Tendencias en el Diseño de Procesadores con Arquitectura ARM

Javier Diaz Bruguera
Senior Principal Microarchitect, Contractor
Arm Cambridge, UK

UCM, Noviembre 2022

¡El Procesador ha cumplido 50 años!

Electronics News Weekly, Noviembre 1971

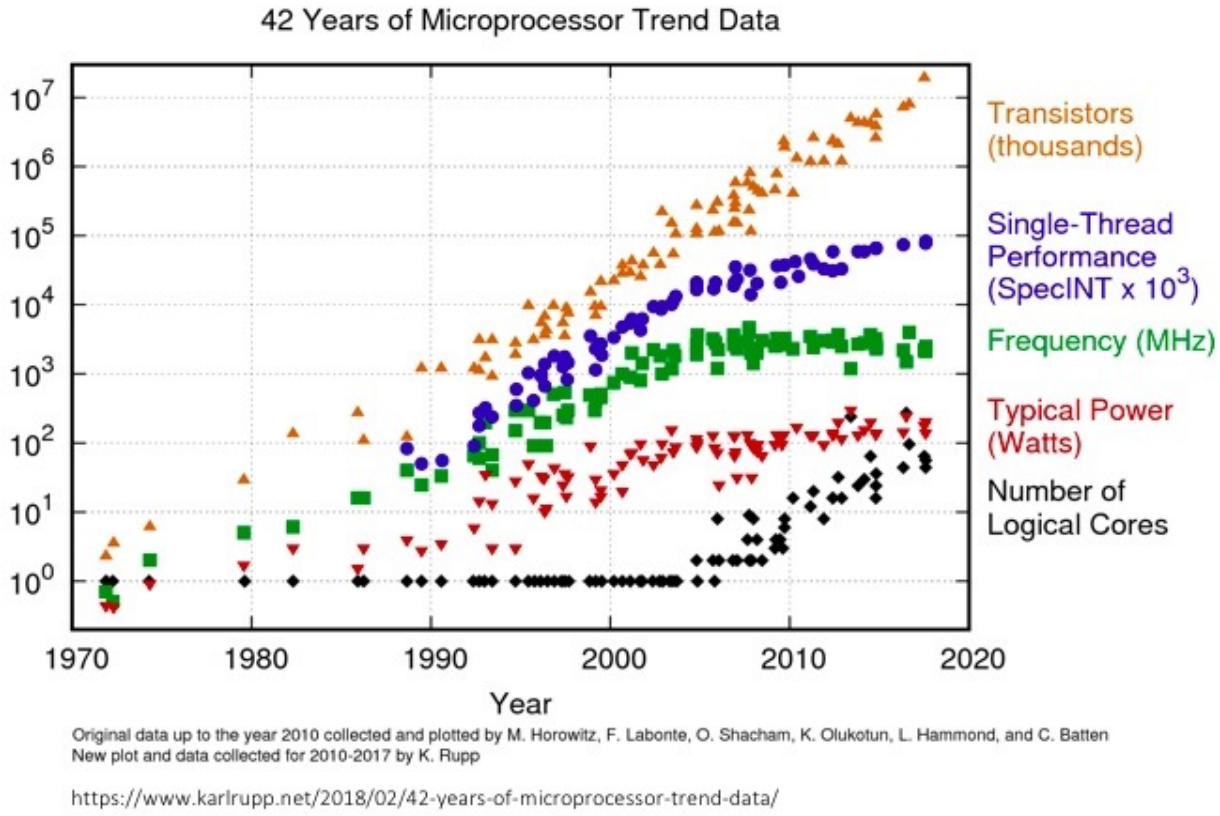
“Microprocessors are getting into everything. We won’t be able to pick up a single piece of equipment in the near future, except maybe a broom that hasn’t got a microprocessor in it.”

Arthur Clarke, 1979

Una evolución vertiginosa

- Posible debido a los avances en
  - Arquitectura
  - Microarquitectura
  - Tecnología
  - Fabricación de CIs

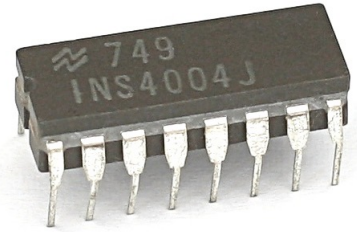
	Intel 4004 (1971)	Apple M1 (2022)
Frecuencia	750 KHz	3.3 GHz
Transistores	2.250	16.000 millones
Tecnología	10 μm	5 nm
Area	12 mm^2	120 mm^2
Datos	4 bits	64 bits



- Límite en la frecuencia (GHz) **debido a la enorme cantidad de calor que se genera en el procesador y al consumo de potencia**

Los procesadores han cambiado

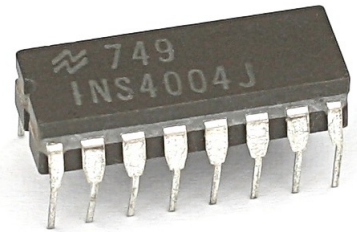
Chip del Intel 4004 de
National Semiconductor



Bus de 4 bits multiplexado
- direcciones de 12 bits
- datos de 4 bits
- instrucciones de 8 bits

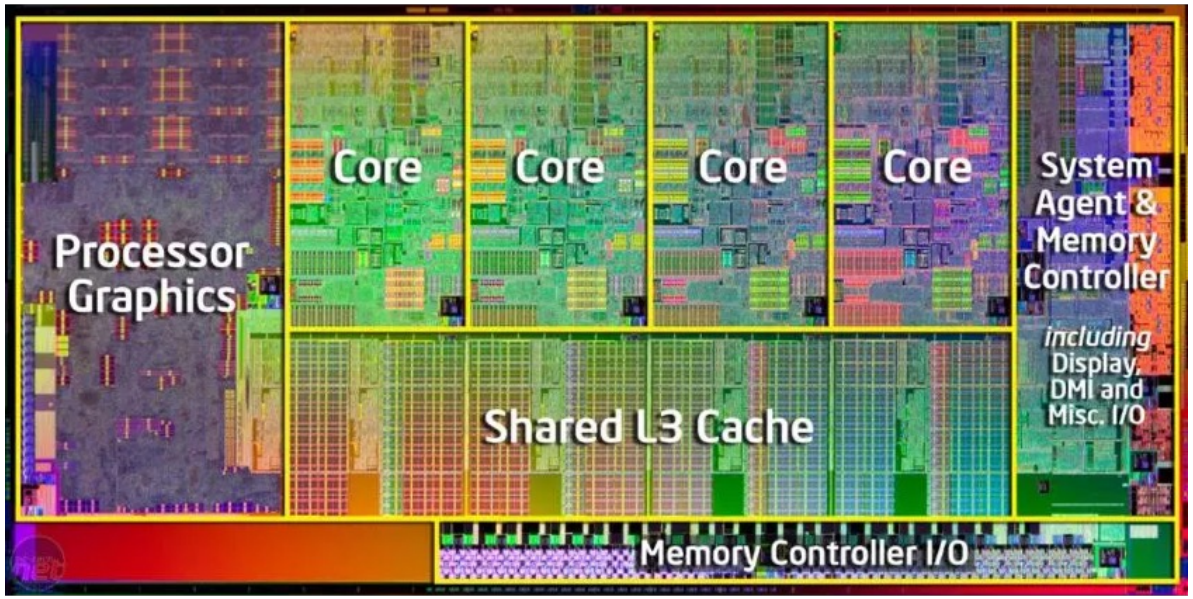
Los procesadores han cambiado

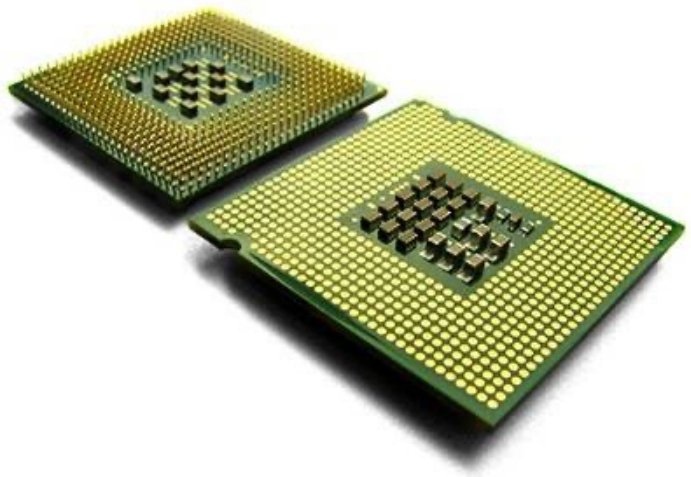
Chip del Intel 4004 de
National Semiconductor



Bus de 4 bits multiplexado
- direcciones de 12 bits
- datos de 4 bits
- instrucciones de 8 bits

Sistemas heterogéneos en un chip (SoC)



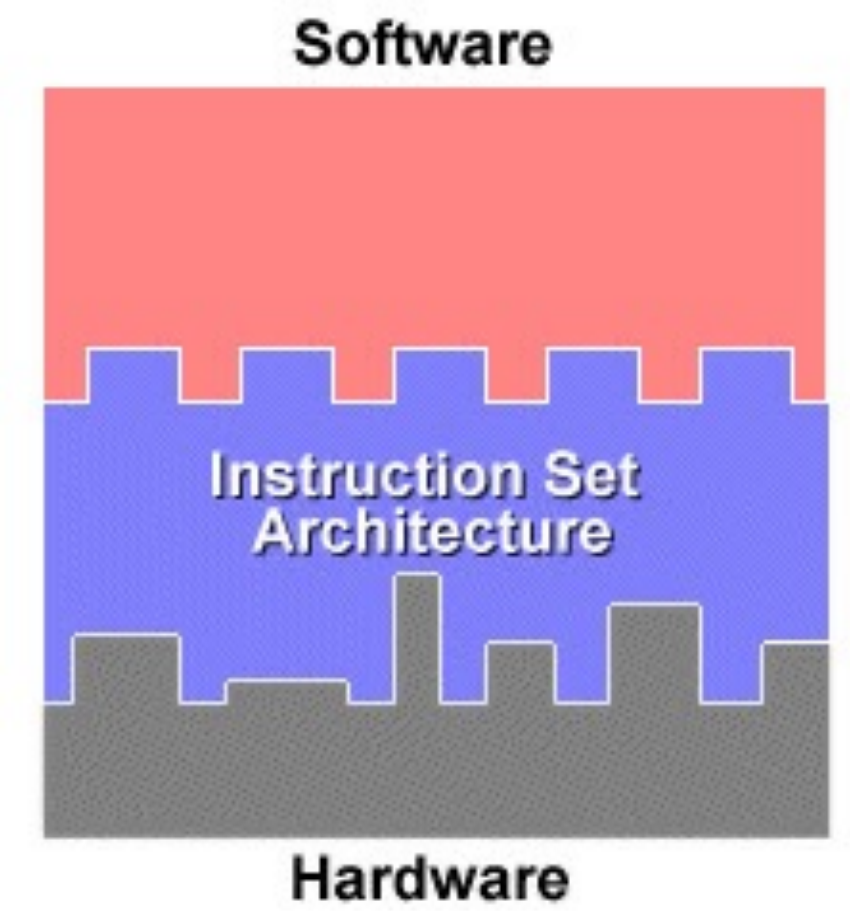


El número de pines
de I/O es > 1000

Arquitectura vs. microarquitectura

La arquitectura del procesador

- Imagen abstracta del sistema como sería visto por un programador en lenguaje ensamblador
- Es muy difícil introducir cambios rompedores en la arquitectura de un procesador
  - Se debe mantener la compatibilidad con versiones anteriores de la arquitectura
- Las arquitecturas suelen evolucionar de forma *suave*
  - Transición de 32 a 64 bits
  - Extensiones vectoriales
  - Protección y seguridad
  - …

