



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Los chips de la próxima década: problemas, soluciones y pruebas de concepto

José Duato

Universitat Politècnica de València,
Departamento de Informática de Sistemas y Computadores

Real Academia de Ciencias Exactas, Físicas y Naturales

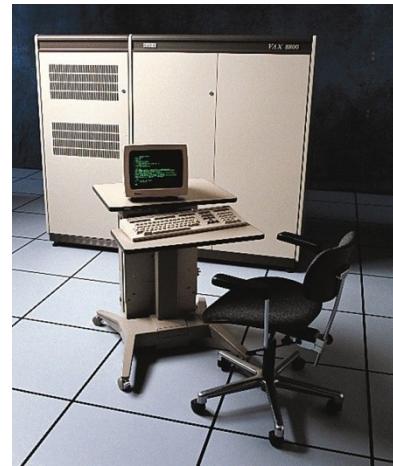


Evolución de las tecnologías de la información



Evolución de las tecnologías de la información

Durante décadas, las tecnologías de la información han progresado de forma espectacular



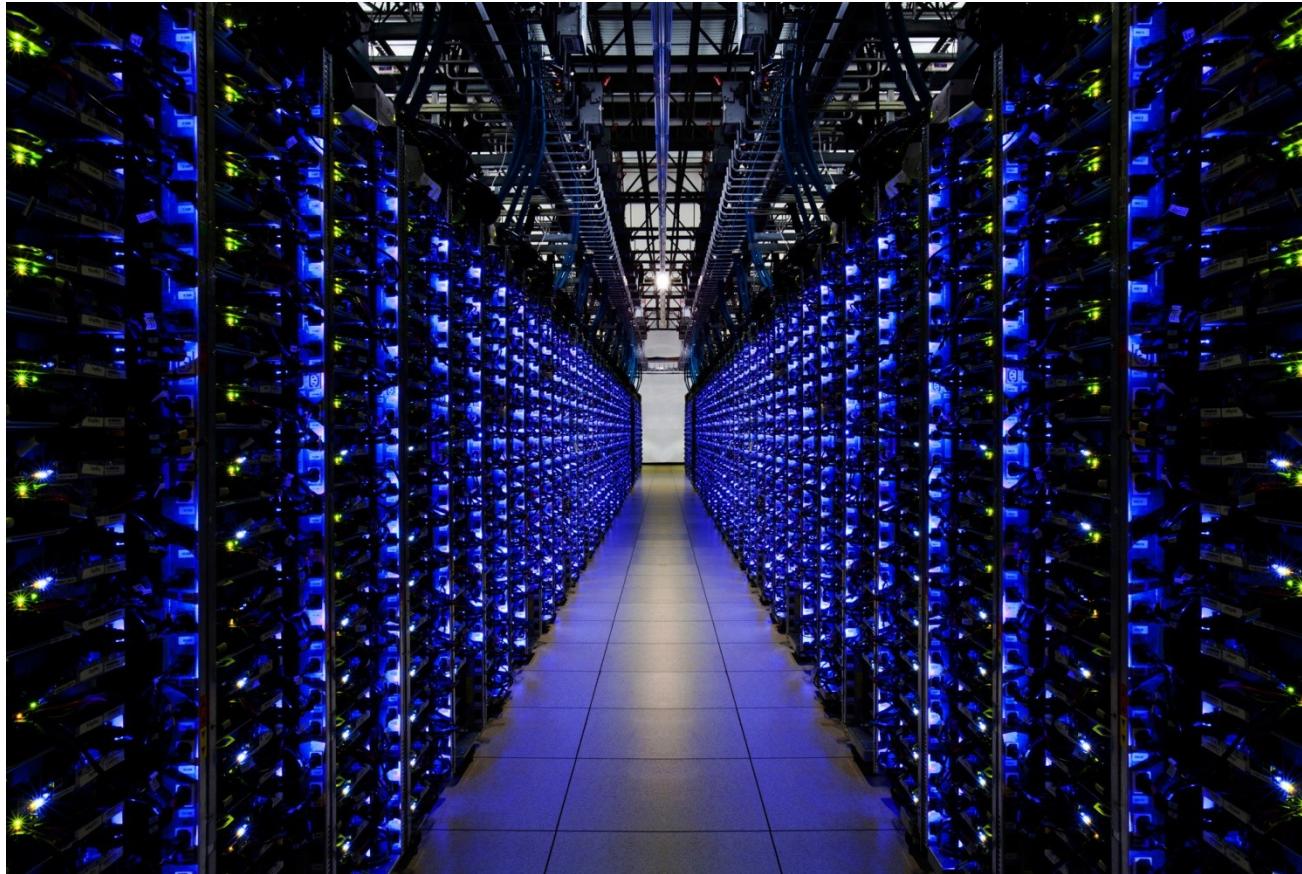
Evolución de las tecnologías de la información

Desarrollando dispositivos cada vez más rápidos y compactos, y con menor consumo de energía

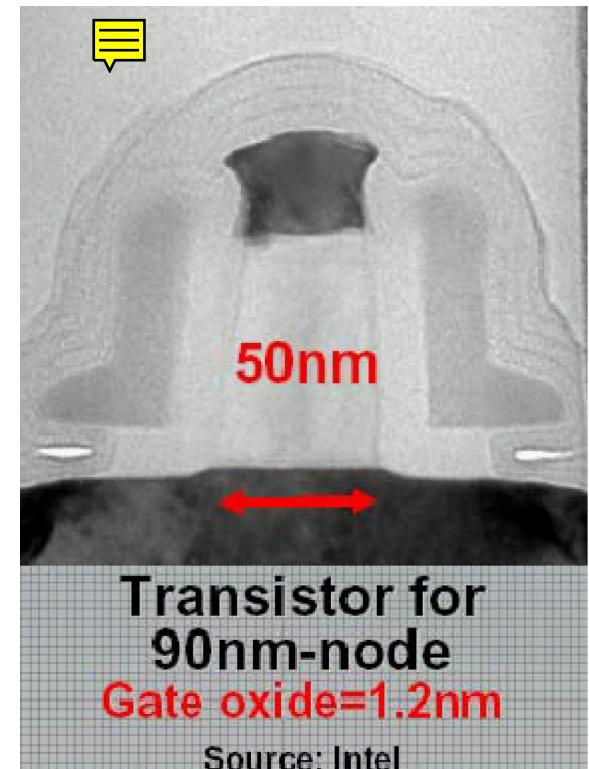


Evolución de las tecnologías de la información

Pasando de usar dispositivos aislados a disponer de enormes cantidades de información, gracias a Internet y a numerosos servidores de gran tamaño

