

Tema 16: Arquitecturas de Dominio Específico

Arquitectura de Computadoras

Ing. Nicolás Majorel Padilla (npadilla@herrera.unt.edu.ar)

<http://microprocesadores.unt.edu.ar/arqcom/>

¿Cuáles temas veremos?



- ▶ Primera parte: **Generalidades**
 - ▶ Introducción. Conceptos generales.
 - ▶ *Performance, Price & Power.*
 - ▶ Conceptos de **Paralelismo y Pipelining.**
- ▶ Segunda parte: **Procesadores**
 - ▶ ISA: qué es, cómo se diseña.
 - ▶ Diseños de procesadores para *performance.*
- ▶ Tercera parte: **Subsistema de Memoria**
 - ▶ Memoria Caché.
 - ▶ Memoria Virtual.
- ▶ Cuarta parte: **Sistema completo**
 - ▶ Subsistemas de Entrada/Salida y de Comunicaciones.
 - ▶ Consumo de energía.
- ▶ Quinta parte: **Nuevos paradigmas**
 - ▶ Arquitecturas específicas.
 - ▶ Seguridad.

Temas que veremos

- ▶ Evolución de performance y de necesidades.
- ▶ Arquitecturas de Dominio Específico.
- ▶ Ley de Huang.
- ▶ Ejemplos de DSA.
- ▶ Tendencias actuales.

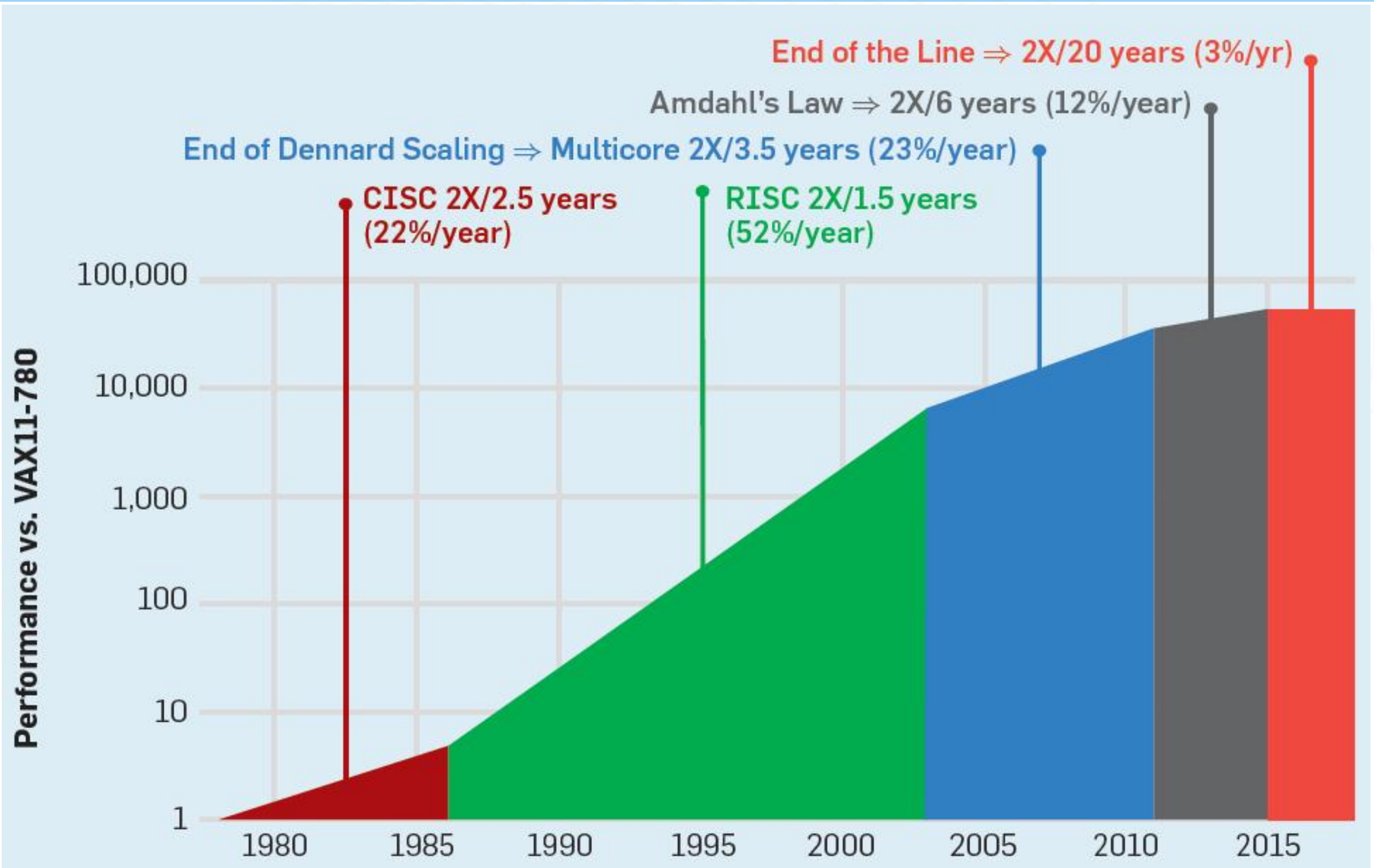
Lectura recomendada

- ▶ Computer Organization and Design, RISC-V Edition (2da ed, 2021):
 - ▶ Sección 6.7: *Domain-specific Architectures*

¿Dónde estamos parados?

- ▶ Vimos que ILP era una manera de incrementar la performance.
 - ▶ Además, es bastante transparente para los programadores.
 - ▶ Pero estaba fuertemente limitado por las dependencias, en códigos mayormente secuenciales.
- ▶ Vimos que la *Memory Wall* empezó a convertirse en un problema grande.
 - ▶ E intentamos reducirla con una jerarquía de memorias caché.
- ▶ Vimos que la *Power Wall* y el fin del escalamiento de Dennard nos impiden seguir aumentando la frecuencia para mejorar la performance.
 - ▶ Generaron un cambio de paradigma.
 - ▶ **Mejorar la performance mediante *multicore*.**
 - ▶ **Mejorar performance/watt.**

Evolución de la performance



Evolución de necesidades (ej: IA)

Exponential growth of computation in the training of notable AI systems

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations.