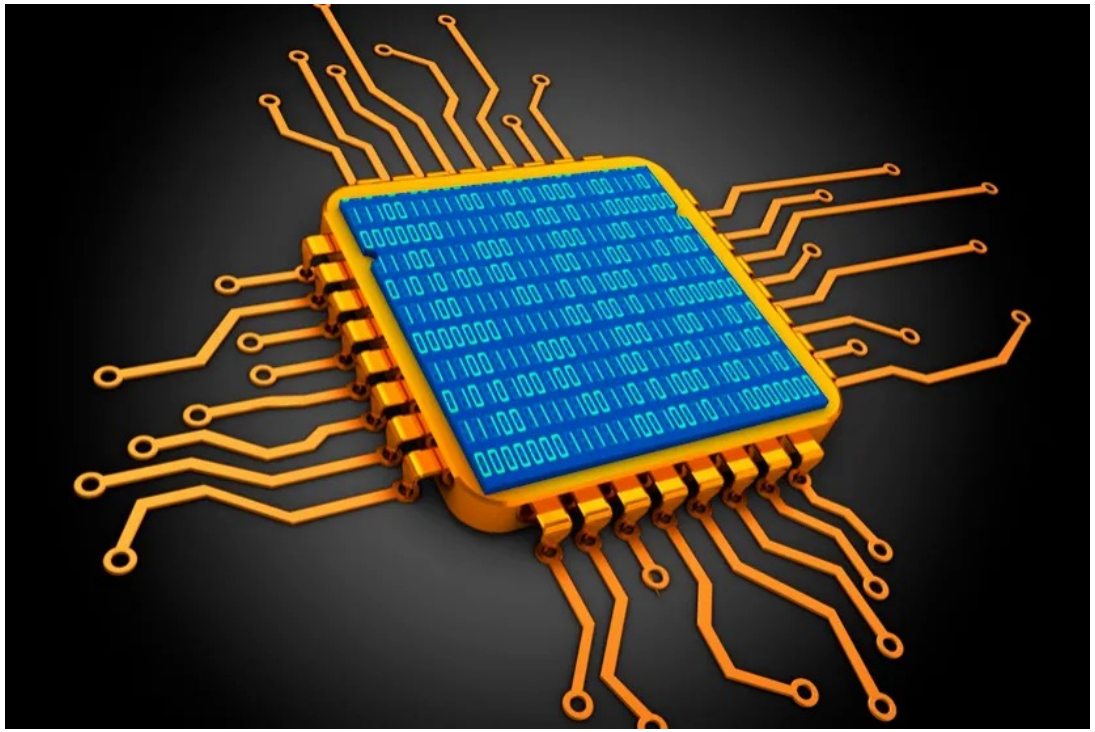


Microarquitectura del procesador

- Implementación de una *arquitectura* en un procesador
  - Una arquitectura dada puede ser implementada con diferentes microarquitecturas
  - Las implementaciones pueden variar debido a diferentes objetivos de diseño
- La microarquitectura ha evolucionado (radicalmente) para mejorar las prestaciones
- Evolución posible gracias a los avances tecnológicos y de fabricación de circuitos integrados
  - Mayor densidad: más transistores en el mismo área de silicio

Las arquitecturas más destacadas

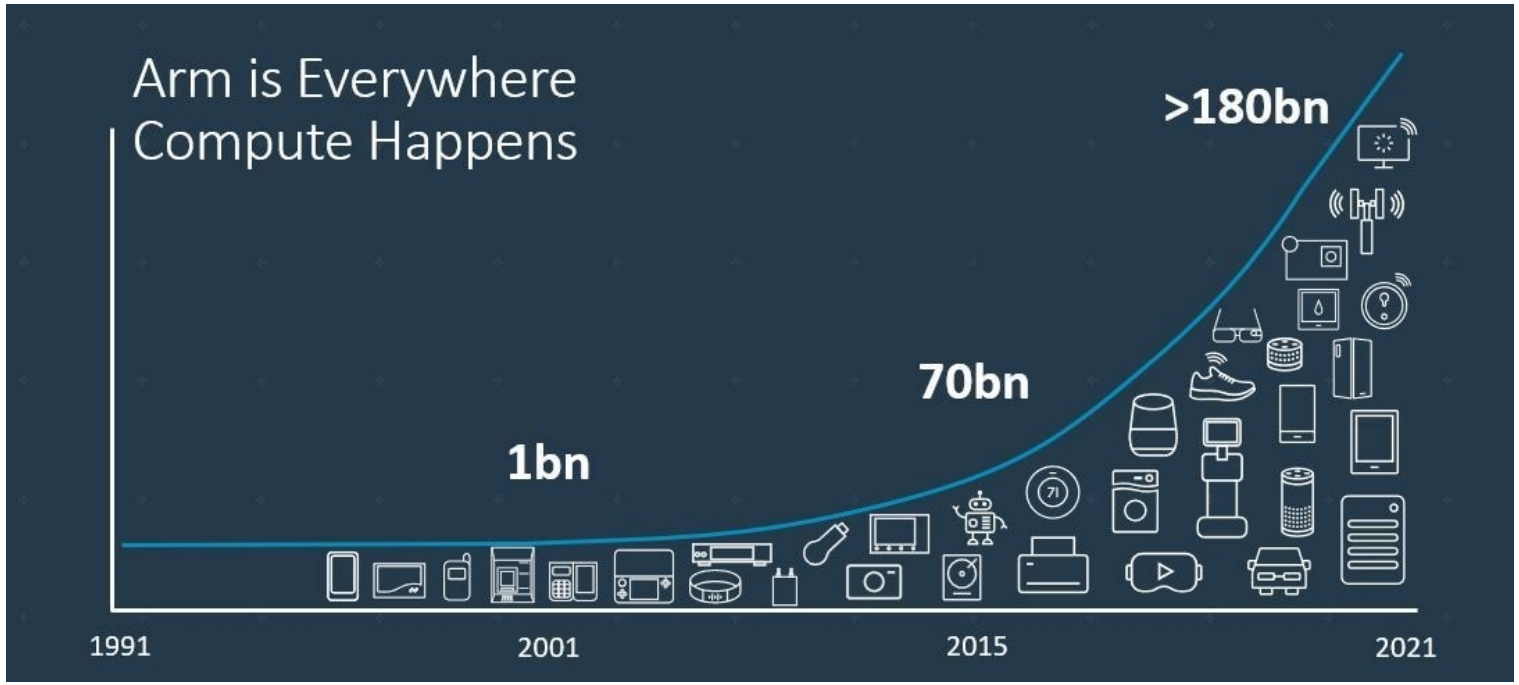
- Tres modelos de uso diferentes
- **x86**
  - Procesadores de Intel y AMD
- **ARM**
  - Se licencia a los *partners* para que la implementen en sus procesadores
  - Qualcomm, Apple, Samsung, MediaTek, NXP, Nvidia, AWS, Google, …
- **RISC-V**
  - Open source
  - SiFive, MIPS, Imagination, …



Licencias ARM

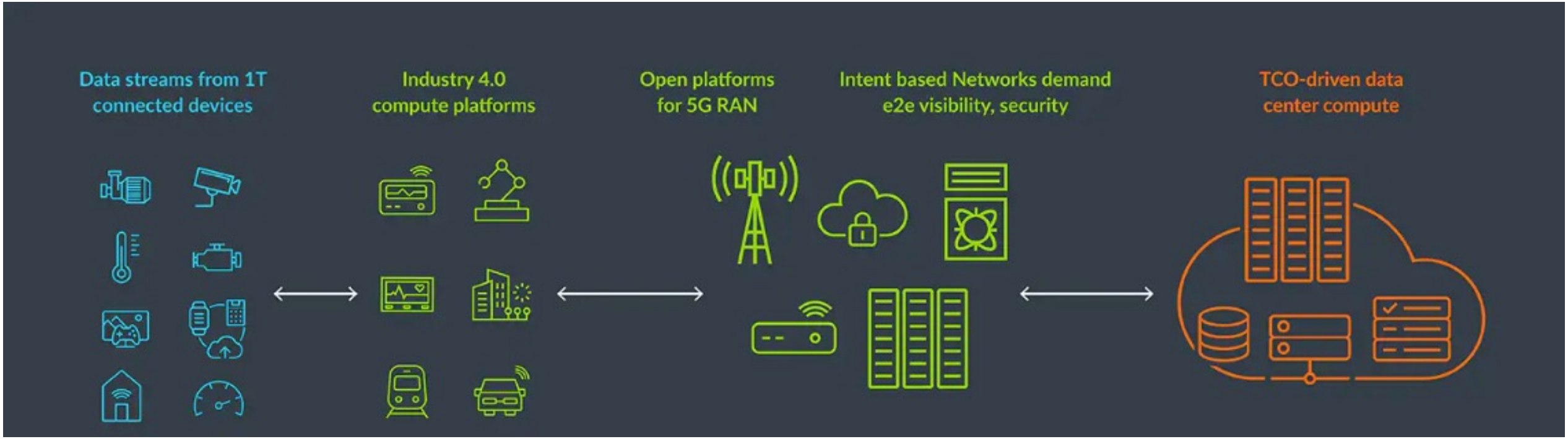
- El núcleo del negocio es la venta (licencia) de **IPs** y los *royalties* de los chips producidos
- **Arquitectura**: El *partner* diseña su propio procesador con arquitectura Arm
- **Microarquitectura**: el *core* ARM se integra en el diseño del *partner*
- ***Cortex CPU***
  - ***Neoverse***: HPC, Cloud, Infraestructura
  - ***Cortex-X, Cortex-A***: De aplicaciones para sistemas computacionalmente complejos
    - Cortex-X: Alto rendimiento
    - Cortex-A: Eficiencia energética (big, little)
- ***Cortex-M***: Procesadores de bajo consumo de potencia para microcontroladores y IOT
- ***Cortex-R***: Procesadores de tiempo real para aplicaciones críticas
- **Flexible Access**: acceso ilimitado a un número elevado de IPs, herramientas y formación de Arm para desarrollo de productos

Licencias ARM



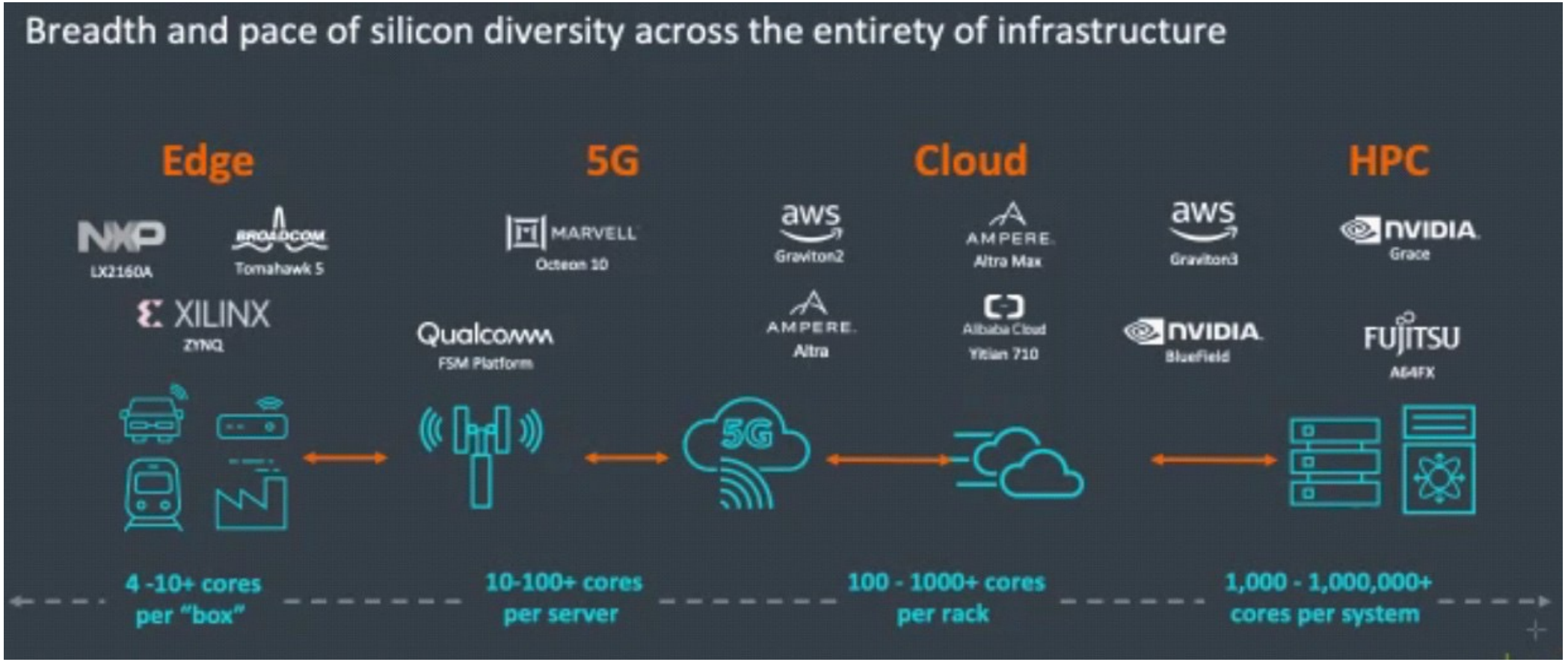
En Octubre de 2021 se alcanzaron los 200.000 millones de chips con tecnología Arm acumulados en 30 años