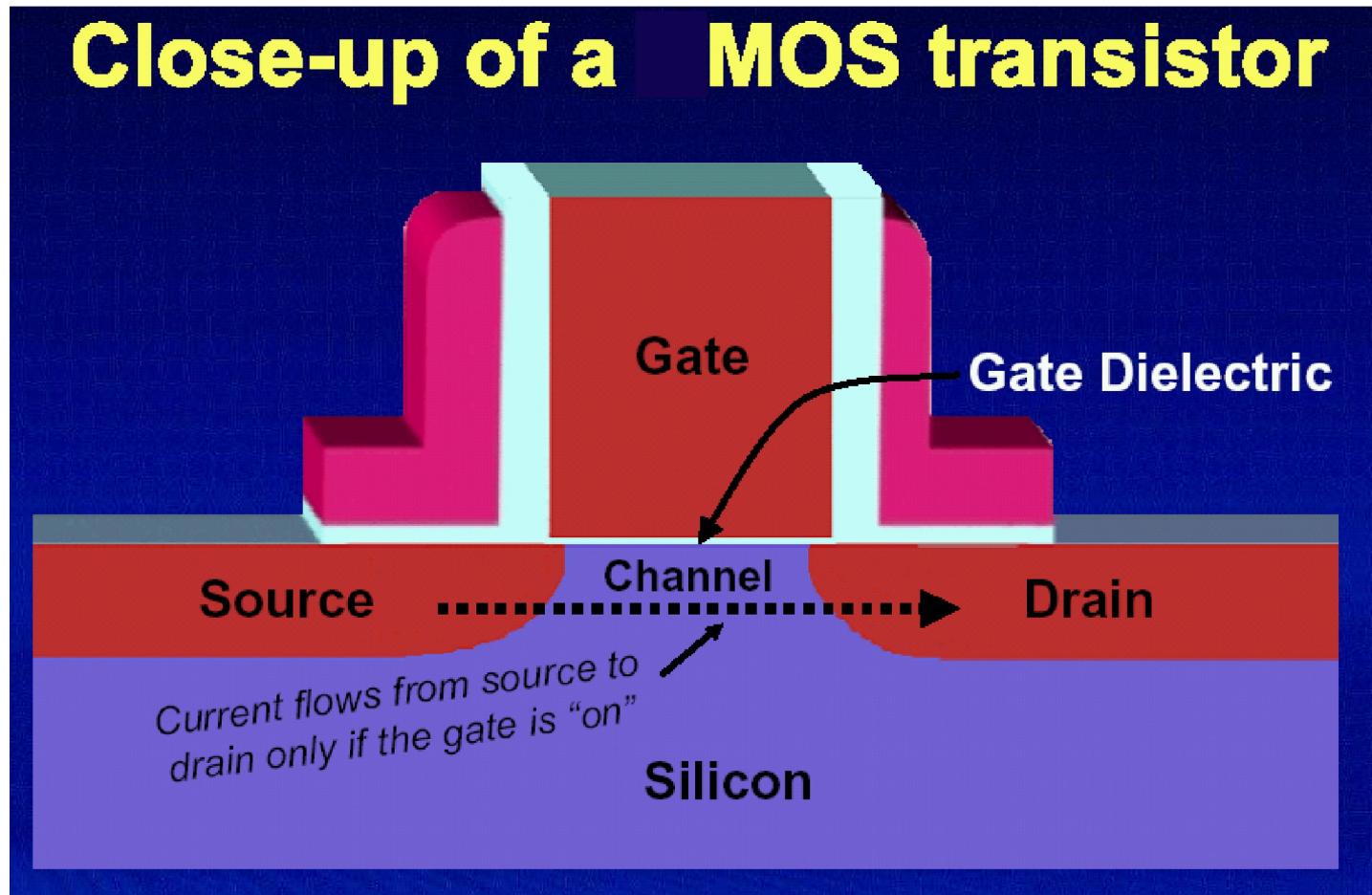


Elemento clave: el transistor



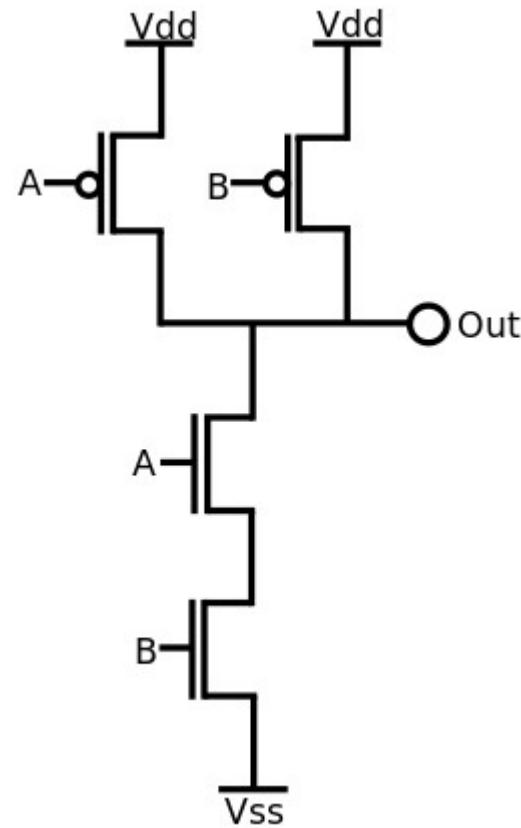
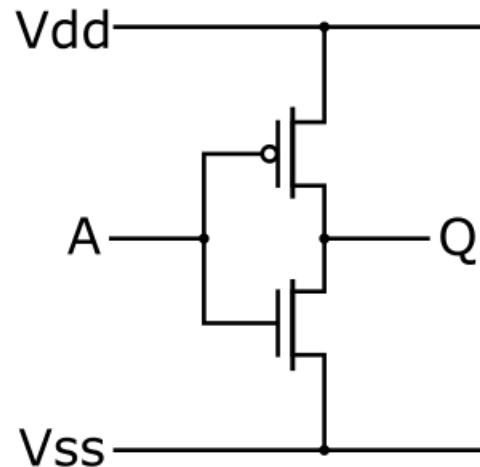
Elemento clave: el transistor

Los transistores MOS y la familia CMOS permiten circuitos digitales rápidos y de muy bajo consumo