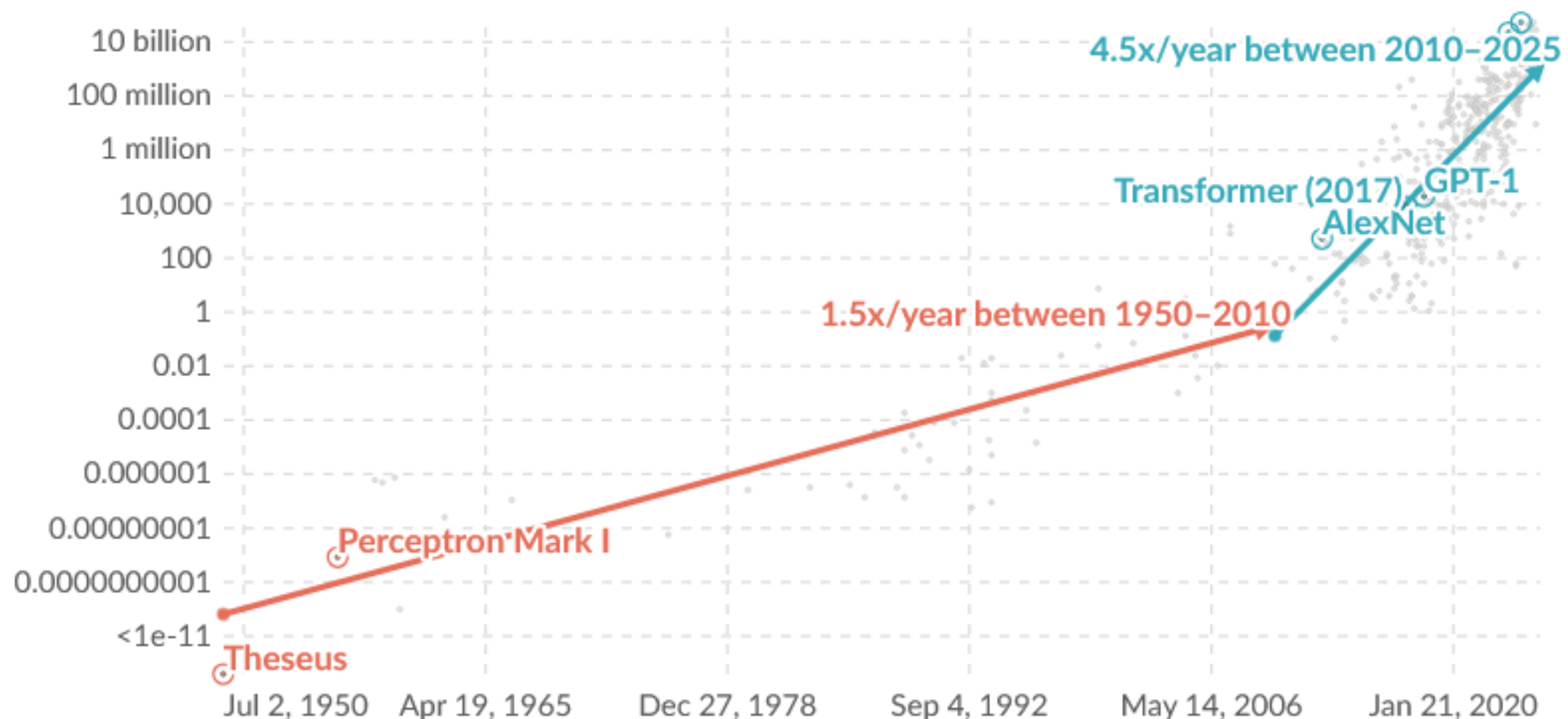
Our World
in Data

Table Chart

Select countries and regions

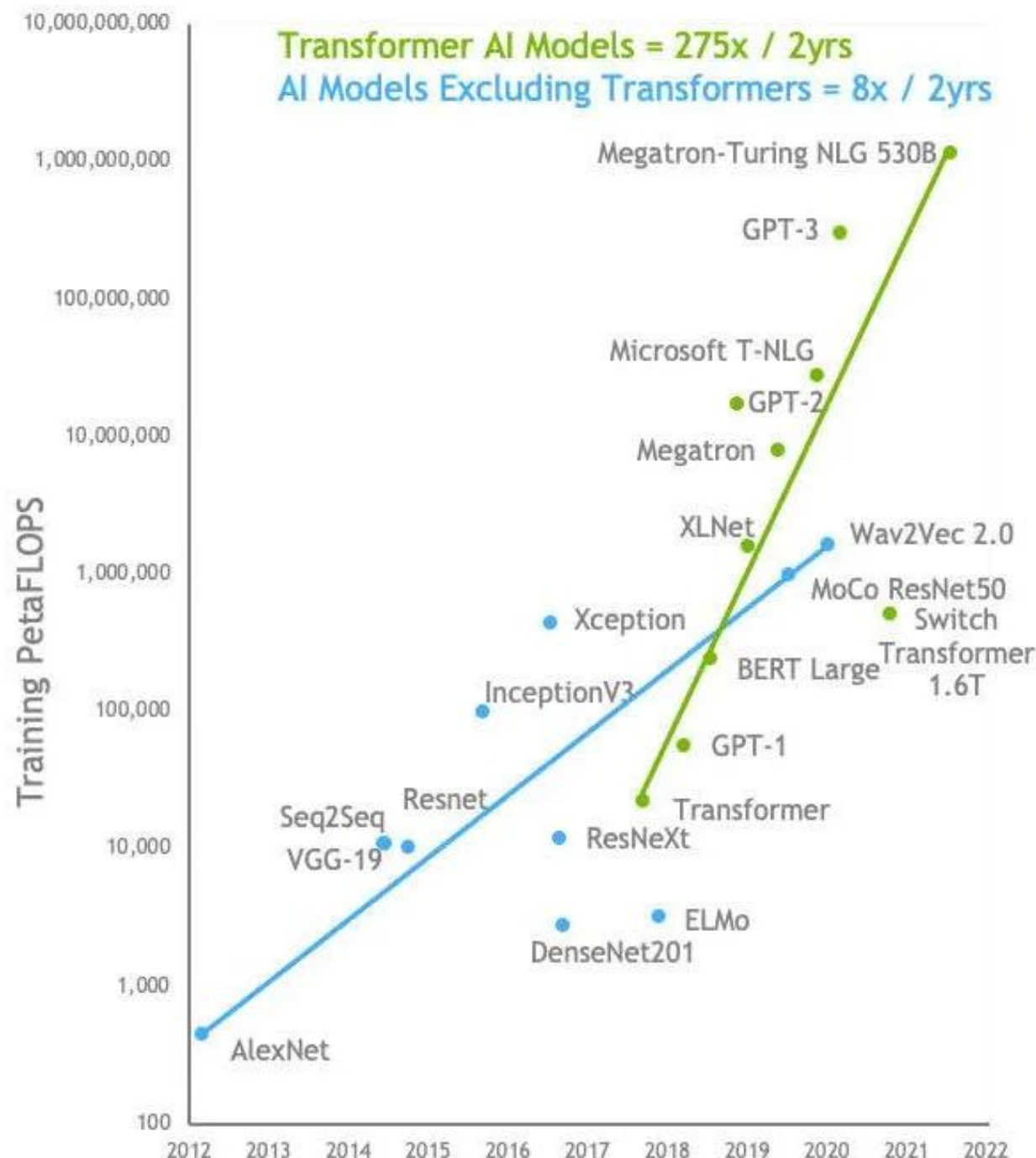
Settings

Training computation (petaFLOP)



Evolución de necesidades (ej: IA)