Computación cloud-edge-endpoint



- Plataformas muy diferentes
- La arquitectura ARM es central en la tecnología

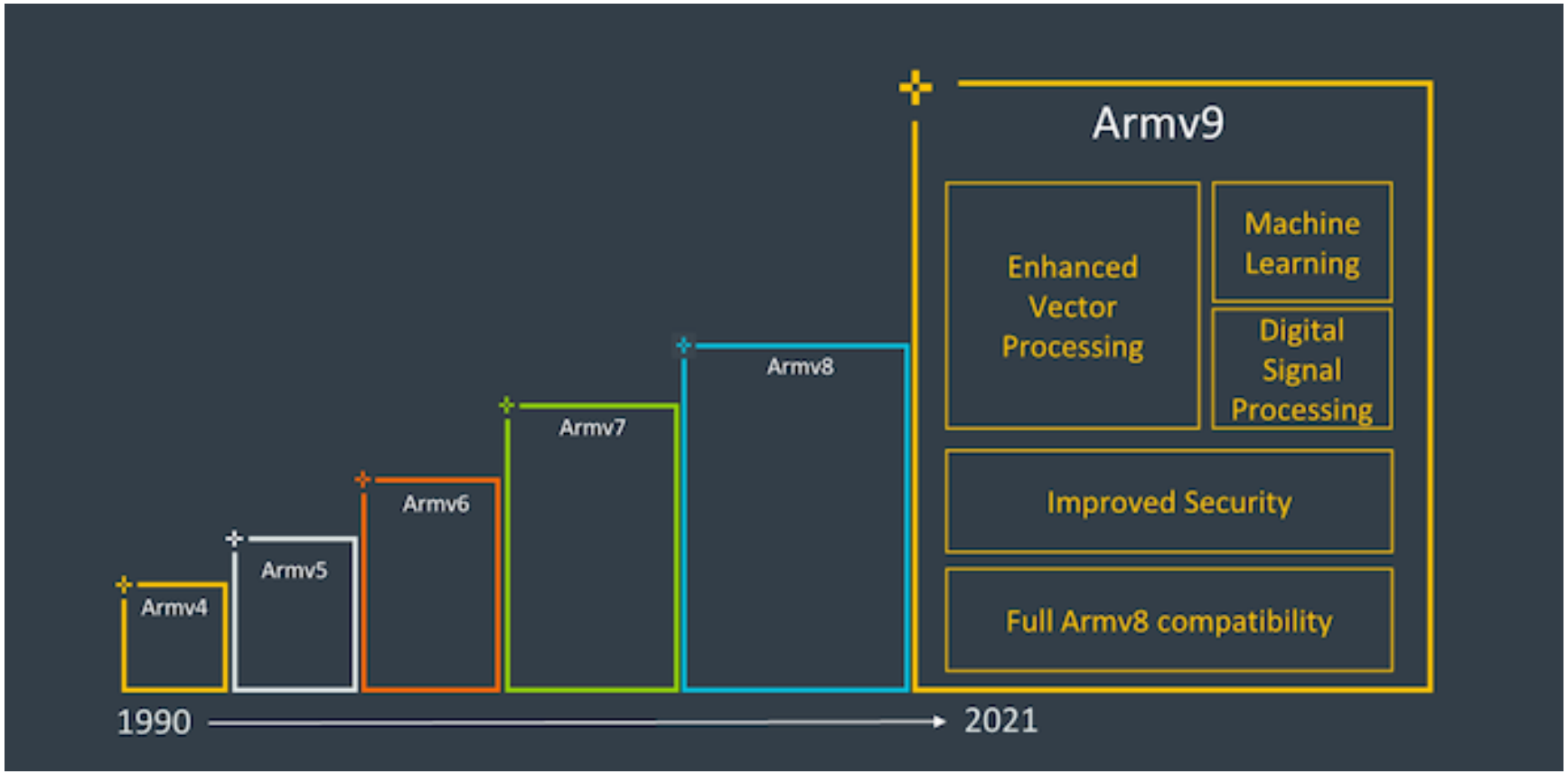
Computación cloud-edge-endpoint



When we look today, I think the trend is that everything is moving to Arm.
Cristiano Amon, CEO of Qualcomm, 2022

La Arquitectura ARM

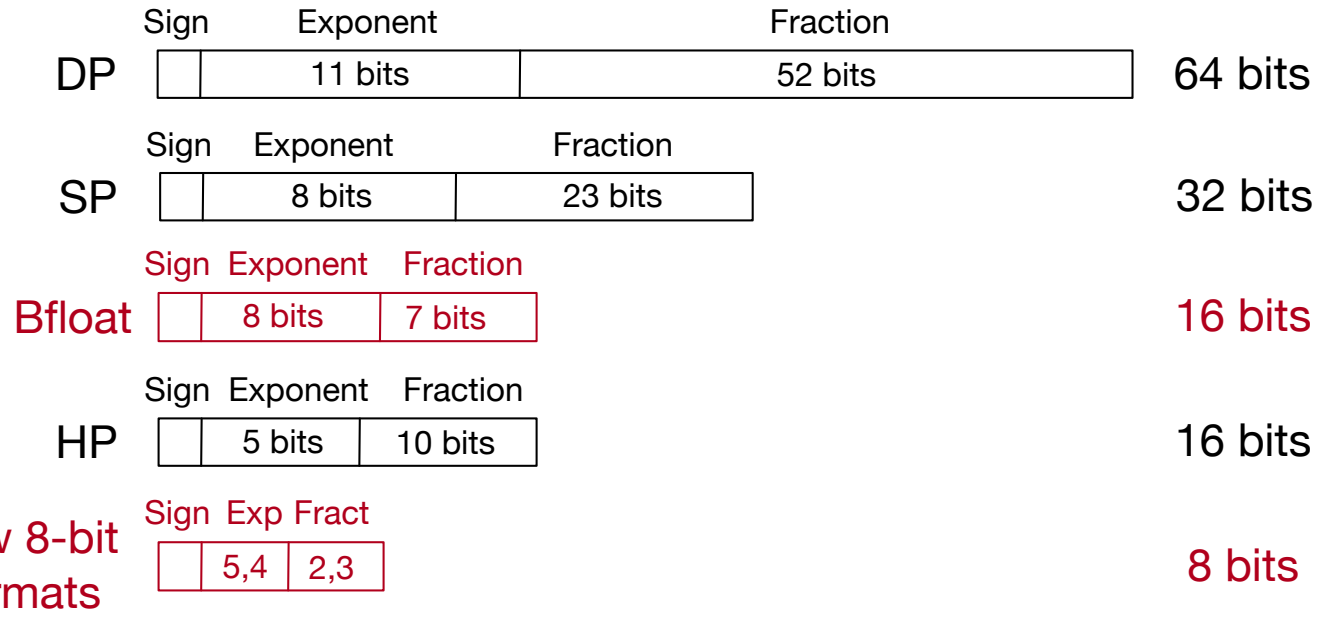
Armv9: Machine learning, SVE y seguridad



- Tres pilares: Inteligencia artificial (*Machine Learning*), mejoras en extensiones vectoriales, y seguridad

Armv9: Machine learning (ML)

- Las cargas de ML son cada vez más importantes
- Cientos de miles de gigabytes de datos se generan cada día con aplicaciones de IA
  - Electrónica de consumo, salud, logística, fabricación inteligente, automoción, …
- Aceleradores de ML, pero algunas tareas de ML se ejecutan en CPU
- Instrucciones de multiplicación de matrices 2x2, dot product (8 y 16 bits)
- Nuevos formatos de datos: bfloat
- En desarrollo
  - Punto flotante de 8 bits
  - Nuevas extensions de multiplicación de matrices

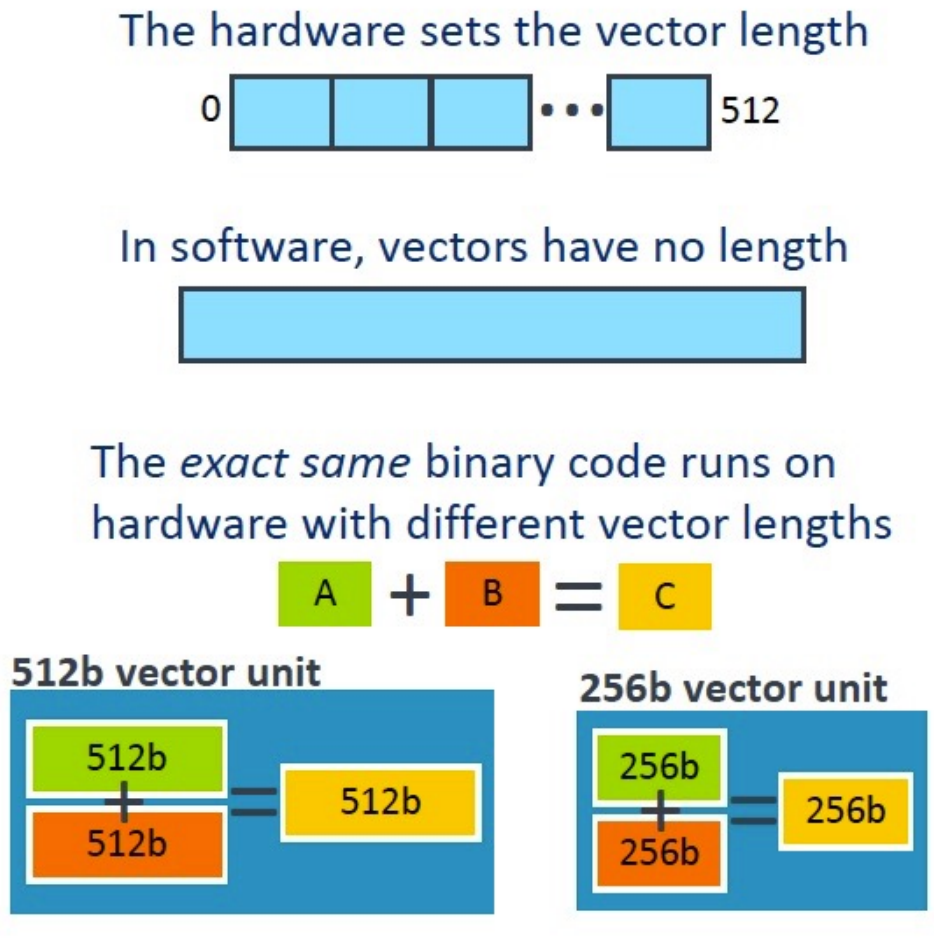


Armv9: scalable vector extensions (SVE/SVE2)

- Sucesor de *NEON*
- Anunciado en 2016 e implementado por primera vez en Fujitsu A64FX CPU (2018)
- La primera version *SVE* muy orientada a HPC
- Menos eficiente que *NEON* en otras cargas
- *SVE2* anunciado en 2019
- Se añadieron instrucciones para DSP, vision por computador, criptografía, comunicaciones, etc.