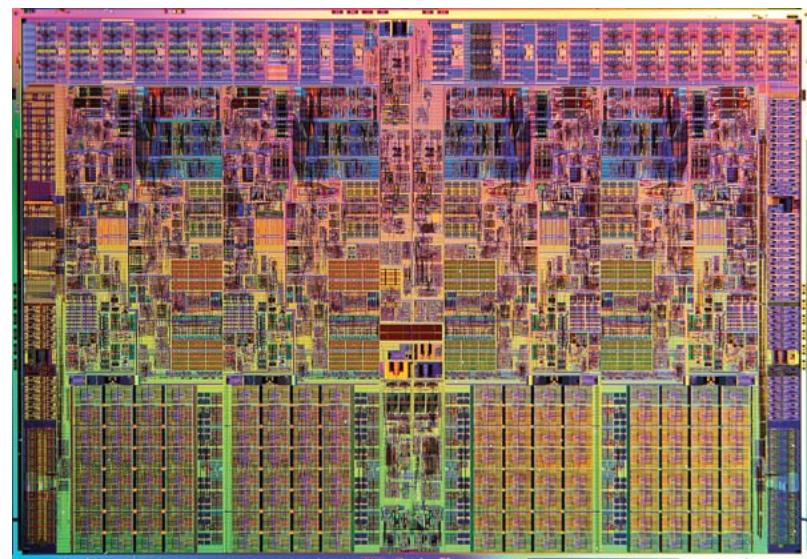
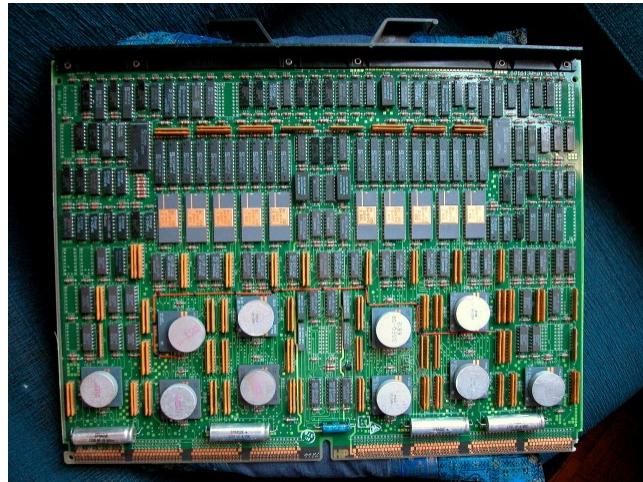
Elemento clave: el transistor

Los transistores MOS y la familia CMOS permiten circuitos digitales rápidos y de muy bajo consumo