EXPLODING COMPUTATIONAL REQUIREMENTS



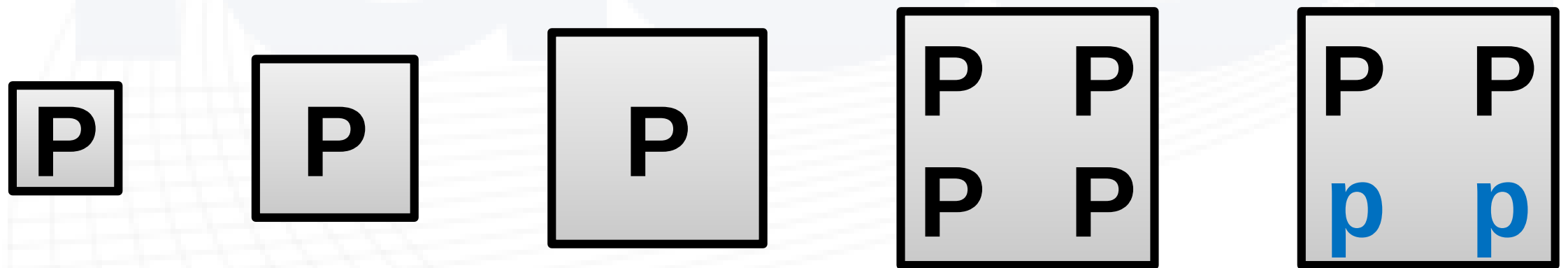
- ▶ En la gráfica anterior, performance 4,5x / año.
- ▶ Según una publicación de NVIDIA (2022), en algunos casos puede ser peor.
- ▶ <https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/>

¿Nuevo cambio de paradigma?

- ▶ Según la tendencia mostrada en los gráficos anteriores, los procesadores recién duplicarán su performance en 2038.
 - ▶ Pero las necesidades de performance están aumentando más rápido.
 - ▶ Las aplicaciones actuales demandan (y las del futuro demandarán) mucha más capacidad de cómputo.
- ▶ La performance no puede mejorar agregando transistores.
 - ▶ Debido la desaceleración de la Ley de Moore, al fin del escalamiento de Dennard y a restricciones de consumo de energía.
- ▶ Tampoco mejorará agregando múltiples núcleos.
 - ▶ Limitados por la Ley de Amdahl.
- ▶ *¿Qué se puede hacer a nivel Arquitectura de Computadoras?*
- ▶ La respuesta actualmente parece ser **Arquitecturas de Dominio Específico (DSA)**.

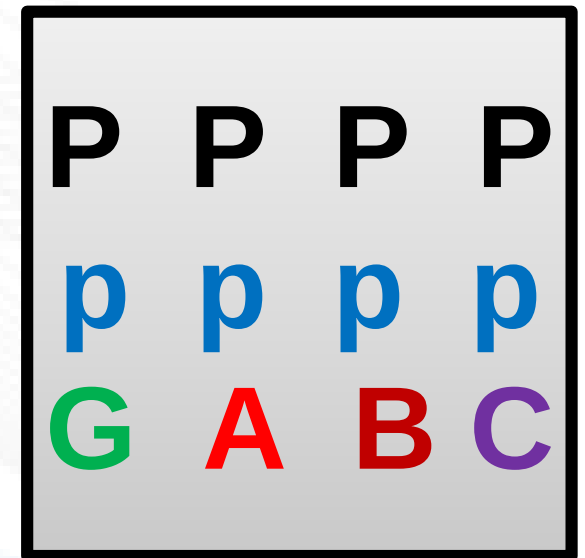
¿Nuevo cambio de paradigma?

- ▶ Hasta hoy en día, los procesadores (y los sistemas) se diseñan bajo la guía de “las tres P”.
 - ▶ Pero hay otro factor implícito: **Generalidad**. Deben poder ejecutar (ser compatibles) cualquier aplicación que el usuario final requiera.
- ▶ Antes, un procesador nuevo anunciaba cuántos transistores traía.
- ▶ Luego, se anunciaba cuántos núcleos homogéneos traía.
- ▶ Luego, los núcleos ya no fueron homogéneos.



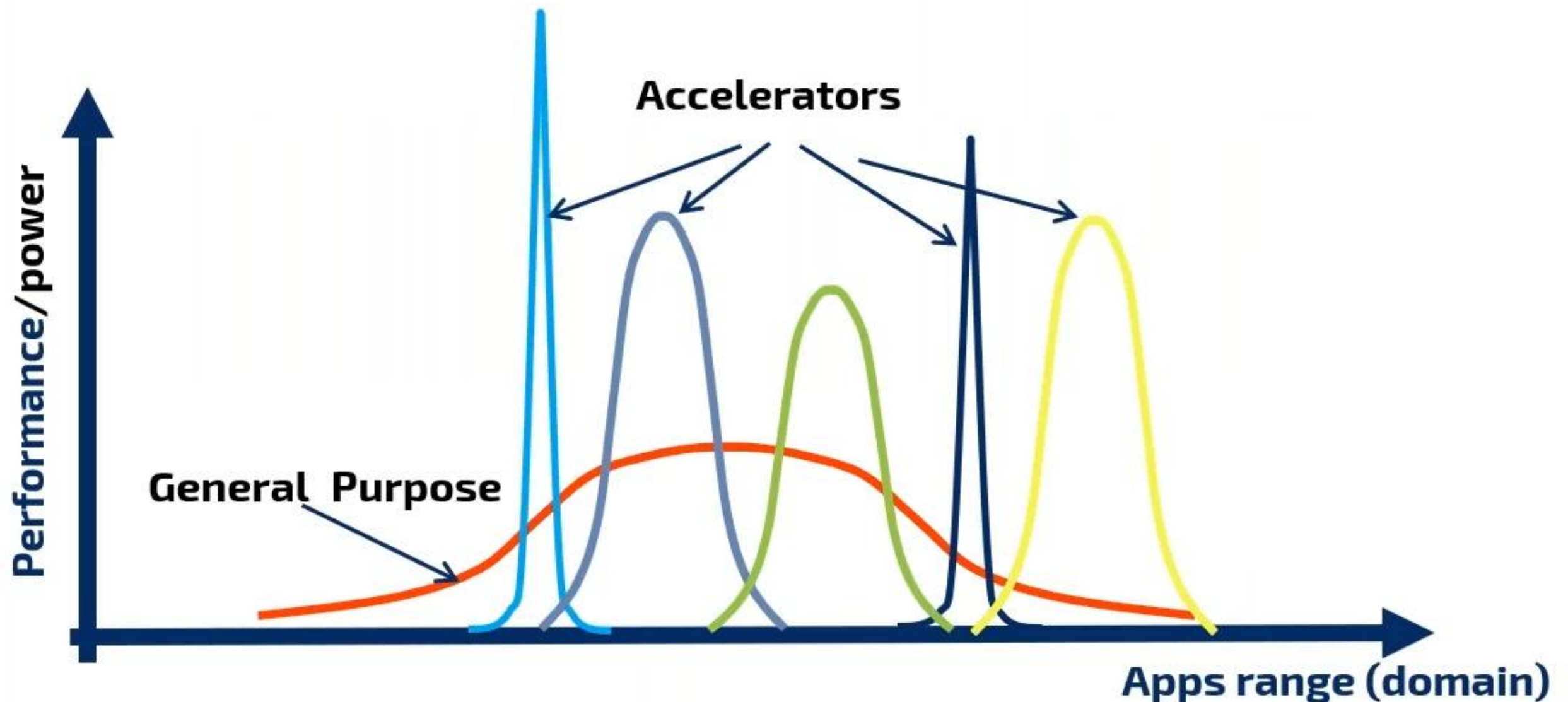
¿Nuevo cambio de paradigma?