Scalable vector extensions (SVE/SVE2)

- Vectores de longitud variable: de 128 a 2048 bits, en incrementos de 128 bits
- Foco en la arquitectura
  - Independiente de la longitud de vector del hardware
  - El *partner* elige la implementación hardware
- Vector-length agnostic (VLA) programming model: El programador no necesita saber la longitud del vector
  - Código de operación único para cada instrucción
  - Longitud del vector en un registro de la arquitectura
  - El código solo necesita compilarse una vez
  - El mismo binario puede ejecutarse en diferentes núcleos, con diferentes SIMD pipelines
- Vectores con elementos de 64, 32, 16 y 8 bits



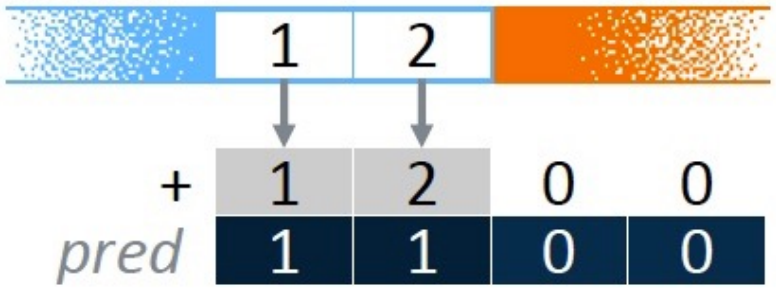
SVE: vector-length agnostic (VLA)

- Predicación por elementos: La operación se hace en elementos individuales bajo control del registro de predicado

```
for (i = 0; i < n; ++i)
```

INDEX i	n-2	n-1	n	n+1
WHILELT n	1	1	0	0

- Control de lazos con predicación: eliminación de control del lazo

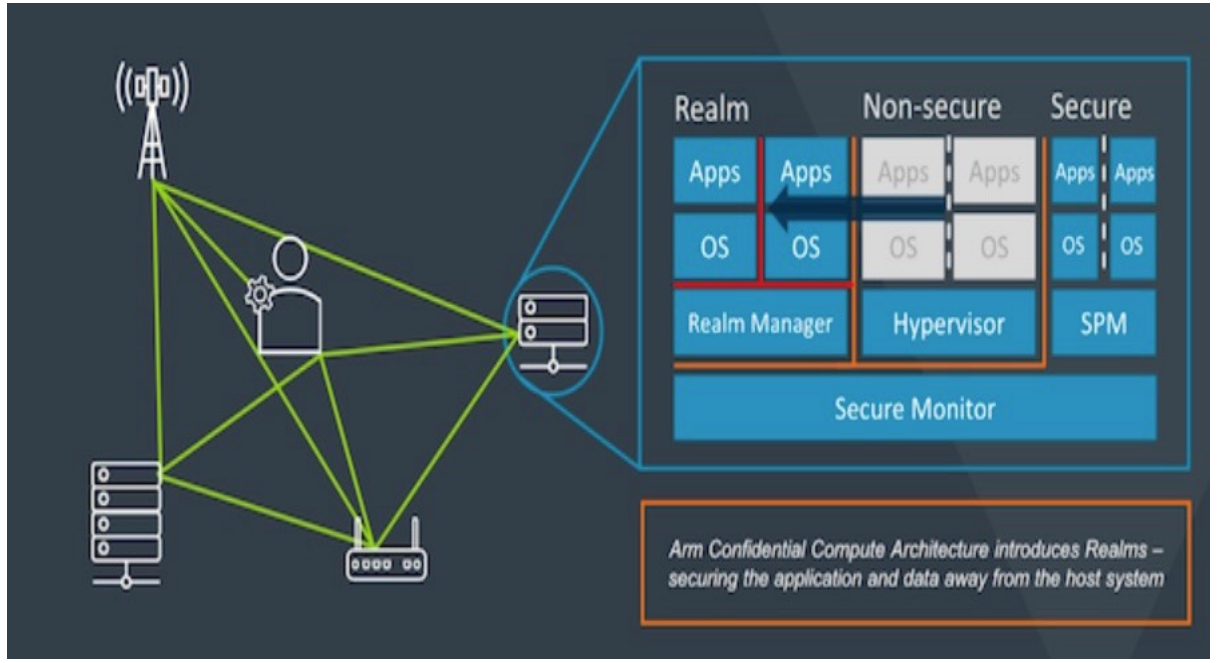


- Partición de vectores

Seguridad: confidential compute architecture (CCA)

- Varias brechas de seguridad en los últimos años (Spectre, Meltdown)
- Rediseño de mecanismos de seguridad

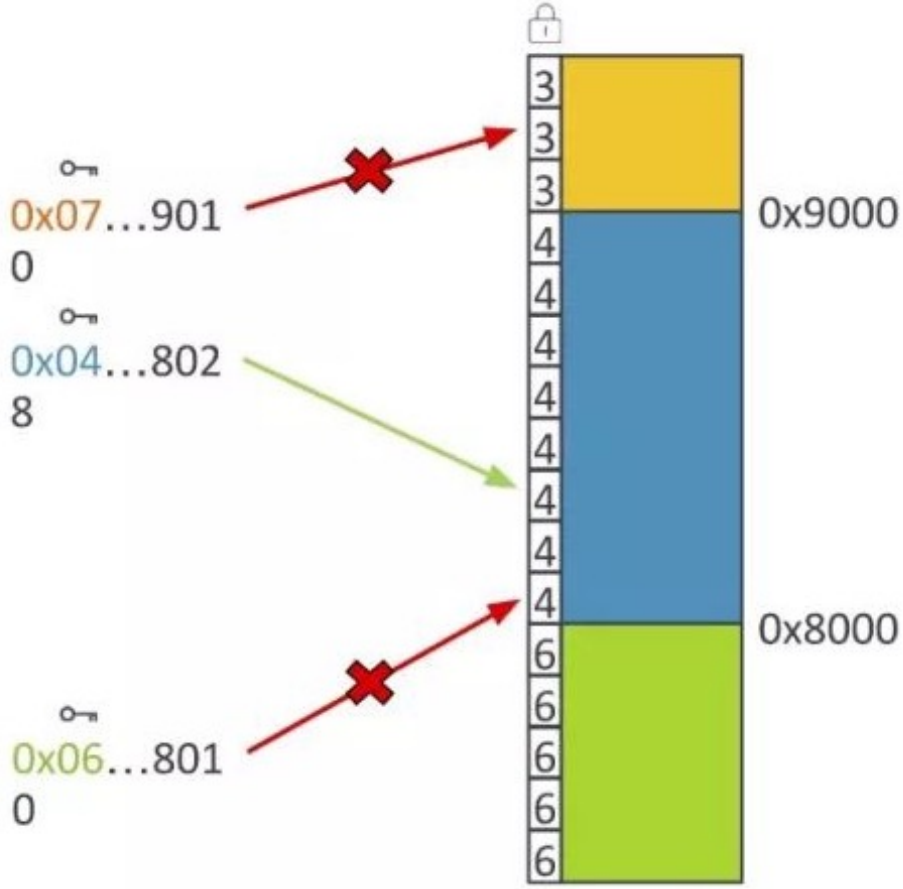
- Modelo tradicional de seguridad:
  - Las aplicaciones confían en el SO y el hipervisor
  - Software con mayores privilegios pueden *intervenir* en la ejecución de software con menores privilegios
  - Brecha de seguridad
- **Arm Confidencial Compute (CCA)**
  - Introduce los "*reinos*" (realms)
  - Contenedores seguros de entornos de ejecución
  - Opacos al SO y el hipervisor
  - Se crean de forma dinámica
  - Area de memoria con protección hardware
  - Aislados del resto del sistema



- Los reinos reducen la interferencia de otros elementos del SO
- Aplicaciones críticas en seguridad no requieren dispositivos específicos

Seguridad: memory tagging extension (MTE)

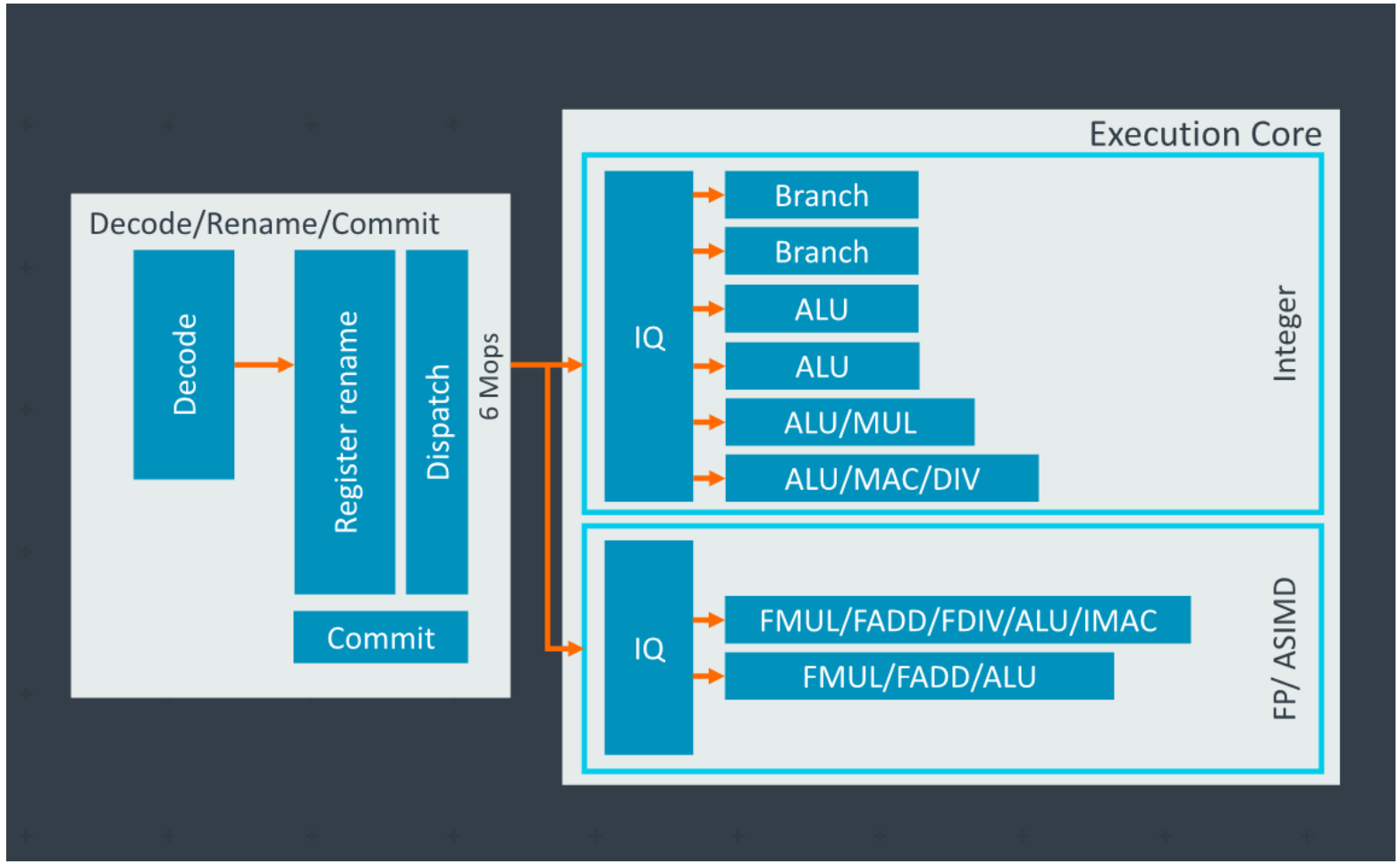
- Protección contra la *corrupción* de la memoria
- MTE: añadir un conjunto de bits a grupos de posiciones de memoria para identificarlas como seguras
  - *Address tagging (key):* 4 bits en cada puntero
  - *Memory tagging (lock):* 4 bits en cada región de 16 bytes de la memoria de aplicaciones
- Comparación de *key* y *lock*
- Las etiquetas de bloques de memoria contiguos son diferentes
- Gestionado en hardware: modificación de las instrucciones de *load* y *store*
- Implementado en la cache