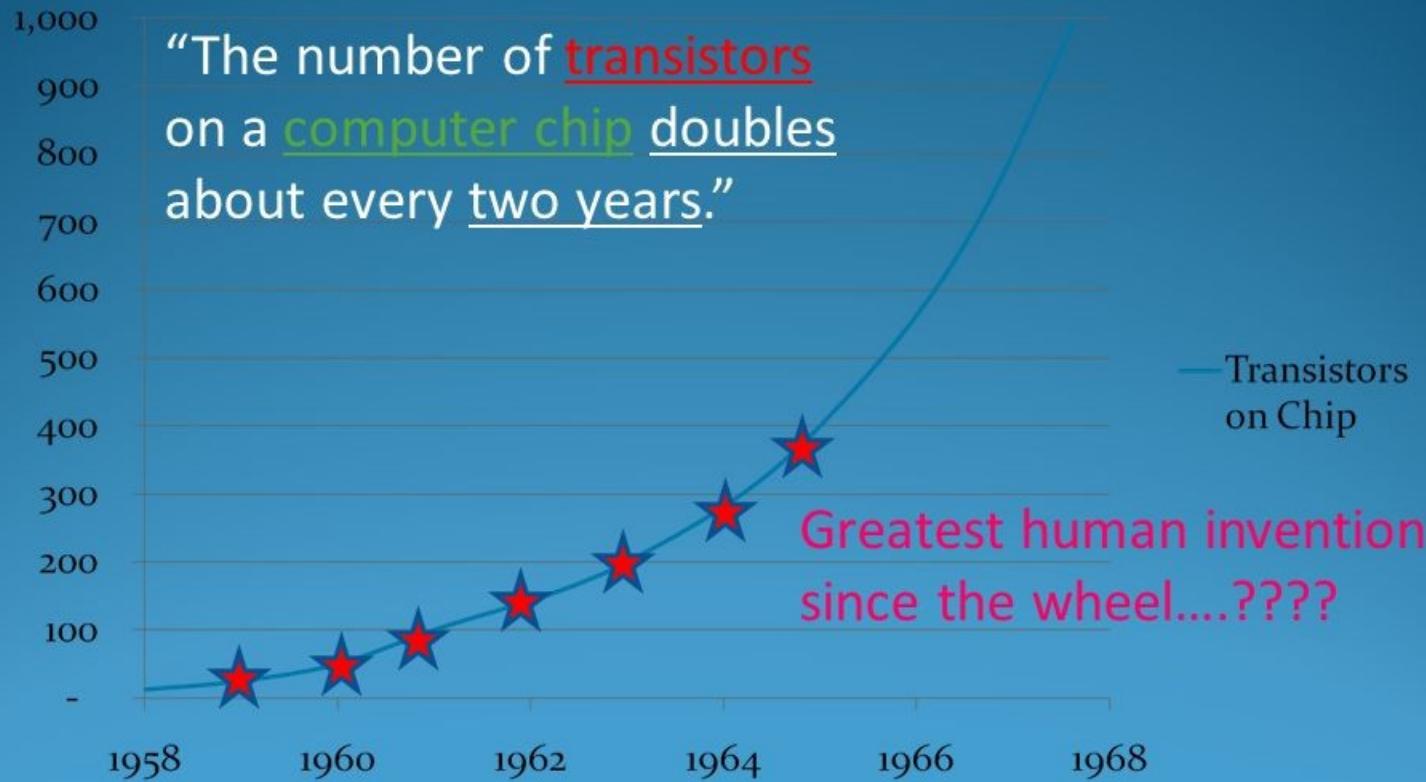


Elemento clave: el transistor

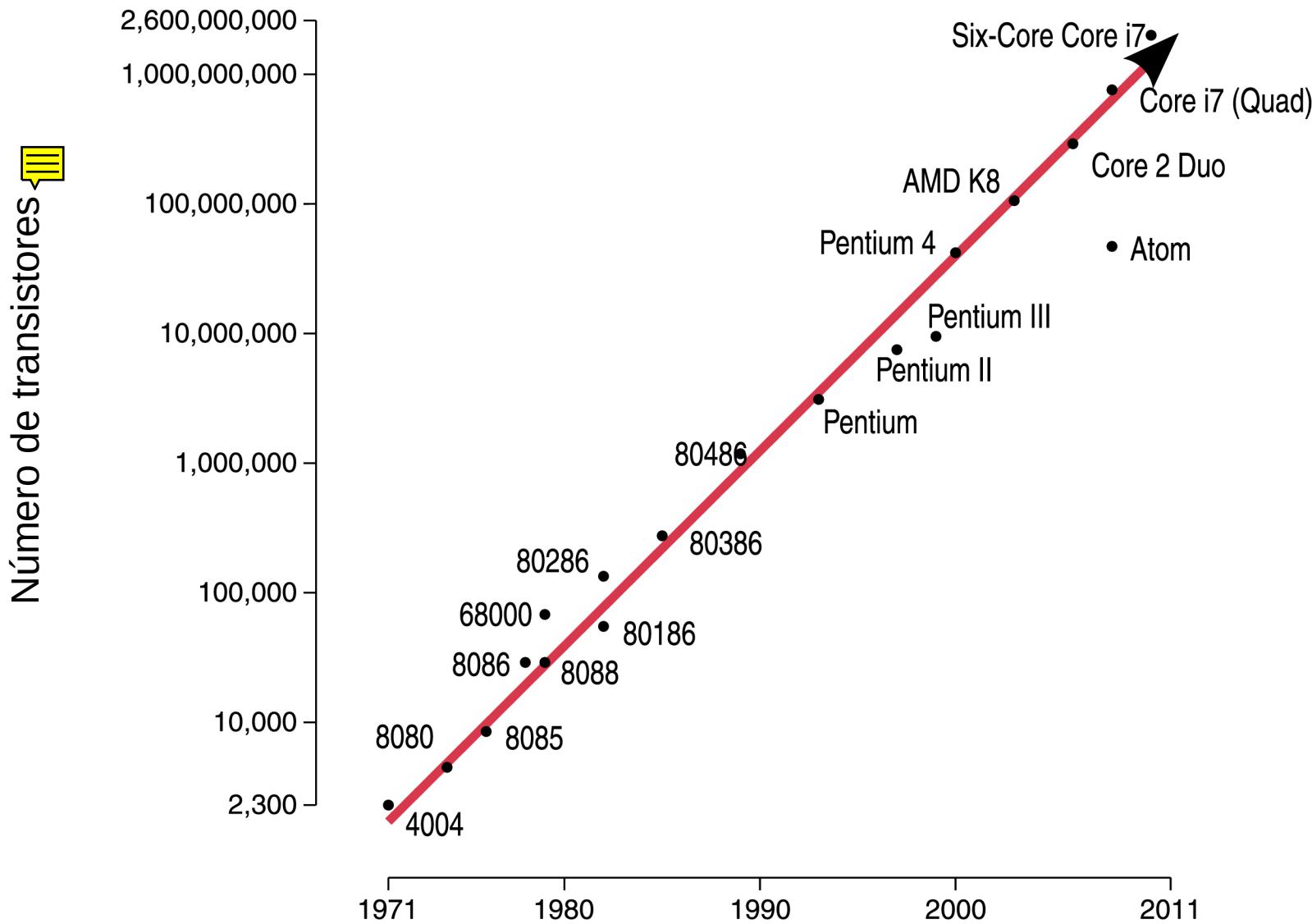
La integración de un número cada vez mayor de transistores ha permitido avances sin precedentes