- ▶ Actualmente, es una **mezcla heterogénea de varios núcleos**.
 - ▶ Cada núcleo dedicado a una aplicación, con su propio ISA, su propio software, sus propios algoritmos.
 - ▶ Así, tenemos CPU, GPU, NPU (Neural), NPU (Network), TPU, DPU, xPU, etc.
 - ▶ Ideales para fabricarse mediante *chiplets*.
- ▶ Estas Arquitecturas de Dominio Específico están siendo dominantes o están en pleno crecimiento.
 - ▶ En todos los tipos de computadoras (desktop, móviles, servidores, datacenters y sistemas embebidos).
 - ▶ En contraposición a las arquitecturas de propósito general.
 - ▶ Hacen muy pocas tareas, pero las hacen extremadamente bien.
 - ▶ *¿Recuerdan los IO Processors que vimos en el Tema 13?*



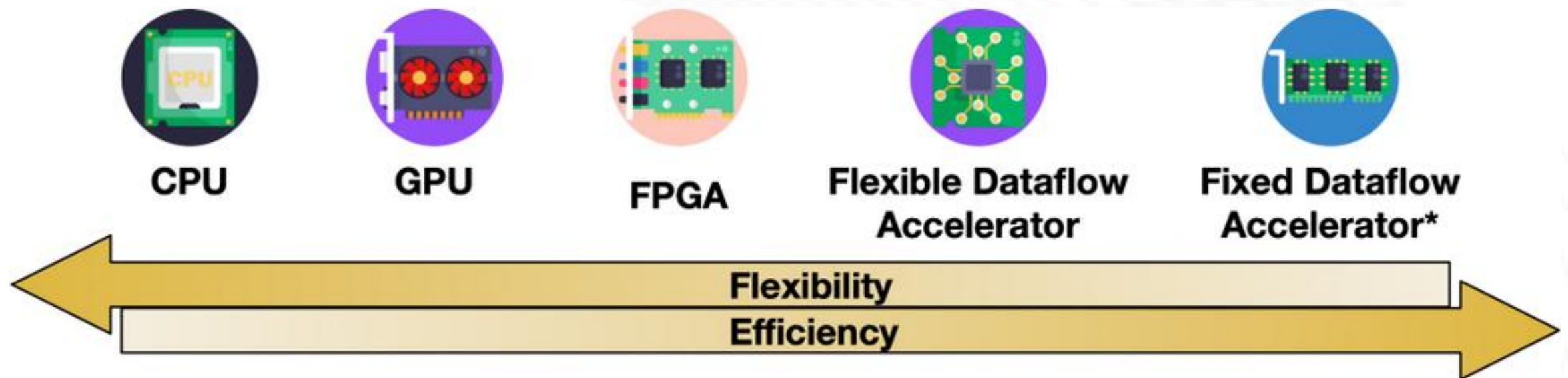
Arquitecturas de Dominio Específico

- ▶ Tienen una mejor relación performance/power, pero para una tarea específica.
- ▶ Reconfiguran (nuevamente) la Ley de Amdahl.



Arquitecturas de Dominio Específico

- ▶ Cambian eficiencia por flexibilidad:



- ▶ Cuando están completamente optimizadas para una tarea, si la tarea o sus datos cambian, pierden su ventaja.
 - ▶ Actualmente (2026), los modelos de IA están cambiando más rápido que los ciclos de desarrollo de los procesadores.
 - ▶ Lo cual genera un nuevo compromiso: *¿es mejor muy específicos o no tan específicos?*
 - ▶ Quizás Flexibilidad (o Generalidad) se convierta pronto en un nuevo criterio de diseño, junto con “las tres P”.

Arquitecturas de Dominio Específico

- ▶ Un ejemplo conocido: CPU vs GPU (del 2018 aprox).

Procesador	Modelo	Max Paralelismo	Frecuencia	Memory Bandwidth	Latencia a Caché L1
CPU	Intel Xeon E5-2690v4	28	3.5 GHz	76.8 GB/s	5 T
GPU	NVIDIA P100	3584	1.1 GHz	732 GB/s	80 T

- ▶ GPU con 3x menor frecuencia, pero con 100x paralelismo.
 - ▶ Ideal para aplicaciones con mucho paralelismo, más lenta para aplicaciones con poco paralelismo.
- ▶ GPU con L1 10x más lento, pero con un ancho de banda a memoria 10x más grande.
 - ▶ Ideal para aplicaciones con cálculos predecibles, en los que se pueda anticipar los accesos a memoria.
- ▶ Otro ejemplo: un procesador de audio, con un ISA que incluya instrucciones para implementar de manera óptima un algoritmo de cancelación de ruido.

Arquitecturas de Dominio Específico

- ▶ Tienen un **paralelismo masivo**: 1000x.
- ▶ **Altamente especializadas**:
 - ▶ Explotan al máximo el principio de localidad.
 - ▶ Optimizan la jerarquía de memorias para estructuras de datos específicas.
 - ▶ Reducen todos los *overhead* por comunicaciones al mínimo.
 - ▶ Tienen operaciones especiales y tipos de datos especiales. Inclusive pueden tener tamaños de palabra no convencionales.
- ▶ Se basan en un **co-diseño entre hardware y algoritmos**.
 - ▶ Los algoritmos deben ser modificados para aprovechar el nuevo Hw.
 - ▶ A su vez, el nuevo Hw debe ser diseñado para aprovechar al máximo los algoritmos y sus estructuras de datos.
 - ▶ El objetivo es acelerar la solución de un problema visto como un todo.