La microarquitectura de los procesadores ARM Cortex CPU

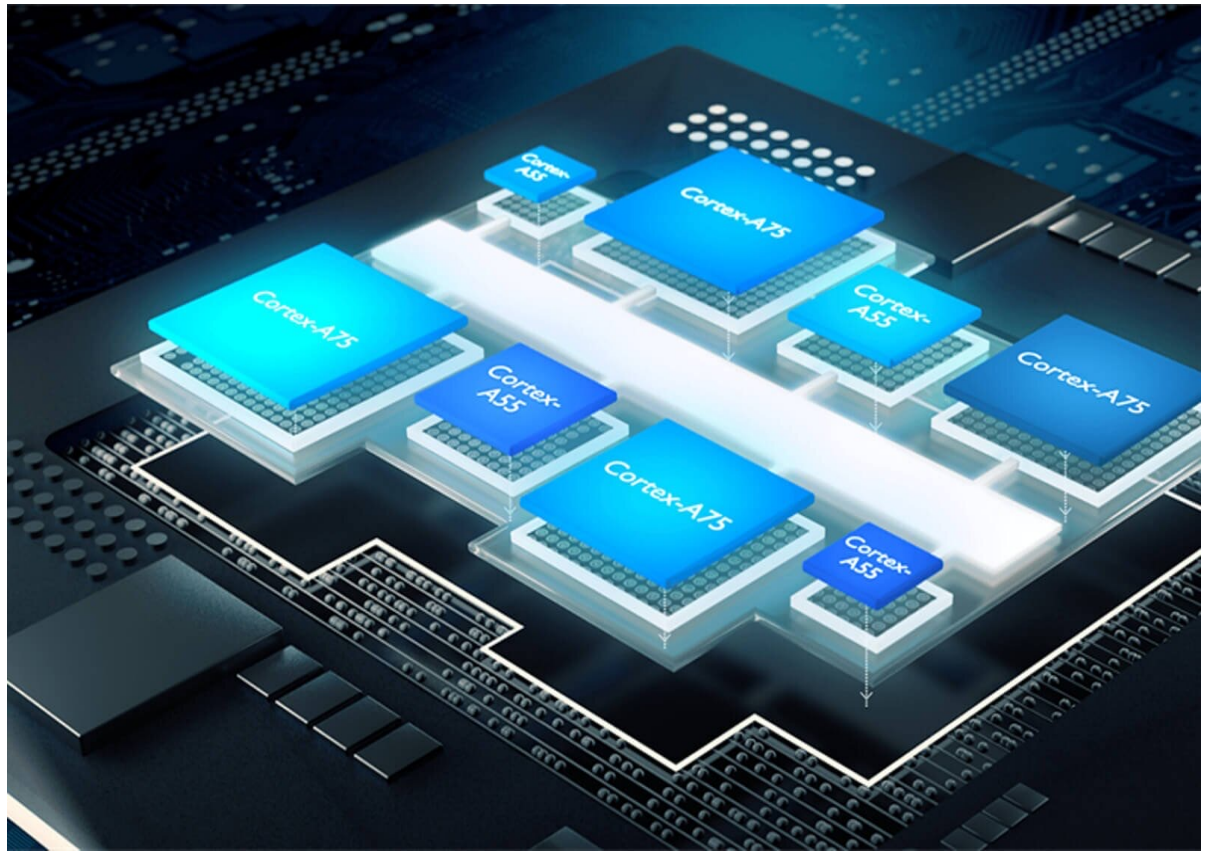
ARM Cortex CPU

- ***Neoverse***: HPC, Cloud, Infraestructura
- ***Cortex-X, Cortex-A***: De aplicaciones para sistemas computacionalmente complejos
  - Cortex-X: Alto rendimiento
  - Cortex-A: Eficiencia energética

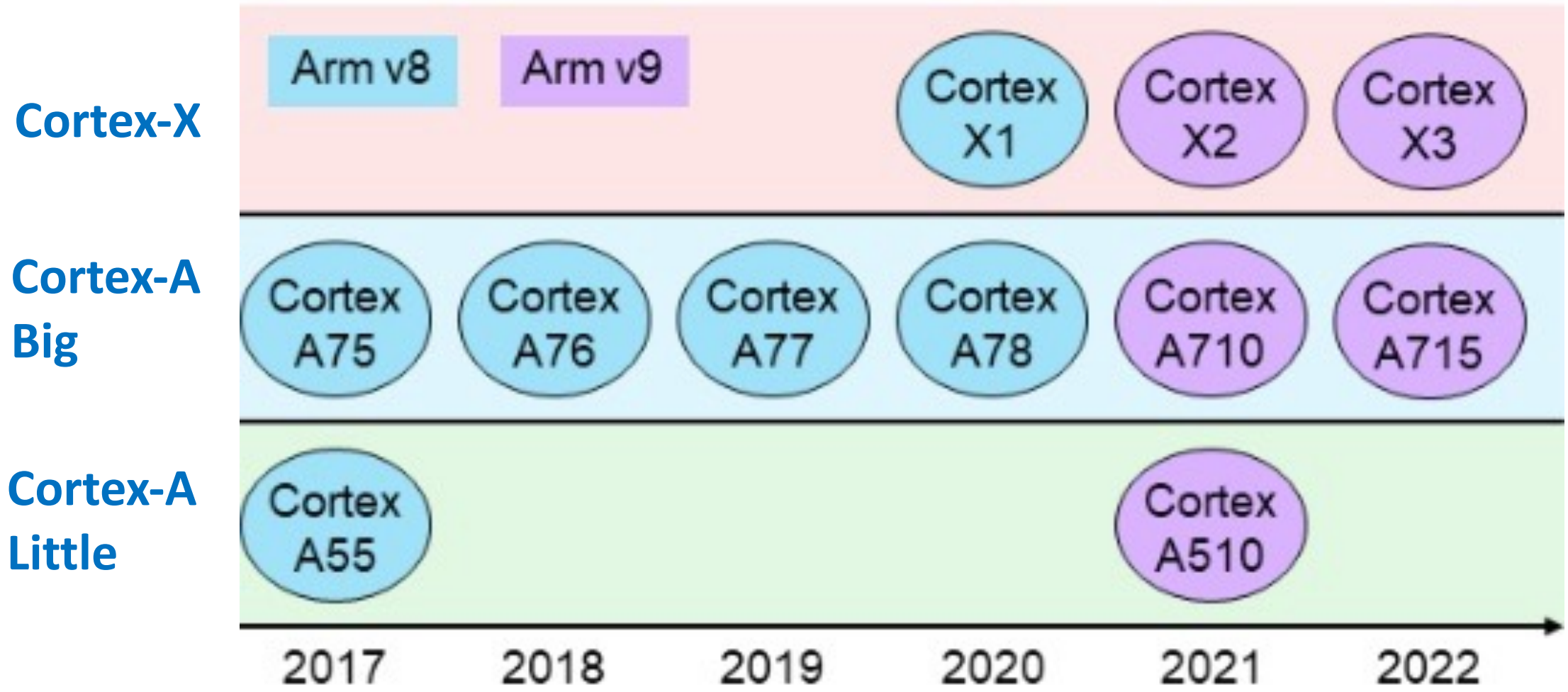


Cortex-A y Cortex-X en smartphones y tablets

- Procesadores multinúcleo heterogéneos
  - Combinación de núcleos de diferentes características para mejorar la eficiencia energética
- Configuración Arm *big.LITTLE*
  - Dos tipos de núcleos en el chip
  - *Big* para maximizar rendimiento
  - *Little* para reducir el consumo de potencia
- Configuracion Tri-cluster

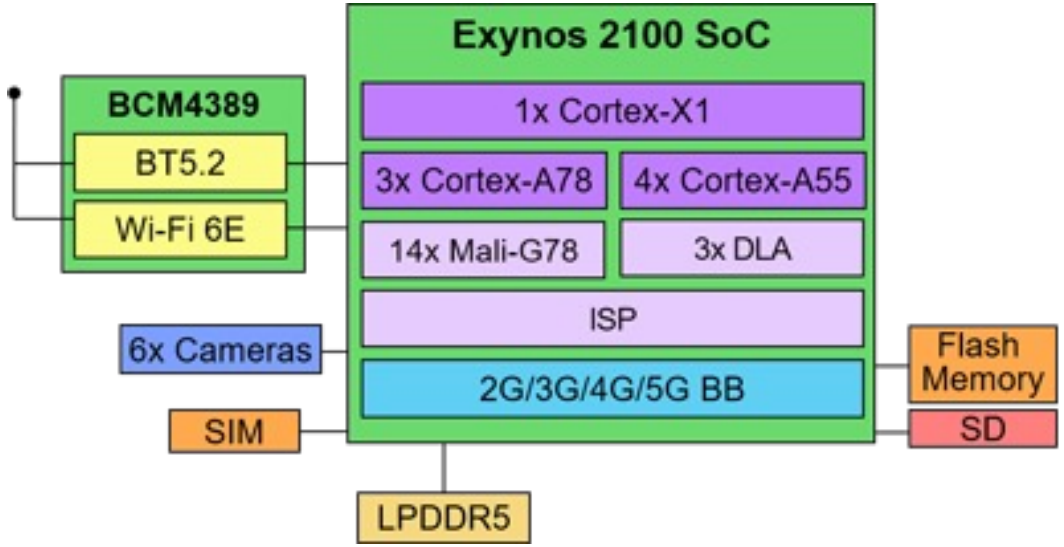


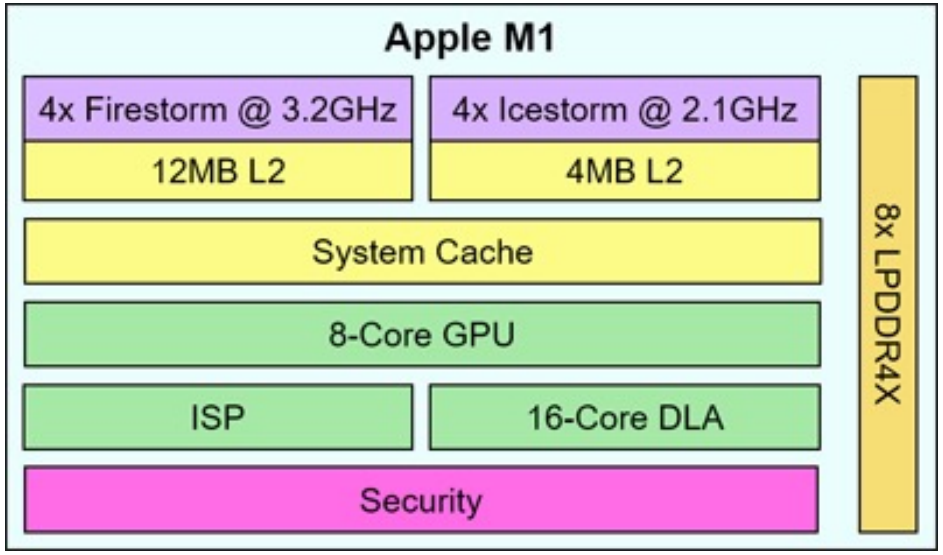
Evolución de los *cores* Cortex-A y Cortex-X