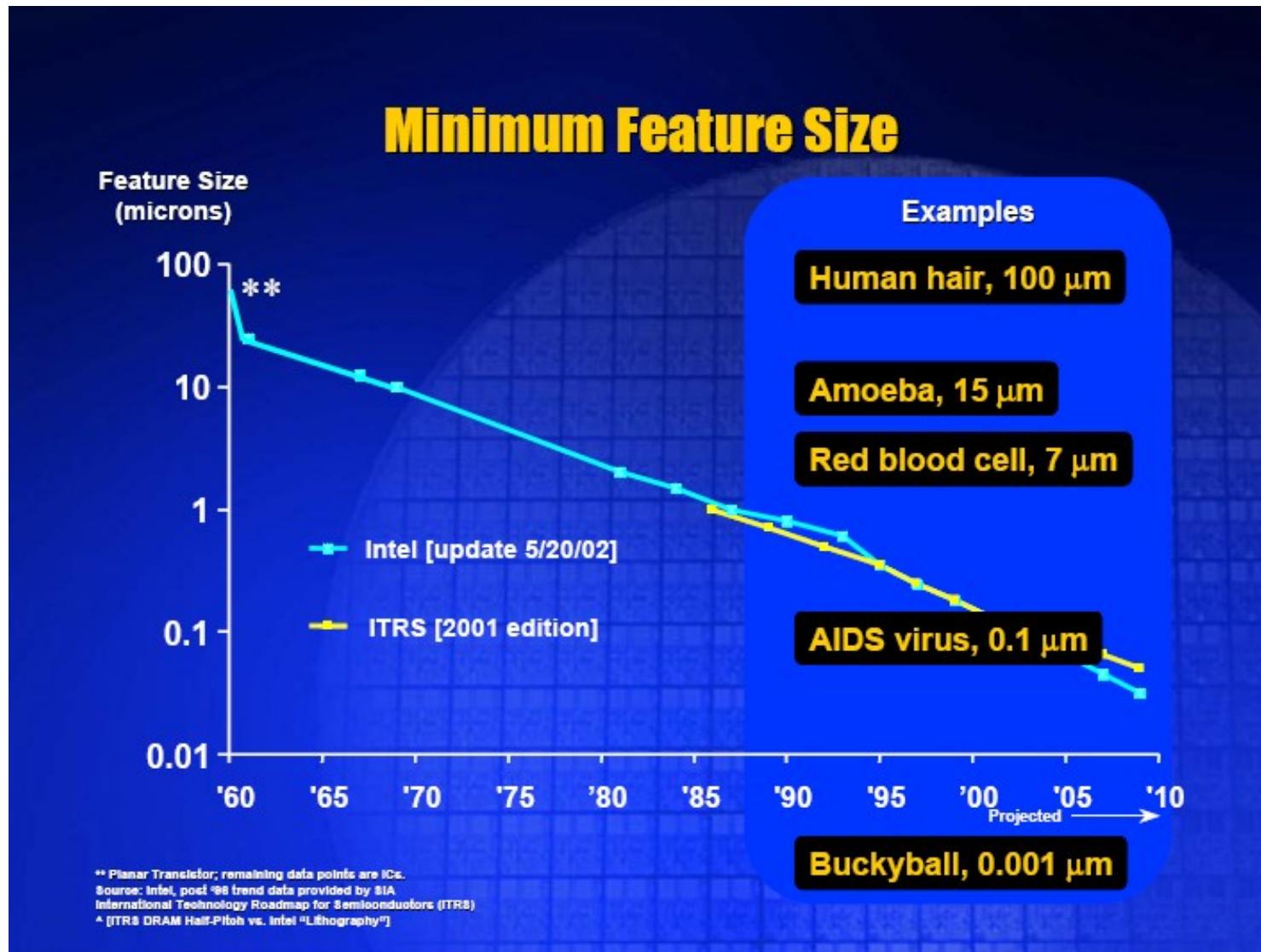
Gordon Moore's Graph – 1960's



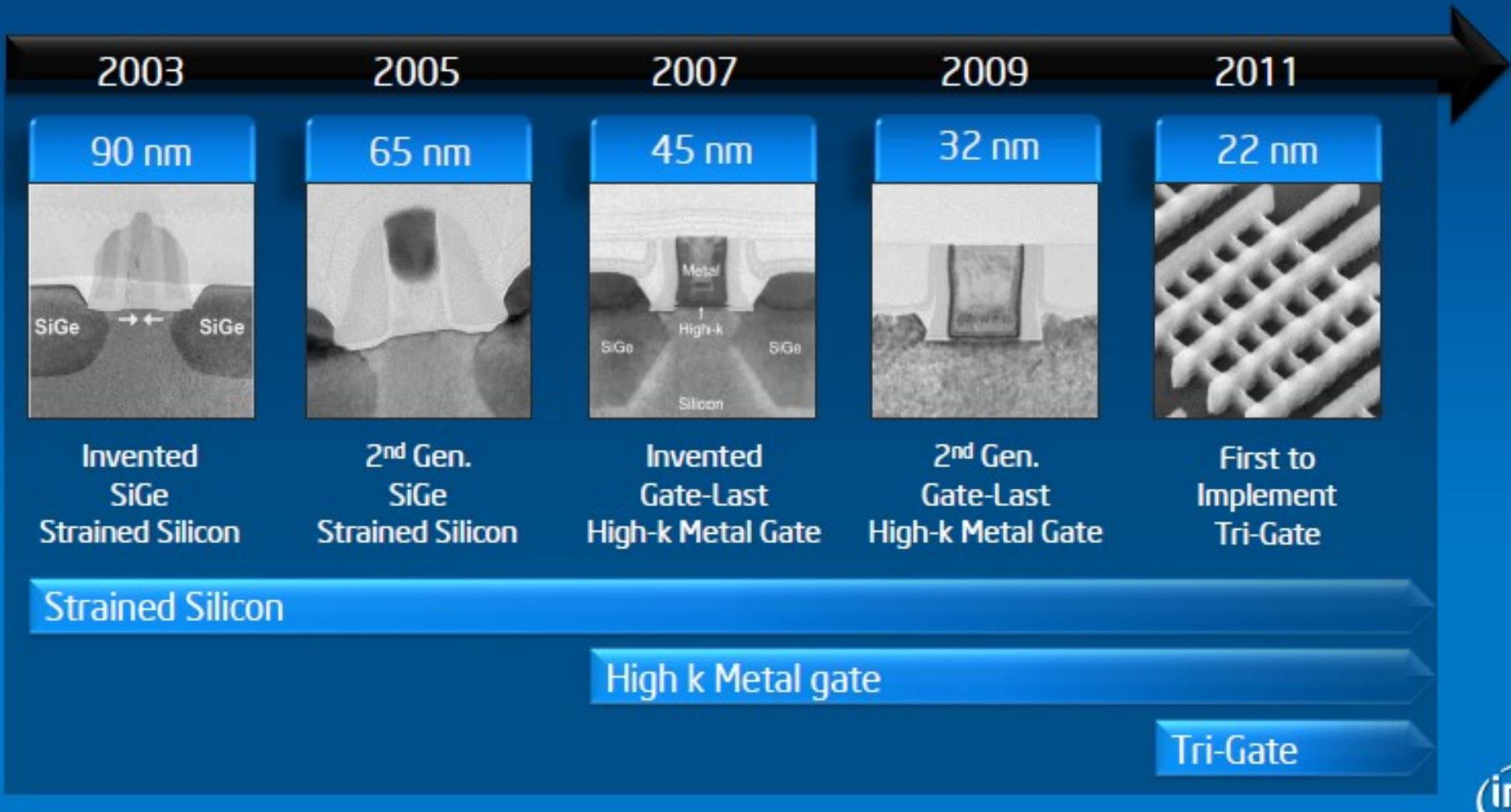
Ley de Moore



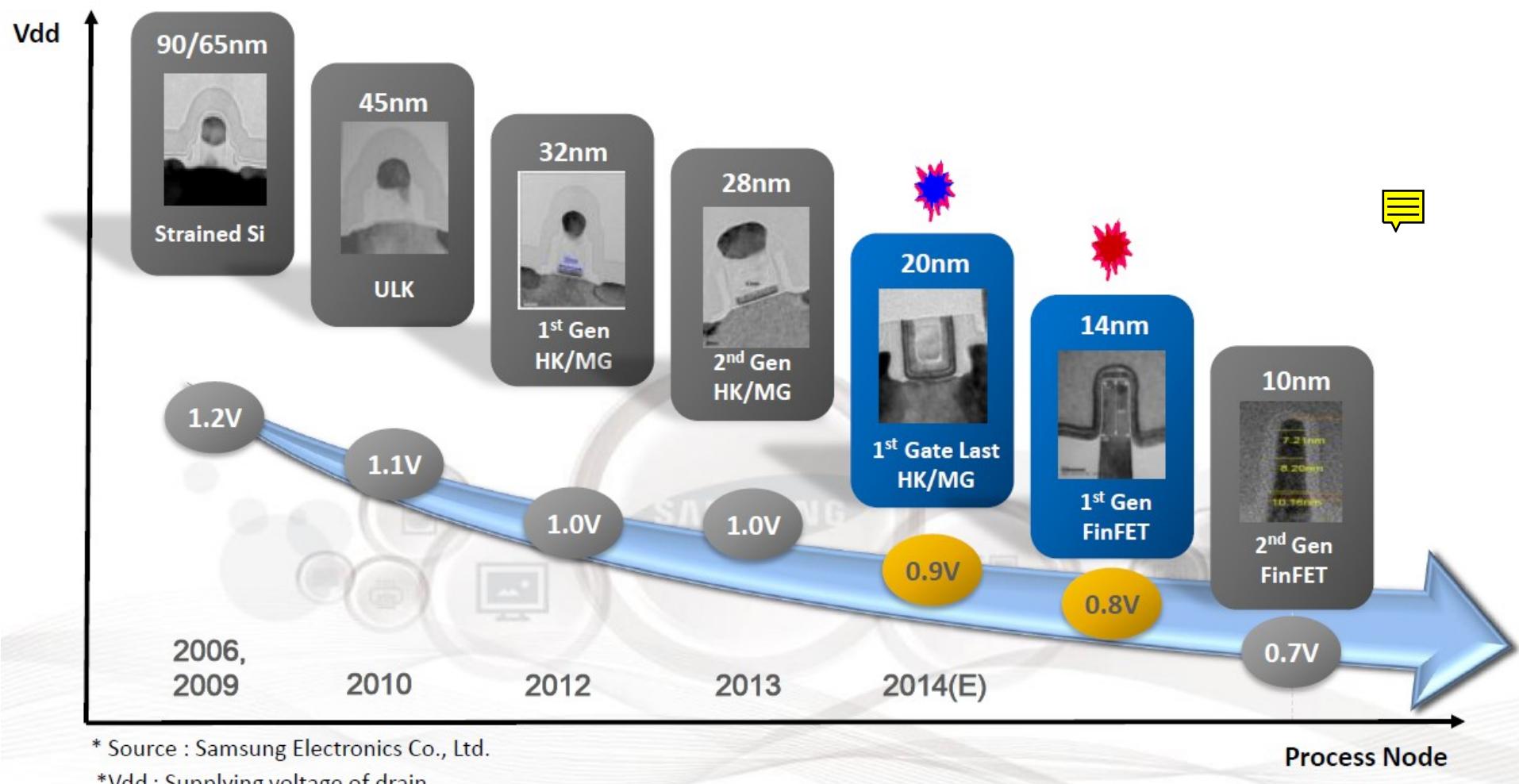
- La ley de Moore no es una ley física
- Es una observación de una tendencia durante los primeros años de fabricación de circuitos integrados
- Los fabricantes lo utilizan como referencia y hacen lo posible para que se cumpla