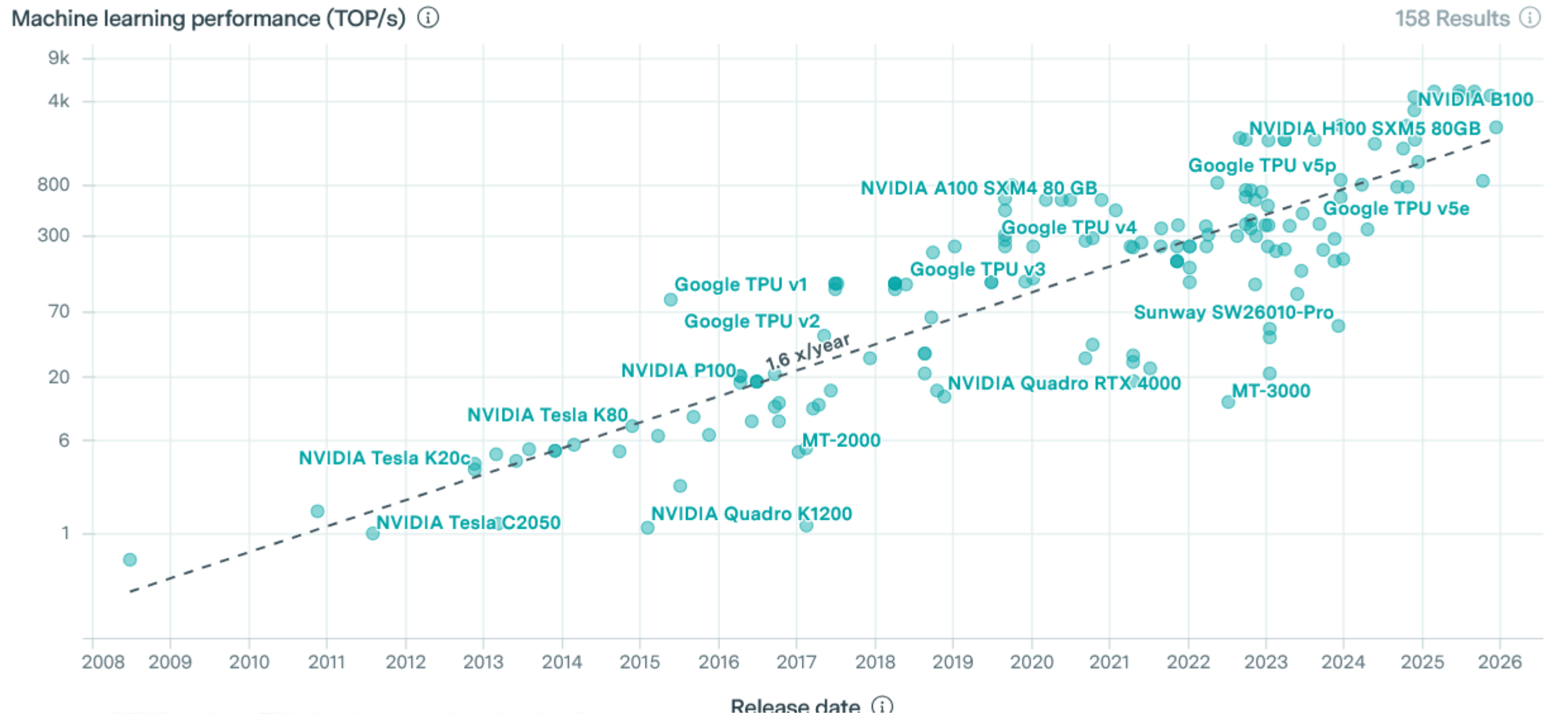
Arquitecturas de Dominio Específico

► Desventajas:

- Dificultad en su programación.
 - Los programadores deben aprender nuevos ISA para cada DSA.
 - *¿Recuerdan al ISA como contrato entre Hw y Sw (Tema 04)?*
 - *“No one uses a new piece of hardware because it’s not supported by software, and it’s not supported by software because no one’s using it.”*
 - *¿Unificar los ISA para disminuir la fragmentación? RISC-V.*
 - Todo el ecosistema de Sw (compiladores, SO, IDEs, lenguajes, debuggers, etc.) debe ser modificado para soportar el nuevo Hw.
 - Los programadores deben aprender librerías de paralelismo no convencionales, con abstracciones de alto nivel.
 - Los programadores deben modificar los algoritmos, comprendiendo a fondo qué es lo que hace el Hw.
- Como son específicas, no disfrutan de las ventajas de la economía de escala, por lo que sus costos fijos las hacen más caras.

¿Una nueva “Ley”?

- ▶ En 2018, el CEO de Nvidia pronosticó que “la performance de los procesadores para IA se duplicará cada año”.



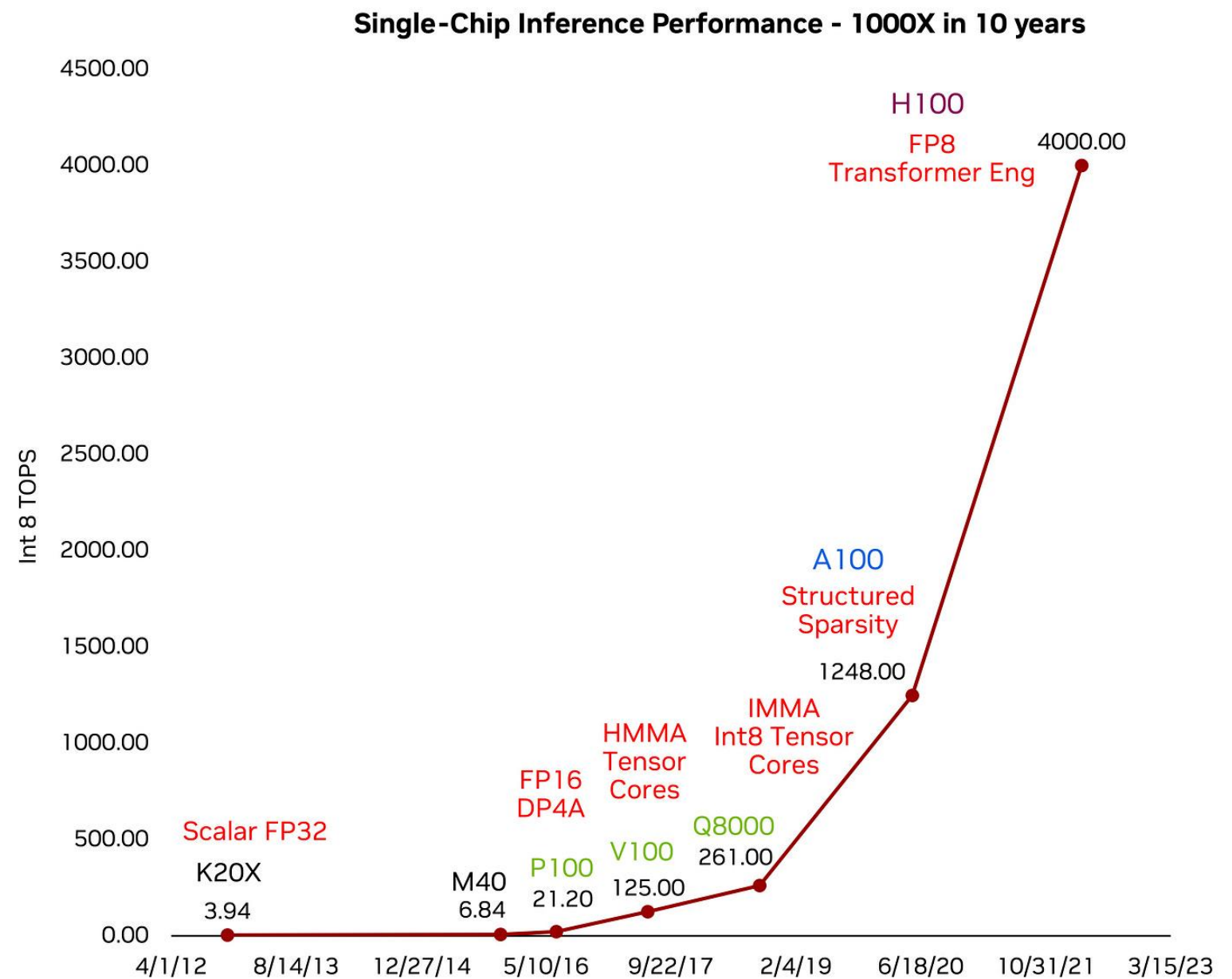
La “Ley de Huang”

- ▶ Se observa que es 1.6x, no 2x.
 - ▶ Nada mal, teniendo en cuenta que pasaron 8 años.
- ▶ No es una ley “general”, es específica: sólo para procesadores para IA.
- ▶ Puesta de otro modo, una mejora de 2x anual representa una mejora de 1000x en 10 años.
 - ▶ Analizada en detalle...