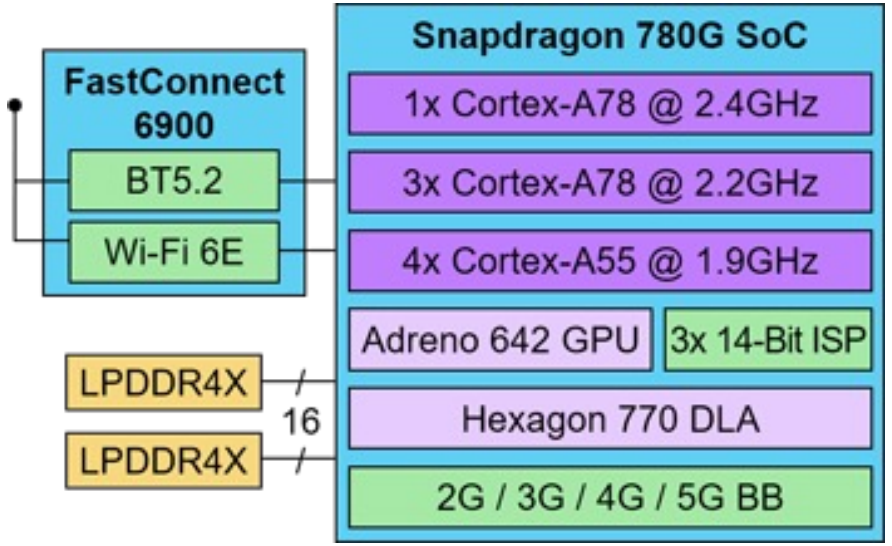


Cortex-A, Cortex-X: *system-on-chip (SoC)*

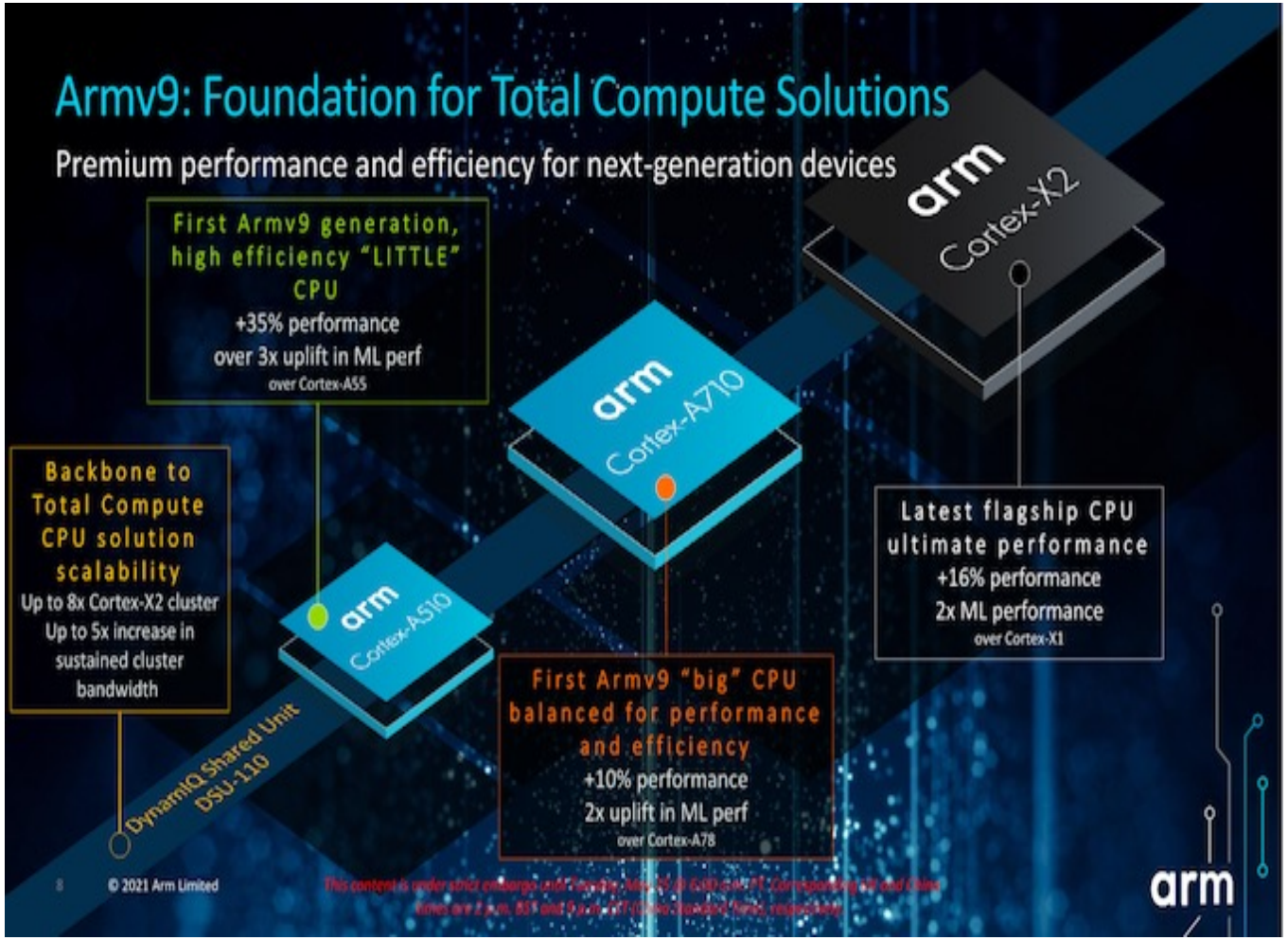
- Varios núcleos de diferentes características
- Graphics processing Unit (GPU)
    - Ray-tracing
- Neural networks processing unit (DLA)
- Otros aceleradores (ISP, 5G, …)







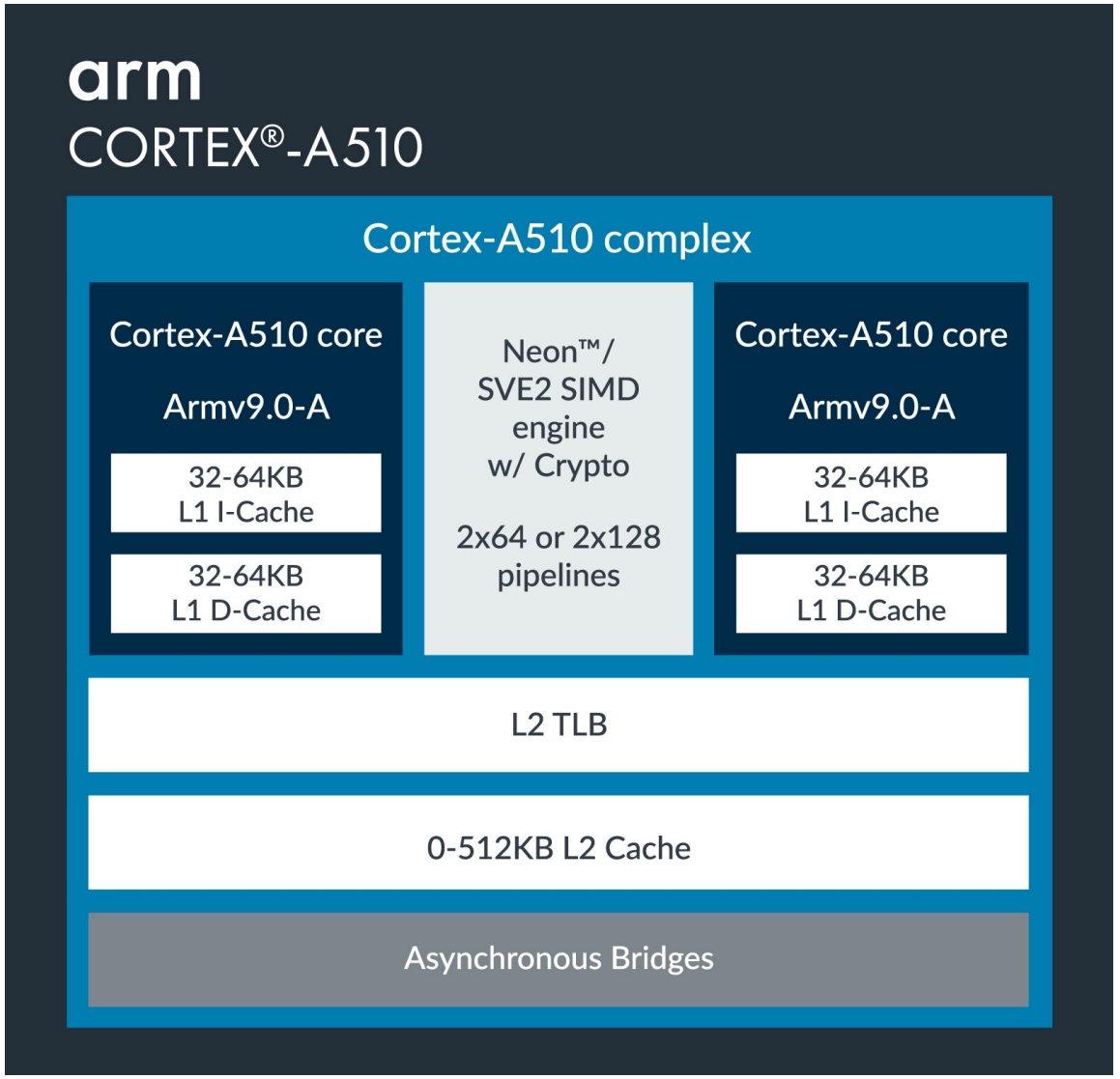
Tri-cluster: Cortex-X2, Cortex-A710, Cortex-A510



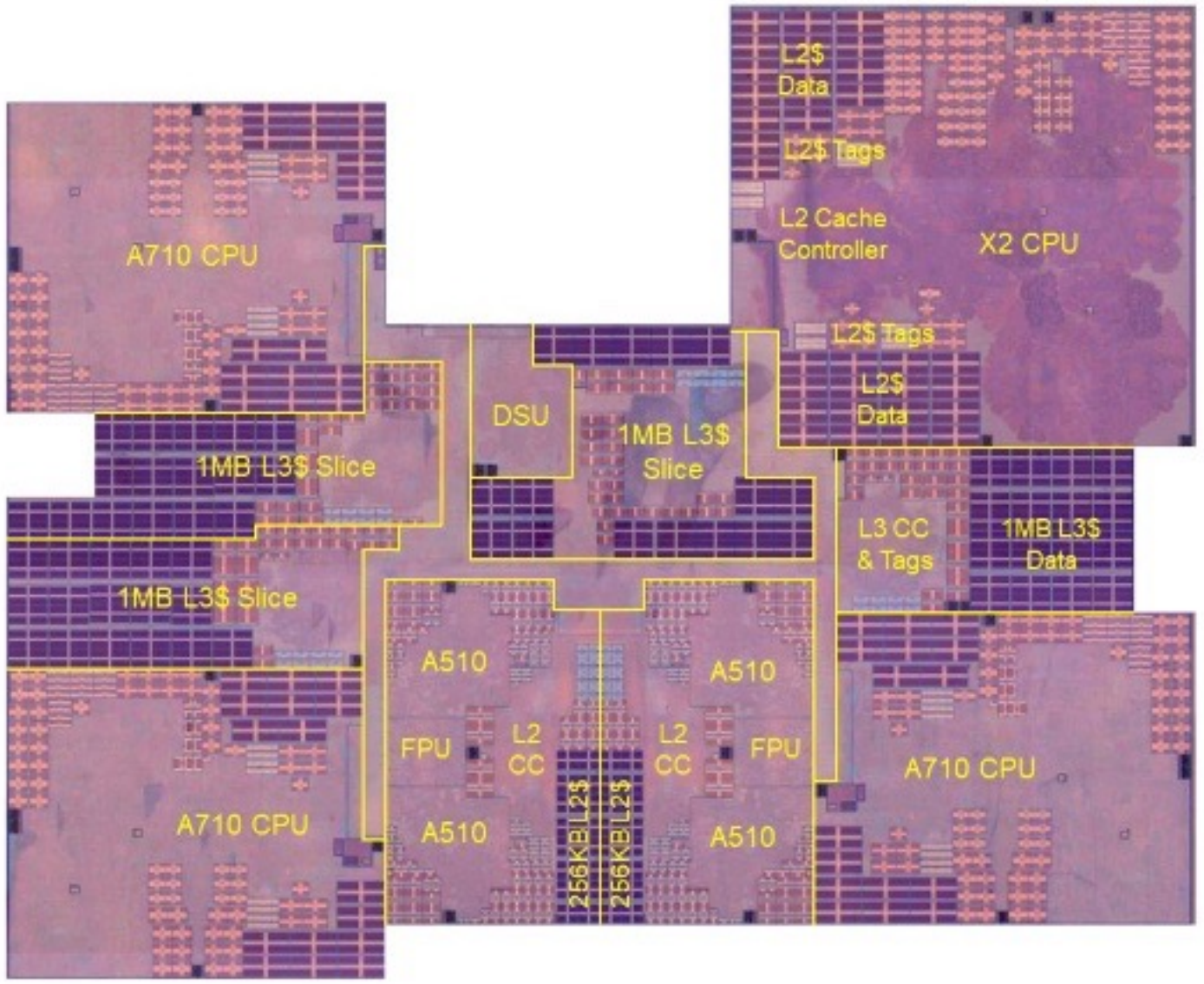
- Evolución de big.LITTLE
- Diferentes puntos en la curva de *PPA* (*power, performance, area*)
- **Cortex-X2**: "*performance first*", a costa de mayor área y potencia
- **Cortex-A710**: optimizado para eficiencia y performance, *PPA* balanceado
- **Cortex-A510**: *power efficiency*
- Soporte para aplicaciones de 32 bits solo en el Cortex A710