Transistor Innovations Enable Technology Cadence

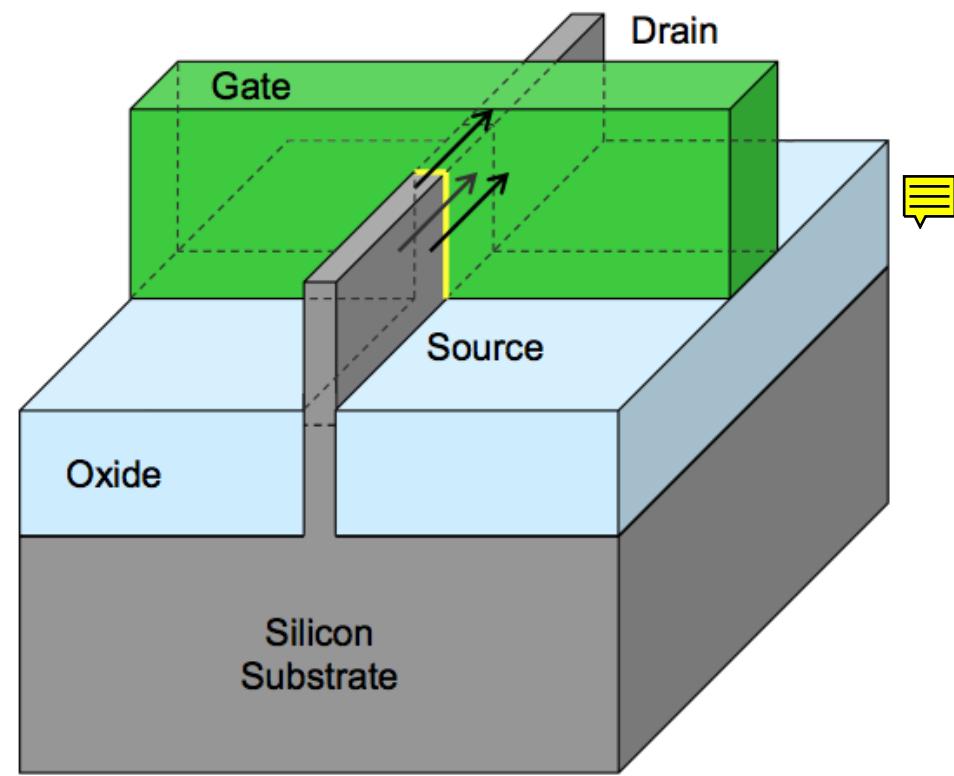
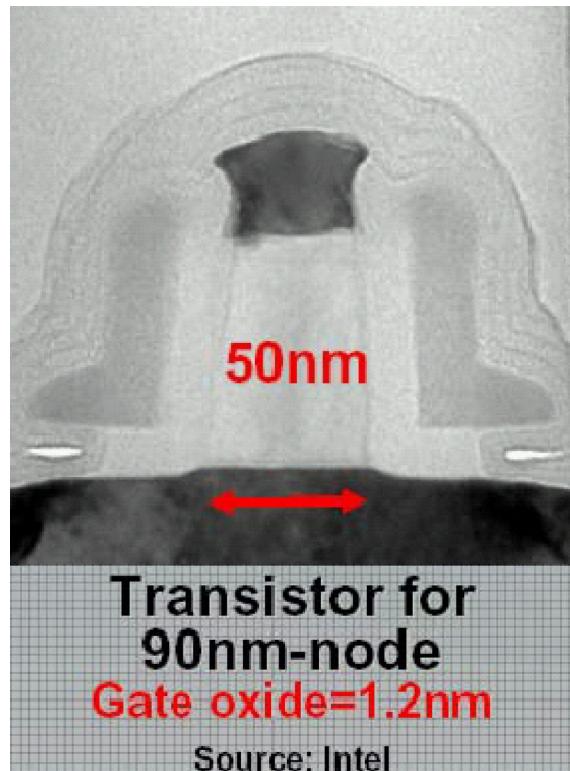


Hoja de ruta

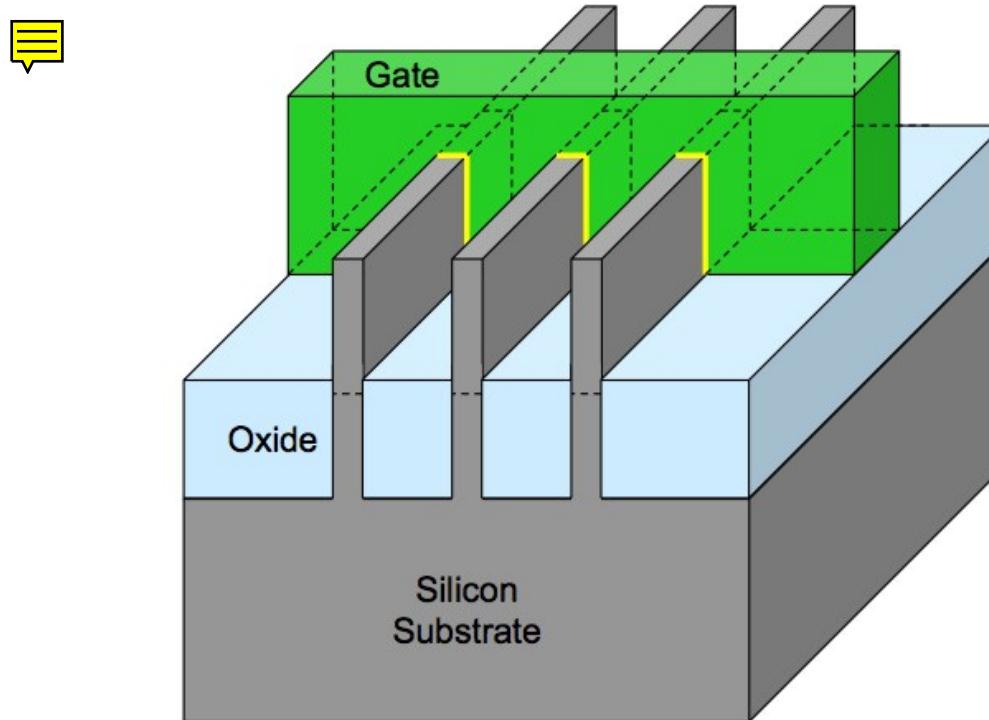


Dificultades tecnológicas

Pero a medida que se reduce su tamaño, cada vez es más difícil diseñar transistores eficientes