La “Ley de Huang”

Gains from

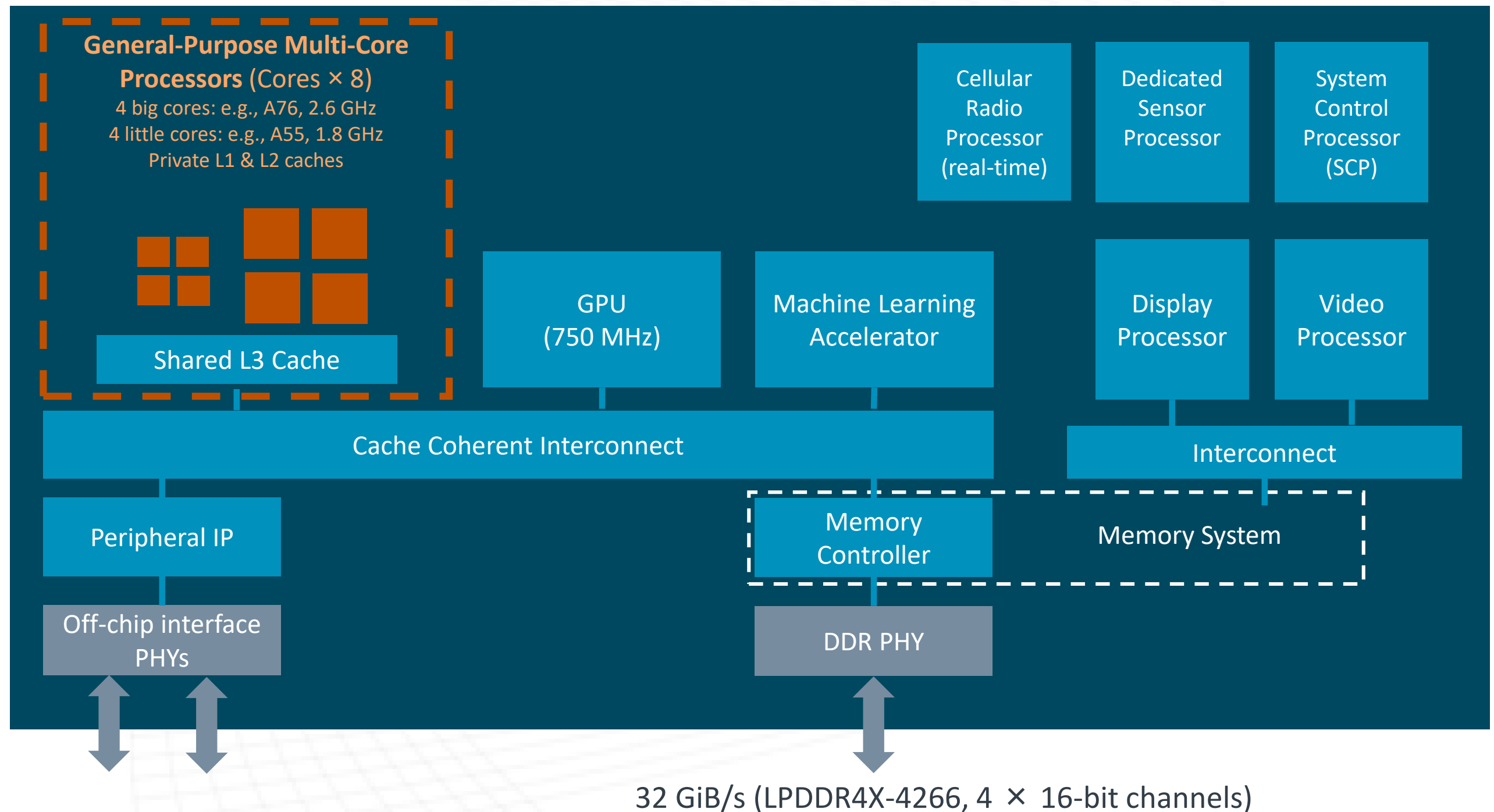
- Number Representation
 - FP32, FP16, Int8
 - (TF32, BF16)
 - ~16x
- Complex Instructions
 - DP4, HMMA, IMMA
 - ~12.5x
- Process
 - 28nm, 16nm, 7nm, 5nm
 - ~2.5x
- Sparsity
 - ~2x
- Model efficiency has also improved – overall gain > 1000x



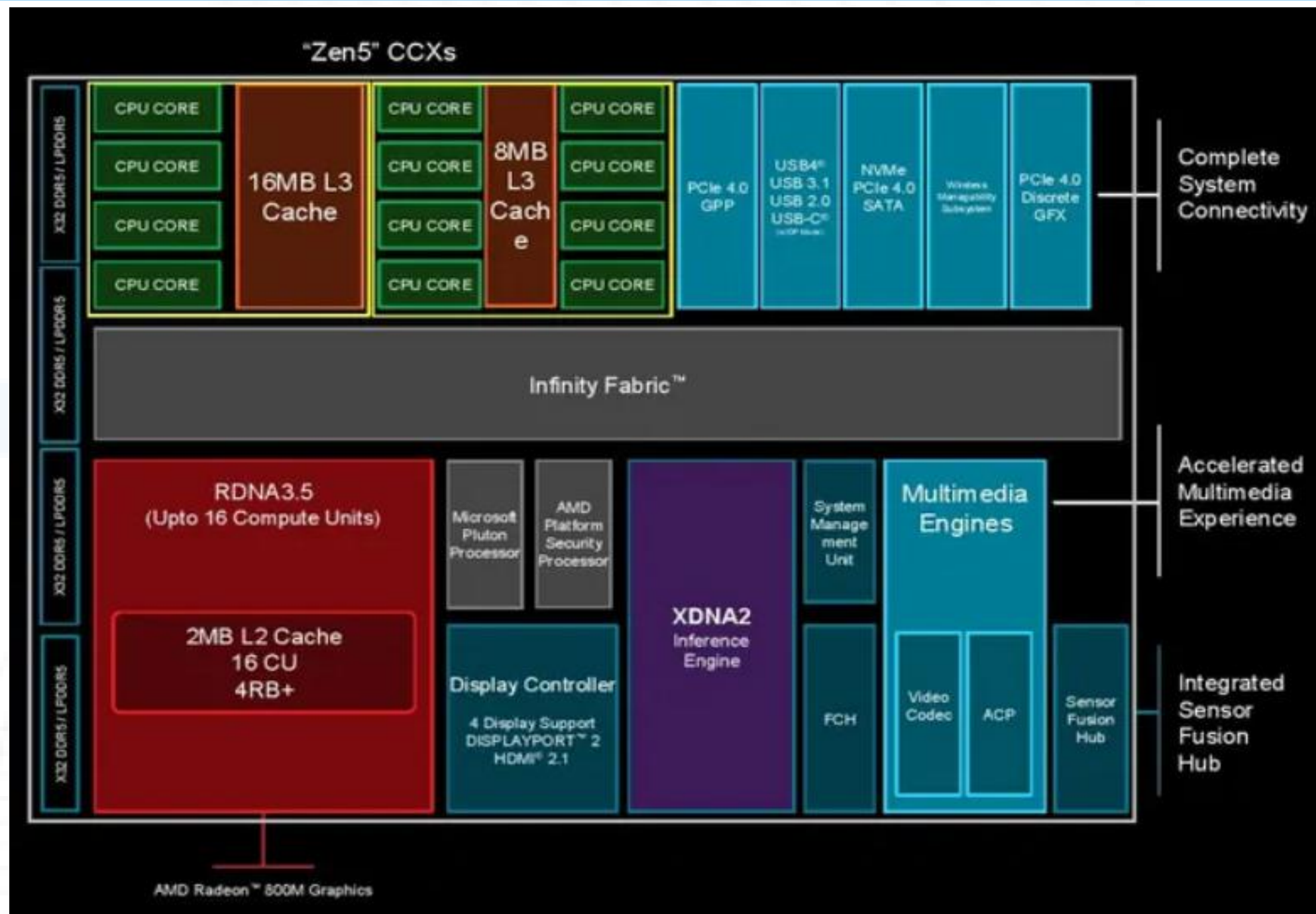
La “Ley de Huang”