Cortex-A510: energéticamente eficiente

- *In-order*
- Microarquitectura *merged core:* dos núcleos que comparten L2, FP, NEON/SVE2
  - Cada núcleo es un núcleo completo con *front-end*, *integer back-end* y L1
  - Las cargas de trabajo más frecuentes para este núcleo son de enteros
  - Mejora en área y rendimiento
- 2x64-bit o 2x128-bit pipelines
  - Menos unidades que en los *big cores*



Samsung Exynos 2200



Comparación del rendimiento de CPUs Arm y x86

1 núcleo

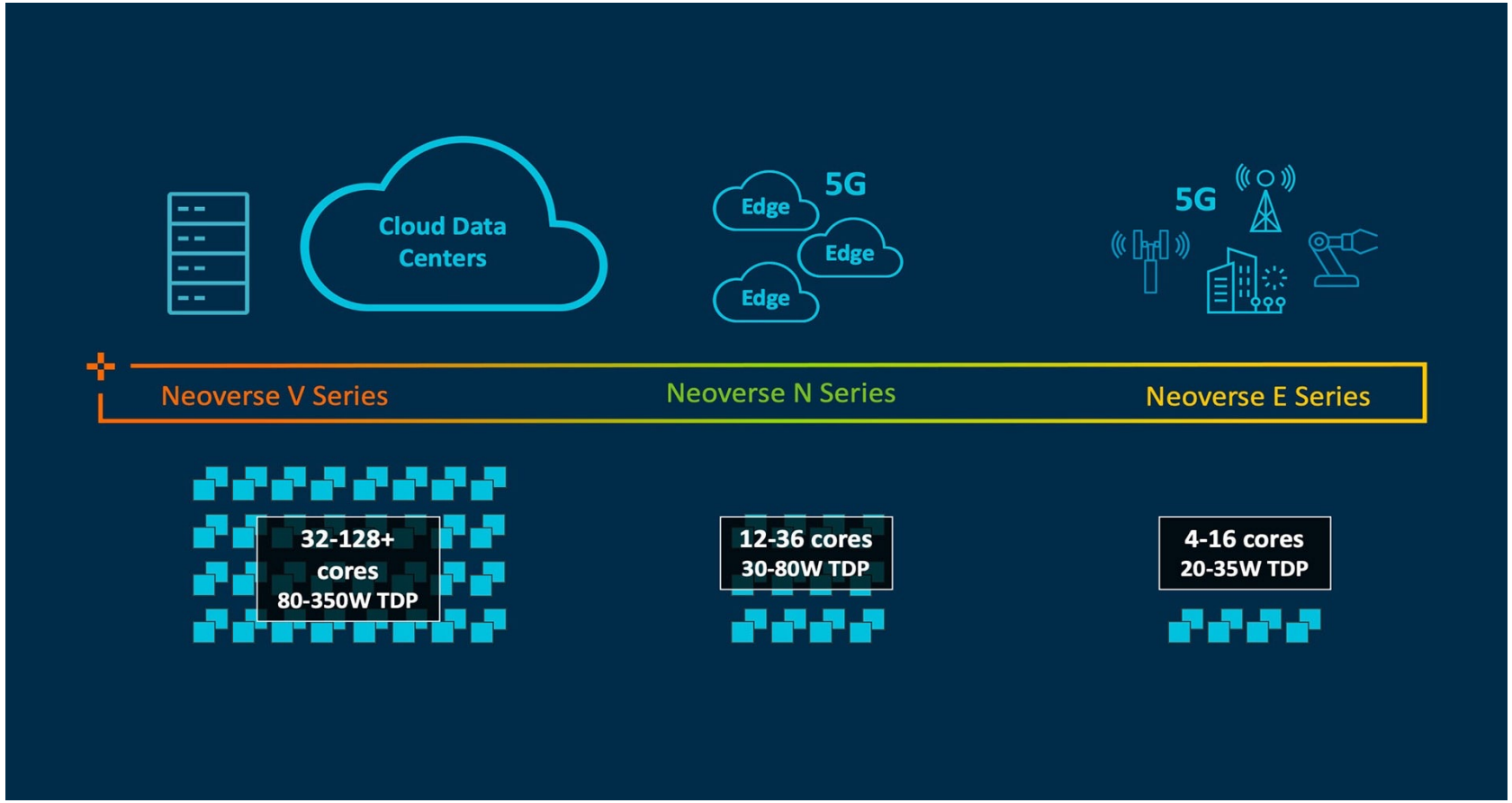
Multinúcleo

	Apple A15 Bionic	Intel Alder Lake i9-12900K	MediaTek Dimensity 9000	Qualcomm Snapdragon 8 Gen 1	Samsung Exynos 2200
Prime CPU(s)	2xAvalanche	8xGolden Cove	1xCortex-X2	1xCortex-X2	1xCortex-X2
Top Speed	3.2GHz	5.2GHz	3.05GHz	3.0GHz	2.8GHz
Private L2 Cache	256KB (L1)	1,280KB	1,024KB	1,024KB	1MB§
CPU Die Area	2.55mm²	6.76mm²	1.68mm²	2.98mm²	2.10mm²
Normalized Area	2.55mm²	4.87mm²	1.68mm²	2.77mm²	2.02mm²
Geekbench 5 SC	1,741	2,005	1,261	1,236	1,145
Other CPUs	4xBlizzard	8xGracemont	3xCortex-A710, 4xCortex-A510	3xCortex-A710, 4xCortex-A510	3xCortex-A710, 4xCortex-A510
Total CPUs	6 CPUs	16 CPUs	8 CPUs	8 CPUs	8 CPUs
Shared L3 Cache	12MB+4MB*	30MB	8MB	6MB	4MB§
Cluster Die Area	14.7mm²	95.6mm²	14.2mm²	15.6mm²	11.7mm²
Normalized Area	14.7mm²	68.8mm²	14.2mm²	14.5mm²	11.2mm²
Geekbench 5 MC	4,818	17,776	4,247	3,602	3,560
Chip Power†	5.0W	268.3W	7.4W	6.9W	6.9W
IC Process	TSMC 5nm	Intel 7	TSMC 5nm‡	Samsung 5nm‡	Samsung 4nm

Table 1. CPU performance comparison. SC=single core; MC=multicore. Die area normalized to TSMC 5nm. *12MB shared L2 for Avalanche CPUs plus 4MB shared L2 for Blizzard CPUs; †Geekbench 5 or Prime95 reported power; ‡advertised as 4nm but actually 5nm. (Source: vendors, except die area from TechInsights teardowns, benchmarks from notebookcheck.net, and §TechInsights estimate)

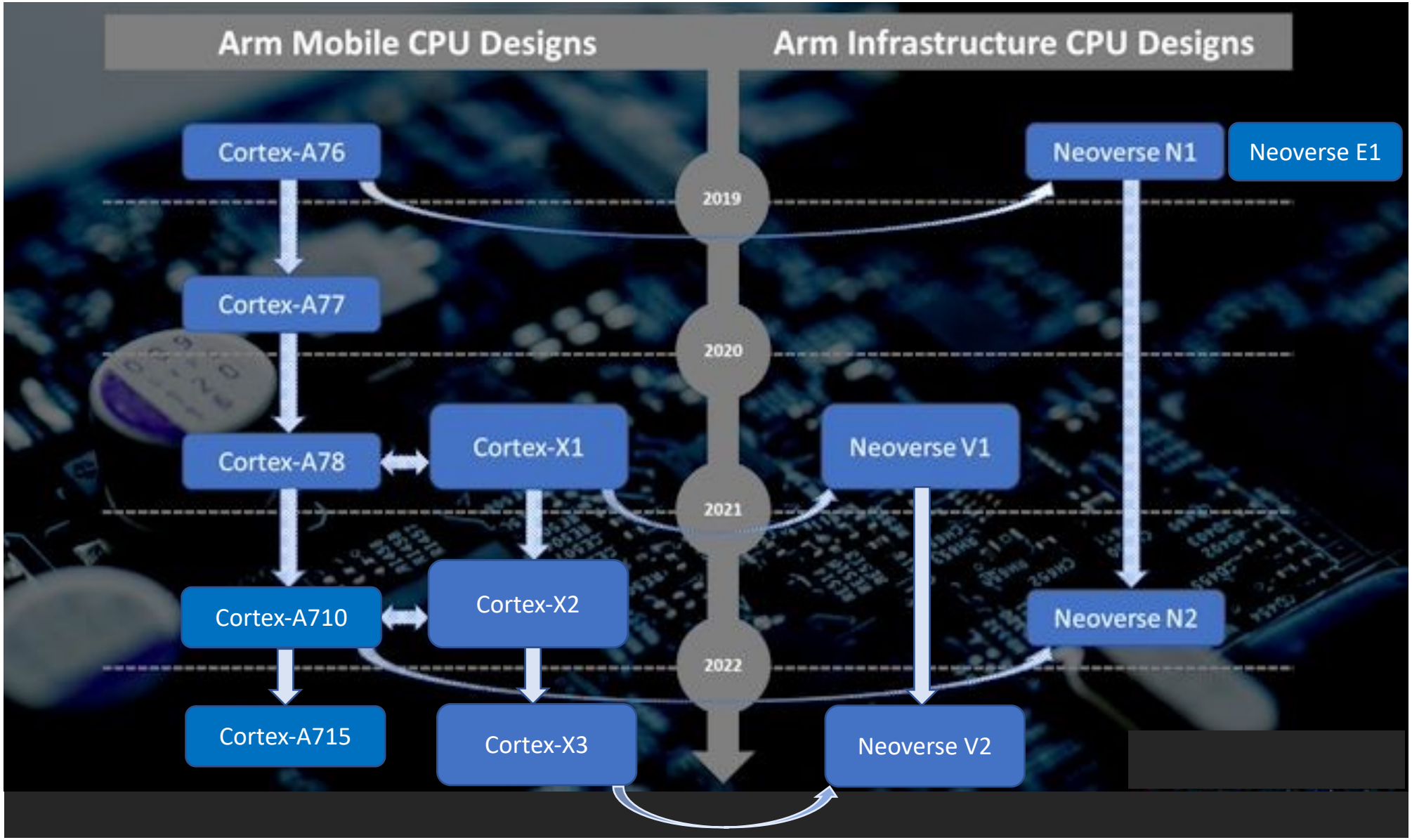
Edge-cloud computing: ARM Neoverse

- *Neoverse V*: Máximo rendimiento
- *Neoverse N*: rendimiento/W
- *Neoverse E*: Eficiencia



- *Neoverse V*: supercomputadores (HPC) y servidores high-end
- *Neoverse N*: servidores main-stream, DPUs y 5G
- *Neoverse E*: sistemas empotrados