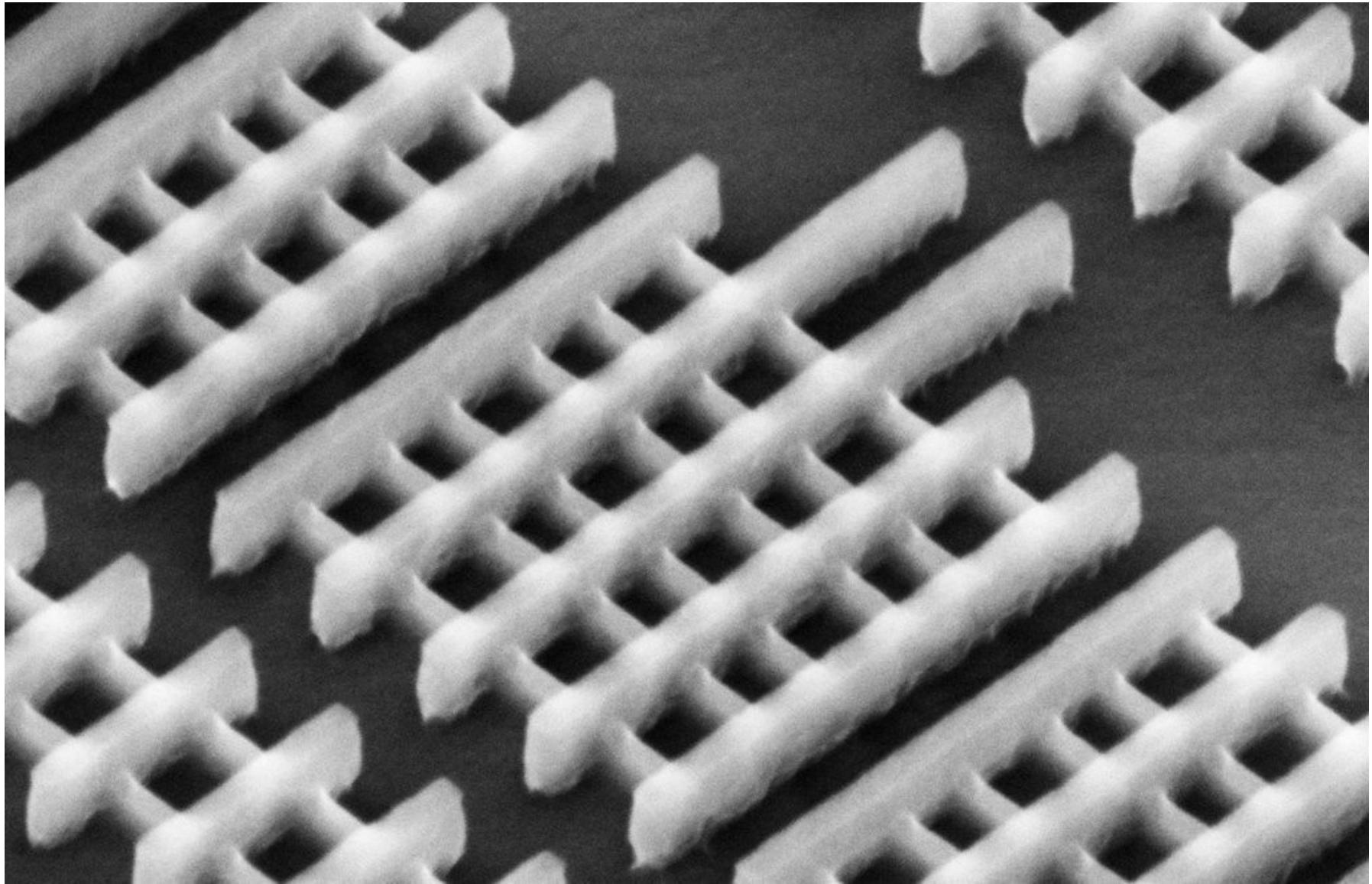


22 nm Tri-Gate Transistor



Tri-Gate transistors can have multiple fins connected together to increase total drive strength for higher performance