- ▶ La mejora de 1000x se compone de:
 - ▶ 2.5x por mejora de proceso de fabricación.
 - ▶ La clásica Ley de Moore
 - ▶ 2x por mejoras algorítmicas.
 - ▶ Co-diseño entre hardware y software.
 - ▶ 16x por representación numérica.
 - ▶ Tipos de datos específicos y diseño específico de hardware.
 - ▶ 12.5x por nuevas instrucciones específicas.
 - ▶ **De tipo CISC**, no de tipo RISC.
 - ▶ Porque el cuello de botella suele estar en el ancho de banda de memoria.

Ejemplo de un SoC con varios DSA

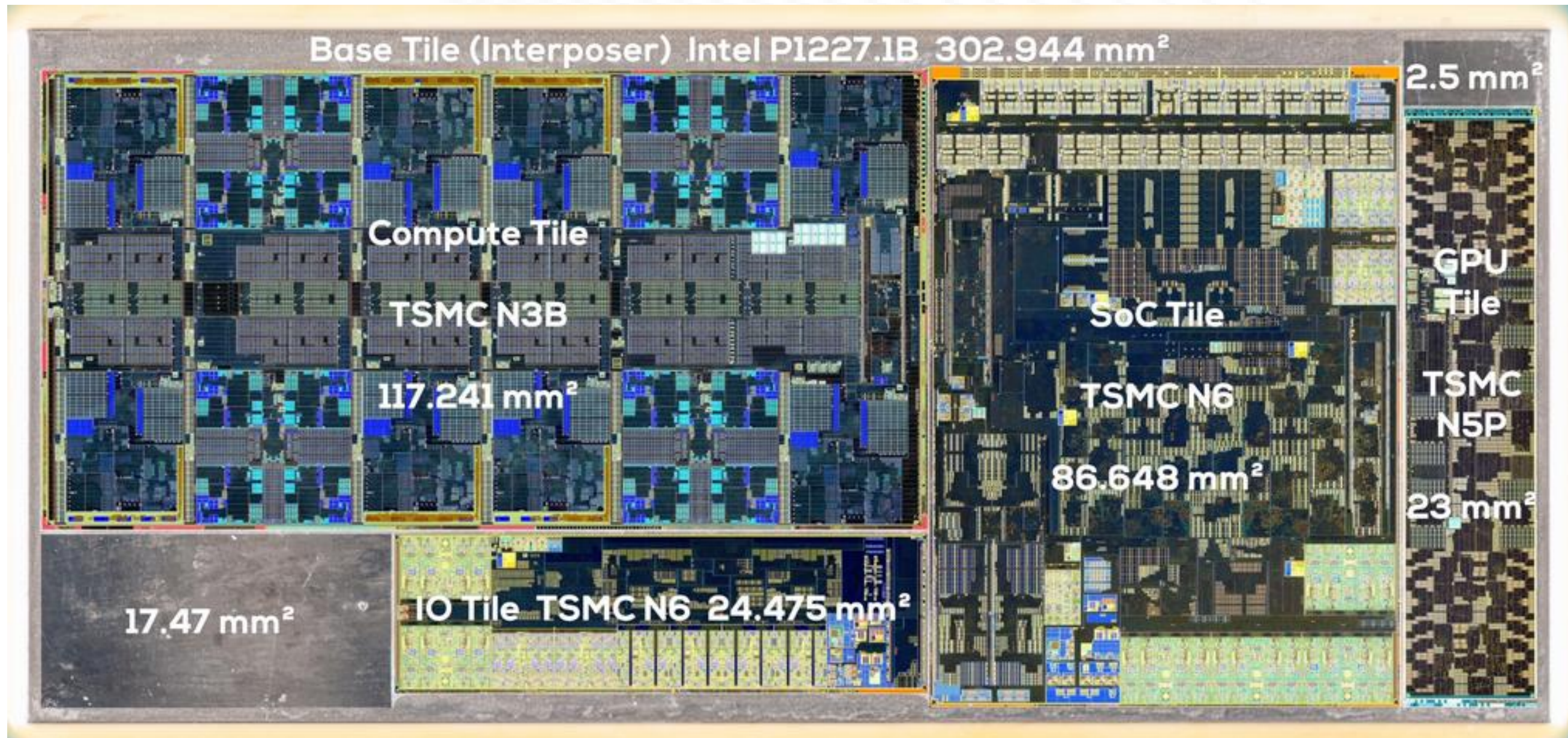


Ejemplo de un SoC con varios DSA



- Es un AMD Strix Point (2024)

Ejemplo de un SoC con varios DSA



- ▶ Intel Core Ultra 9 285k (Arrow Lake, 2024).
- ▶ Nótese los distintos procesos de fabricación (y los rellenos).