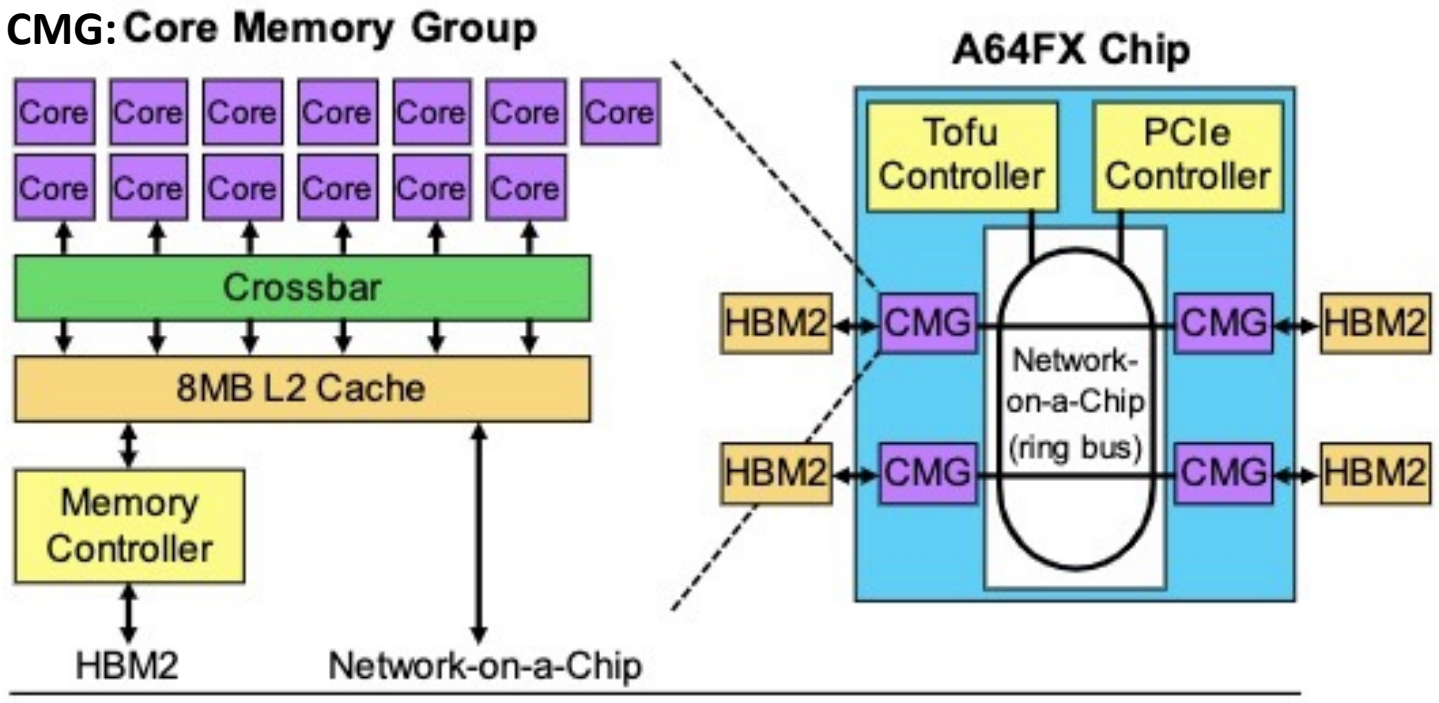
Edge-cloud computing: cliente e infraestructura



HPC con ARM Neoverse: Fujistu Fugaku

- Número 1 en el TOP500: junio 2020 - junio 2022
- *156.976* nodos NUMA A64FX, *52* núcleos por nodo, *7.630.848* núcleos
- Tofu NOC: topología de red en forma de *torus* desarrollado por Fujitsu

- A64FX: 1 nodo -> 4 CMGs con
  - 13 núcleos
  - 8 MB L2 compartida
- Núcleos **Neoverse V1**
- SVE de 512 bits en cada núcleo
- Frecuencia de operación baja para limitar el consumo de potencia ($\sim 2.2 GHz$)



HPC y cloud con ARM Neoverse

- Otros supercomputadores y cloud con Arm
  - ETRI
  - Ampere computing
  - Nvidia A100
  - Google Cloud
  - Microsoft Azure Cloud
  - Tencent Cloud
  - Oracle Cloud
  - Alibaba Cloud
  - Sandia National Laboratories
  - Amazon Web Services (AWS)
  - SiPearl (European Processor Initiative)
  - Marvell
  - …

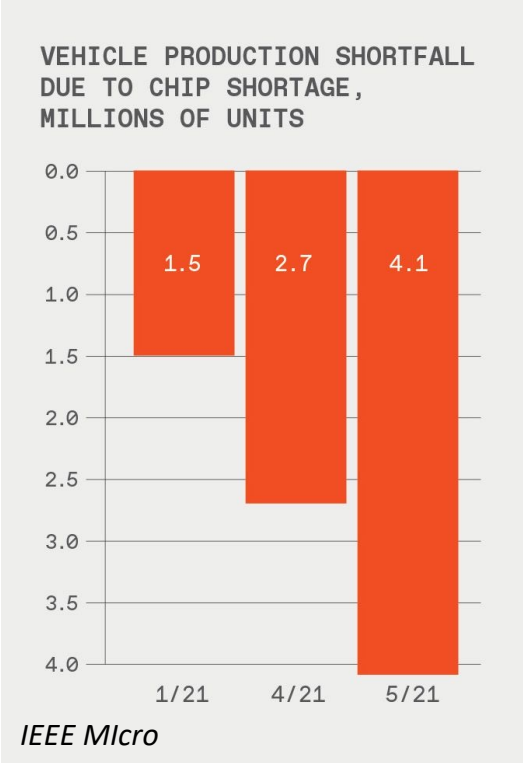
Procesadores ARM en automoción

Procesadores en automoción

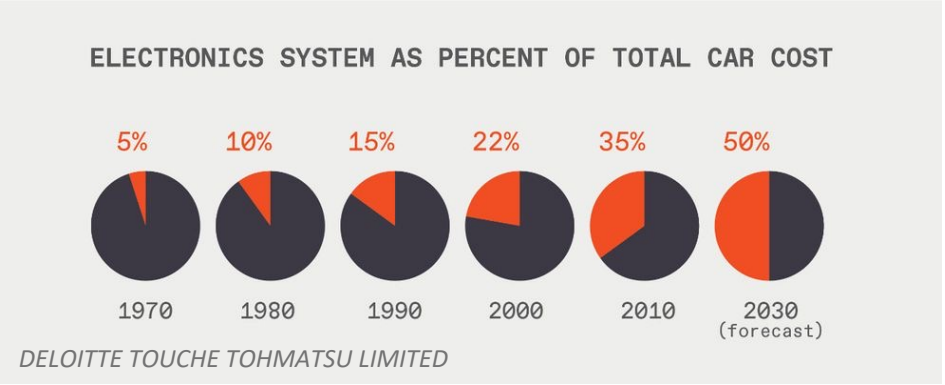
- Toma de autónoma de decisiones y terminación segura de las tareas
- En algunos entornos cualquier fallo o malfunción puede ser una amenaza vital
  - **Vehículos autónomos (AV) y ayudas a la conducción (Advanced Driver – Assistance Systems, ADAS)**
  - Automatización industrial
- Se requiere alta potencia computacional
  - Captura y análisis de grandes cantidades de datos
  - Numerosos sensores (cámaras, LIDAR, GPS, …)
  - Respuesta en tiempo real para predecir situaciones y posibilitar una respuesta o reacción
  - Se necesitan soluciones unificadas
- La seguridad demanda redundancia a todos los niveles

Procesadores en automoción

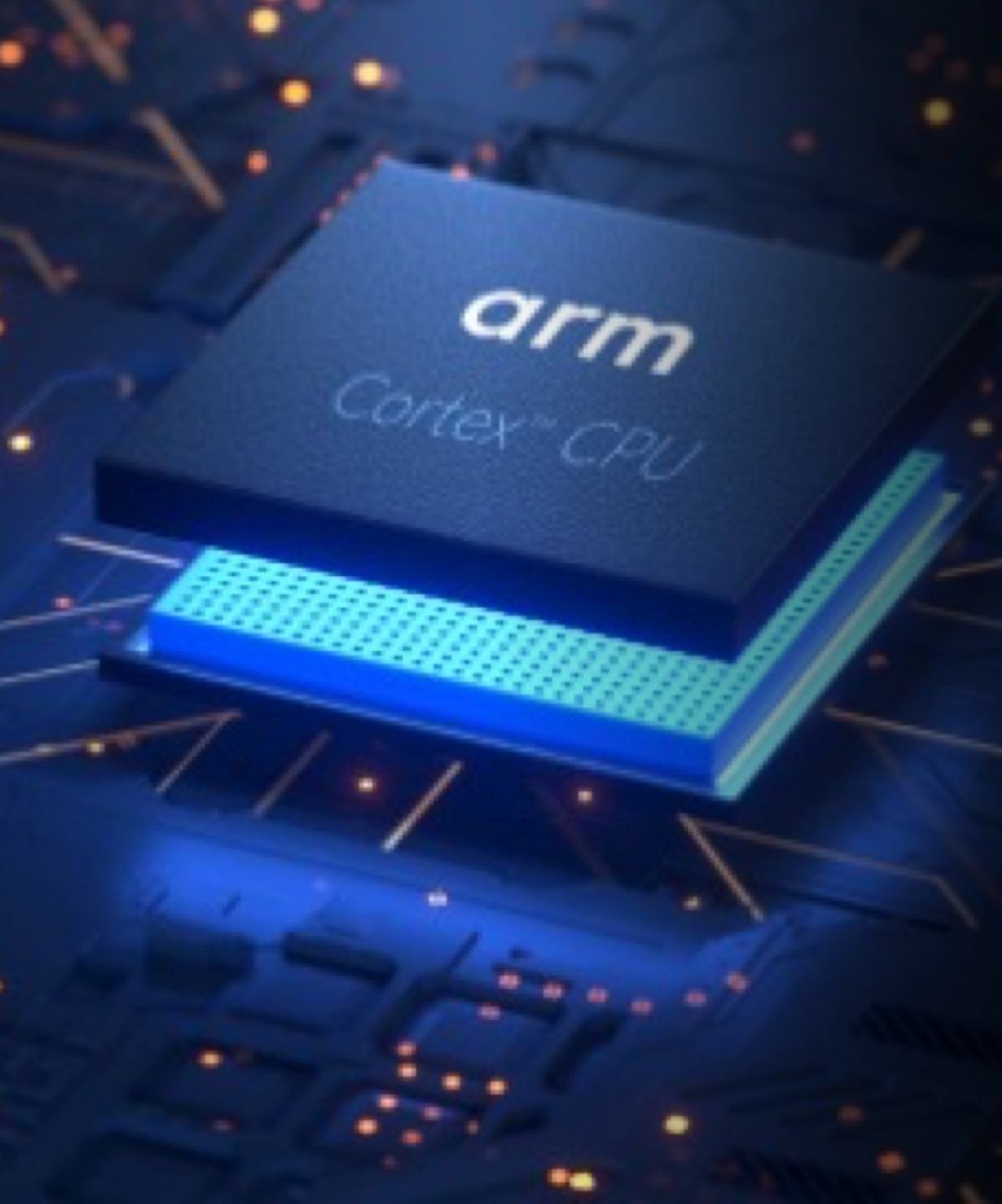
- La industria del automóvil utiliza decenas de *ordenadores* empotrados en los vehículos
- Cambio tecnológico sin parangón en ninguna otra industria
  - Muy afectada por la carencia de chips y la fragilidad de sus cadenas de suministros
- Electrónica y software son esenciales en los vehículos
  - 100-150 *unidades de control electrónico* (ECUs)
  - > 100 millones de líneas de código
  - 3000 chips de todo tipo
- Cerca del 40% del coste es atribuible a electrónica y software (*Deloitte*)
- Esfuerzo de integración de sistemas de diferentes OEMs
- El test es un reto



IEEE MIcro



DELOITTE TOUCHE TOHMATSU LIMITED



Muchas gracias por vuestra atención