International Studies: https://www.csis.org/analysis/blocking-chinas-access-ai-chips-matters-us-national-security

Baietto, A., & Bihl, T. (2025). Generative Data for Neuromorphic Computing. *Proceedings of the 58th Hawaii International Conference on System Sciences | 2025*, (págs. 7248-7258).

Baruzzi, V., Indiveri, G., & Sabatini, S. (January 2025). Recurrent models of orientation selectivity enable robust early-vision processing in mixed-signal neuromorphic hardware. *Nature Communications*, 1-15.

Batra, G., Jacobson, Z., Madhav, S., Queirolo, A., & Santhanam, N. (December 2018). *Artificial-Intelligence Hardware: New opportunities for semiconductor companies.* Sidney: Mc Kinsey & Company.

Bourzac, K. (January 2025). The U.S. will start manufacturing advanced chips. *Spectrum IEEE*, 40-41.

Burkacky, O., Patel, M., Pototzky, K., Tang, D., Vrijen, R., & Zhu, W. (March 2024). *Generative AI: The next S-curve for the semiconductor industry?* Mc Kinsey & Company.

Channamadhavuni, S., Thijssen, S., Jha, S., & Ewetz, R. (June 22-25, 2021). Accelerating AI Applications using Analog In-Memory Computing: Challenges and Opportunities. *GLSVLSI '21, Virtual Event, USA* .

Evanson, N. (2025, January 2). *What Are Chiplets and Why They Are So Important for the Future of Processors.* Retrieved from TECHSPOT: https://www.techspot.com/article/2678-chiplets-explained/

Fogarty, K. (2019, August 15th). *GaN versus Silicon for 5G.* Retrieved from Semiconductor Engineering: https://semiengineering.com/gan-versus-silicon-for-5g/#:~:text=%E2%80%9CMost%20of%20the%20difference%20is,IC%20products%20at%20Analog%20Devices.

Guo, Z., Xue, X., Zhou, Q., & Zeng, X. (2024). Computing in Memory for Accelerating Light-Weighted On-Chip Learning in IoT Devices. *2024 IEEE 17th International Conference on Solid-State & Integrated Circuit Technology (ICSICT) |*.

Gupta, S., & Jolly, X. (2024). Neuromorphic Photonic On-chip Computing. *Preprints.org (www.preprints.org) |*.

IRDS. (2023). *IRDS More than Moore.* IEEE.

IRDS BC. (2023). *International Roadmap Devices and Systems (IRDS) Beyond CMOS.* IEEE.

IRDS MM. (2023). *IRDS (International Roadmap Devices and Systems) More Moore.* IEEE.

IRDS MtM. (2023). *International Roadmap Devices and Systems (IRDS) More than Moore.* IEEE.

Isik, M. (2023). A Survey of Spiking Neural Network Accelerator on FPGA. *Journal of Latex Class, IEEE*, 1-15.

Jeroen Kusters, D. B. (2025, February 04). *Deloitte Center for Technology, Media & Telecommunications.* Retrieved from 2025 global semiconductor outlook: https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-telecom-outlooks/semiconductor-industry-outlook.html

Kelleher, A. (February de 2022). *Moore's Law - Now and in the future.* Obtenido de INTEL: https://download.intel.com/newsroom/2022/manufacturing/Intel-Moores-Law-Investor-Meeting-Paper-final.pdf

Khaki, A. M., & Choi, A. (2025). Optimizing Deep Learning Acceleration on FPGA for Real-Time and Resource-Efficient Image Classification. *Applied Sciences MDPI*, 1-13.

Khan, S. M., & Mann, A. (April 2020). *AI Chips: What they are and why they matter.* CSET (Center for Security and Emergency Technology.

Lacey, G., Taylor, G., & Areibi, S. (2016). Deep Learning on FPGAs: Past, Present and Future. *researchegate.net*.

Li, B., Gu, J., & Jiang, W. (2019). Artificial Intelligence (AI) chip technology review. *International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 114-117). Taiyuan (China): IEEE.

Li, C., Lammie, C., Le Gallo, M., & Rajendran, B. (Dec 2024). Efficient Deployment of Transformer Models in Analog In-Memory Computing Hardware. *arXiv:2411.17367* .

Li, D. (2023). Analysis the principle of AI Chip principle and the State-of-the-art applications. *Highligths in Science, Engineering and Technology*, Volume 76 (2023).

Li, Z., Zhang, Y., & Wang, J. (2020). A survey of FPGA design for AI era. *Journal of Semiconductors*.

Meng, Y., Xudong, Z., Jianwen, Z., Xinxin, X., Changling, W., & Fang, W. (2023). A Ultra-Low Power System Design Method of AI Edge Computation. *2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*.

Moore, G. E. (April 1965). Craming more components onto integrated circuits. *Electronics*.

Ondrej Burbacky, J. D. (2022, April). *The semiconductor decade: A trillion-dollar industry.* Retrieved from Mc Kinsey & Company: https://www.mckinsey.com/~/media/mckinsey/industries/semiconductors/our%20insights/the%20semiconductor%20decade%20a%20trillion%20dollar%20industry/the-semiconductor-decade-a-trillion-dollar-industry-v3.pdf?shouldIndex=false

Pang, G. (March-April 2022). The AI Chip Race. *IEEE Intelligent Systems*, 111-112.

Shastri, B., Talt, A., Ferreira de Lima, T., Pernice, W., Bhaskaran, H., Wright, C., & Prucnal, P. (February 2021). Photonics for artificial intelligence and neuromorphic computing. *Nature Photonics*, 102-114.

SIA (Semiconductor Industry Association) & BCG (Boston Consulting Group). (2025, febrero 28). *Strengthing the U.S. Semiconductor Supply Chain .* Retrieved from Semiconductors Industry Association (SIA): https://www.semiconductors.org/

Sipola, T., Alatalo, J., Kokkonen, T., & Rantonen, M. (April 2022). Artificial Intelligence in the IoT Era: A Review of Edge AI Hardware and Software. *Proceedings of the 31st Conference of Open Innovations Association FRUCT & IEEE*, 320-331.

Smith, M. S. (October 2024). Challenges are coming for Nvidia's crown. *SPECTRUM IEEE* , 40-44.

Van Maarseveen, H. (Julio 2023). *XILINX Vitis: AI A journey through the evolution of AI Hardware.* Kindle e-book.

Young, A., Dean, M., Plank, J., & Rose, G. (2019). A Review of Spiking Neuromorphic Hardware. *IEEE Access*, 135606-135620.

Zhenling, S., Qi, L., KANEKO, H., Li, H., & Meng, L. (June 2024). Optimization and Deployment of DNNs for RISC-V-based Edge AI. *Proceedings of The 2024 IEEE International Conference on Real-time Computing and Robotics*, 200-205.

Zhong, S., Su, L., Mingkun Xu, M., Loke, D., Yu, B., & Yishu Zhang, Y. (2025). Recent Advances in Artificial Sensory Neurons: Biological Fundamentals, Devices, Applications and Challenges. *Micro-Nano Letters*, 1-49.