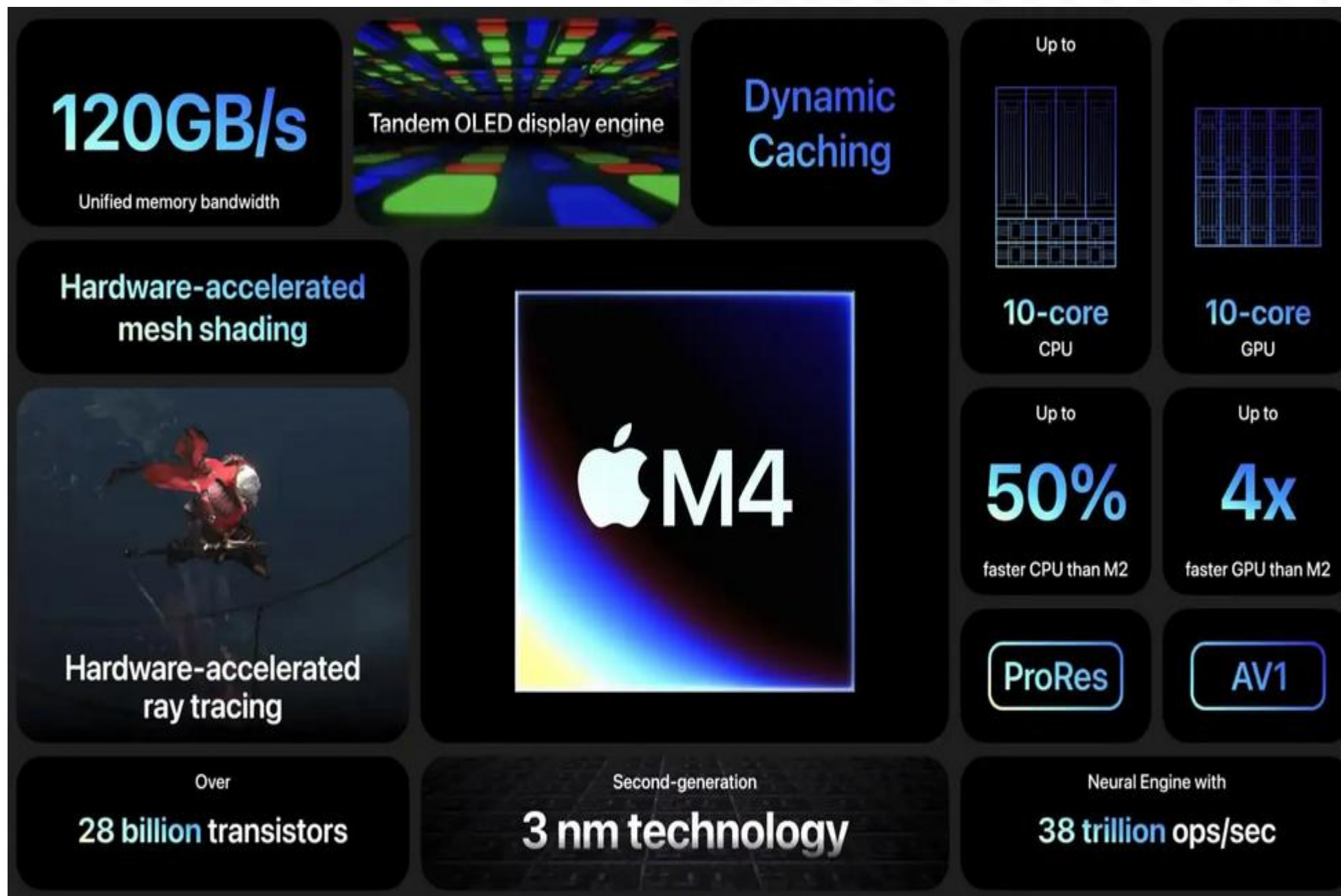
Ejemplos de DSA

- ▶ El Procesador M1 de Apple (2020) contiene:
 - ▶ 4 procesadores de alta performance (Firestorm) + 4 procesadores de alta eficiencia (Icestorm).
 - ▶ 1 procesador de video (GPU) de 8 núcleos + 1 procesador para redes neuronales de 16 núcleos.
 - ▶ 1 procesador para procesamiento de imágenes + 1 procesador para seguridad.
- ▶ Pasando a Google:
 - ▶ Diseñó sus propios chips para IA en 2015 (TPU). En abr-2026 anunció la 8va versión.
 - ▶ En 2021 anunció que diseñó sus propios chips para video (VPU), específicamente para códecs de YouTube.

Ejemplos de DSA

- ▶ Intel anunció en sep-25 una nueva versión de sus IPU (*Infrastructure Processing Units*):
 - ▶ Adaptadores de red con un montón de funciones adicionales.
 - ▶ 24 núcleos ARM de propósito general, y varios aceleradores para funciones específicas de redes.
 - ▶ Asignan máquinas virtuales, realizan encriptación, gestionan métricas de calidad de servicio.
 - ▶ “Cada núcleo que se dedique a esas tareas, es un núcleo menos que se pueda alquilar a clientes”.
- ▶ Otro ejemplo: ¿se acuerdan del procesador *WaferScale* que vimos en el Tema 02?
 - ▶ El chip más grande de la historia ejecuta una única tarea.
 - ▶ Un sistema de 32 chips posee 27 millones de núcleos, y puede procesar hasta 2 exaFLOPS con un consumo de solamente 16 kW.

Repaso: ¿Para qué aprender esos temas?



- ▶ Comprender a fondo la noticia.
- ▶ Discernir cuán cierta es.
- ▶ Discernir cuán útil me resulta.
- ▶ Discernir qué costo estaría dispuesto a pagar.

Tendencias actuales

- ▶ Un “procesador” actual es casi idéntico a un servidor NUMA de hace 20 años.
 - ▶ Sólo que mucho más pequeño, con todo lo que eso implica.
- ▶ Un rack de servidores actual se parece bastante a una supercomputadora de hace 30 años.
- ▶ Quizás podría decirse que los núcleos actuales son como las ALU de hace 25 años.
 - ▶ Y las ALU actuales, como si fuesen los transistores.
- ▶ *¿Recuerdan lo de las “fuerzas en constante movimiento”?*
 - ▶ Nadie sabe realmente qué esperar para dentro de 20 años.

Tendencias actuales

- ▶ Las DSA dominan en la actualidad.
 - ▶ Motivadas por la ambición de *Performance*.
 - ▶ Dejando completamente de lado *Price* y *Power*.
- ▶ No debe perderse del todo la Generalidad en pos de la Especificidad.
 - ▶ La historia de Arquitectura de Computadoras está plagada de cadáveres de máquinas de propósito específico.
 - ▶ Ideas brillantes que fueron opacadas por el avance inevitable de la Ley de Moore o porque no encontraron mercado.
 - ▶ En terminología de IA, se corre el riesgo de caer en *overfitting*.
 - ▶ Siempre debe haber lugar para la innovación.
 - ▶ Para evitar una “inercia tecnológica”, y que lo que hoy es un círculo virtuoso se convierta en uno vicioso.

Resumen final

- ▶ Nuevas aplicaciones demandan cada vez más performance en cada vez menos tiempo.
- ▶ Arquitecturas de Dominio Específico (DSA), o aceleradores.
 - ▶ Brindan mayor performance y mayor eficiencia a cambio de generalidad.
 - ▶ Paralelismo masivo, altamente especializadas.
 - ▶ Basadas en un co-diseño entre Hw y Sw.
- ▶ Ley de Huang: procesadores para IA duplicarán su performance cada año.
- ▶ Los procesadores actuales son un *System on a Chip* con una mezcla de muchos núcleos heterogéneos.
- ▶ Fuerzas en constante movimiento.