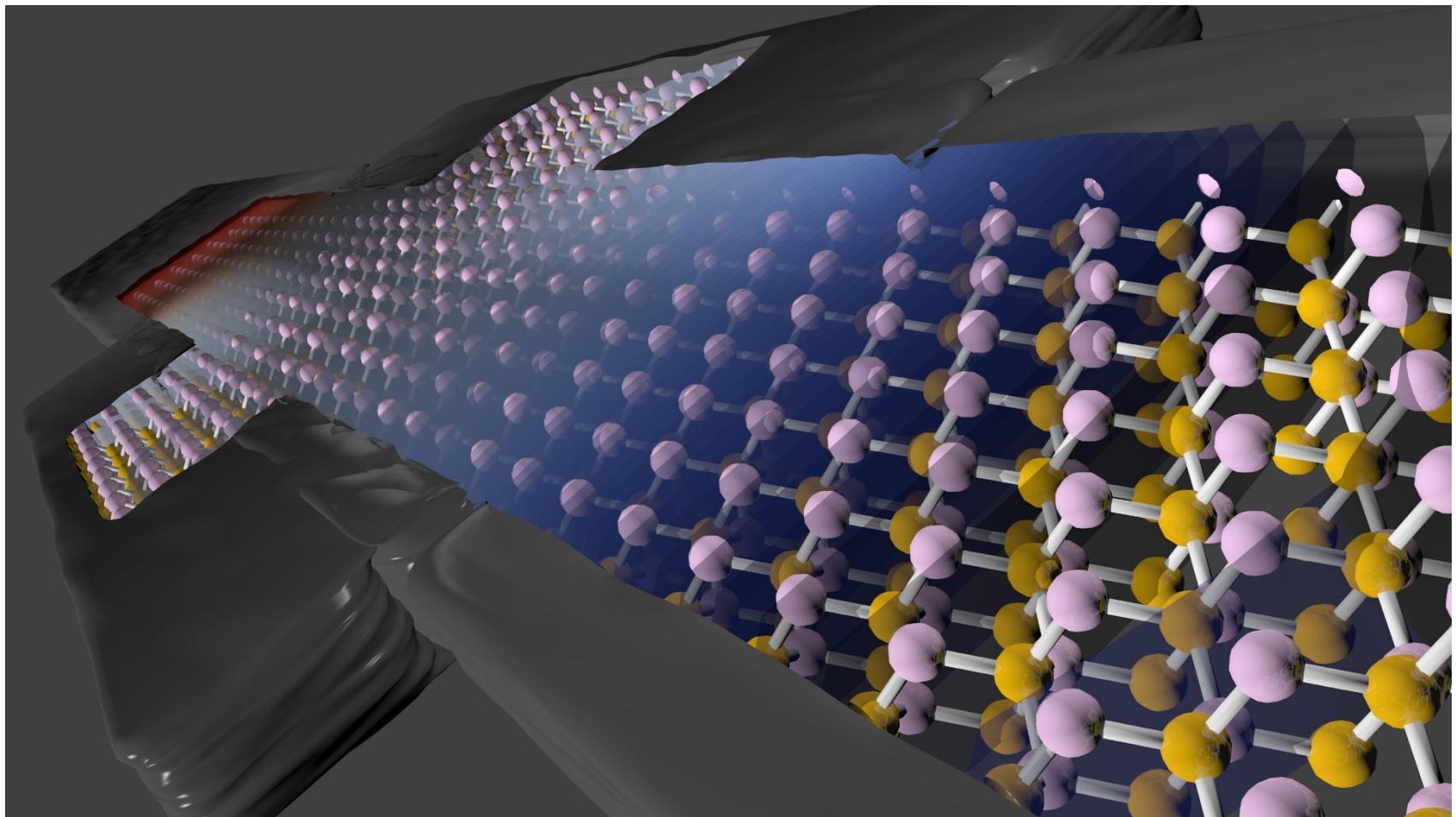
Dificultades tecnológicas



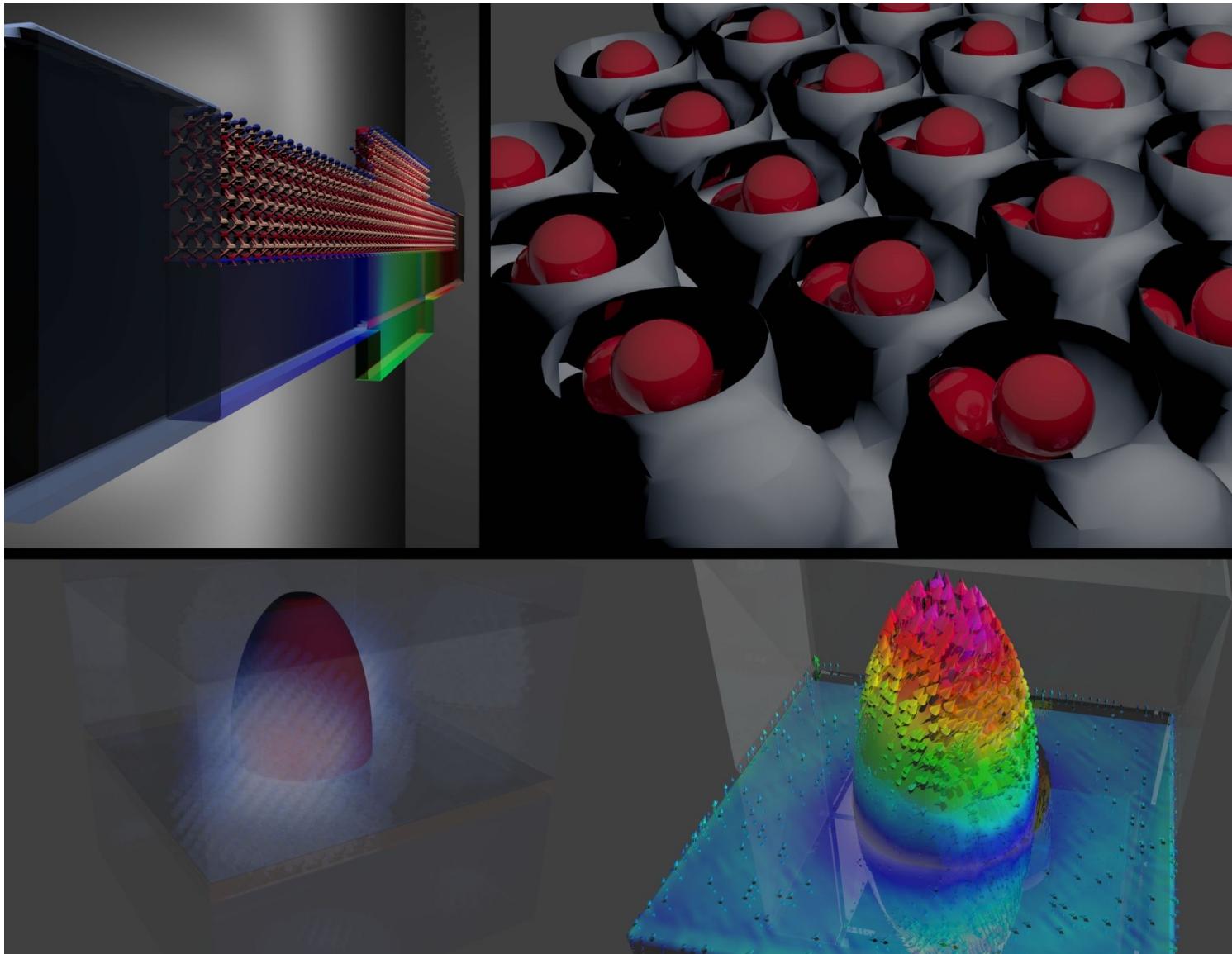
Dificultades tecnológicas

- La ITRS (International Technology Roadmap for Semiconductors) identifica los retos y necesidades de la industria de los semiconductores durante los próximos 15 años
- La ITRS pronostica que los transistores prodrán hacerse más pequeños durante ocho años más, alcanzando los 5 nm
- Más allá es muy difícil predecir porque aparecen efectos cuánticos
- Se utilizan supercomputadores para modelar los transistores y su comportamiento

NEMO5 (NanoElectronics MOdeling Tools)



NEMO5 (NanoElectronics MOdeling Tools)



Además de modelar transistores de silicio y su comportamiento (incluyendo fenómenos cuánticos), se están estudiando nuevos materiales:

- Arseniuro de indio y antimoníuro de indio
- Grafeno
- Nanotubos de carbono

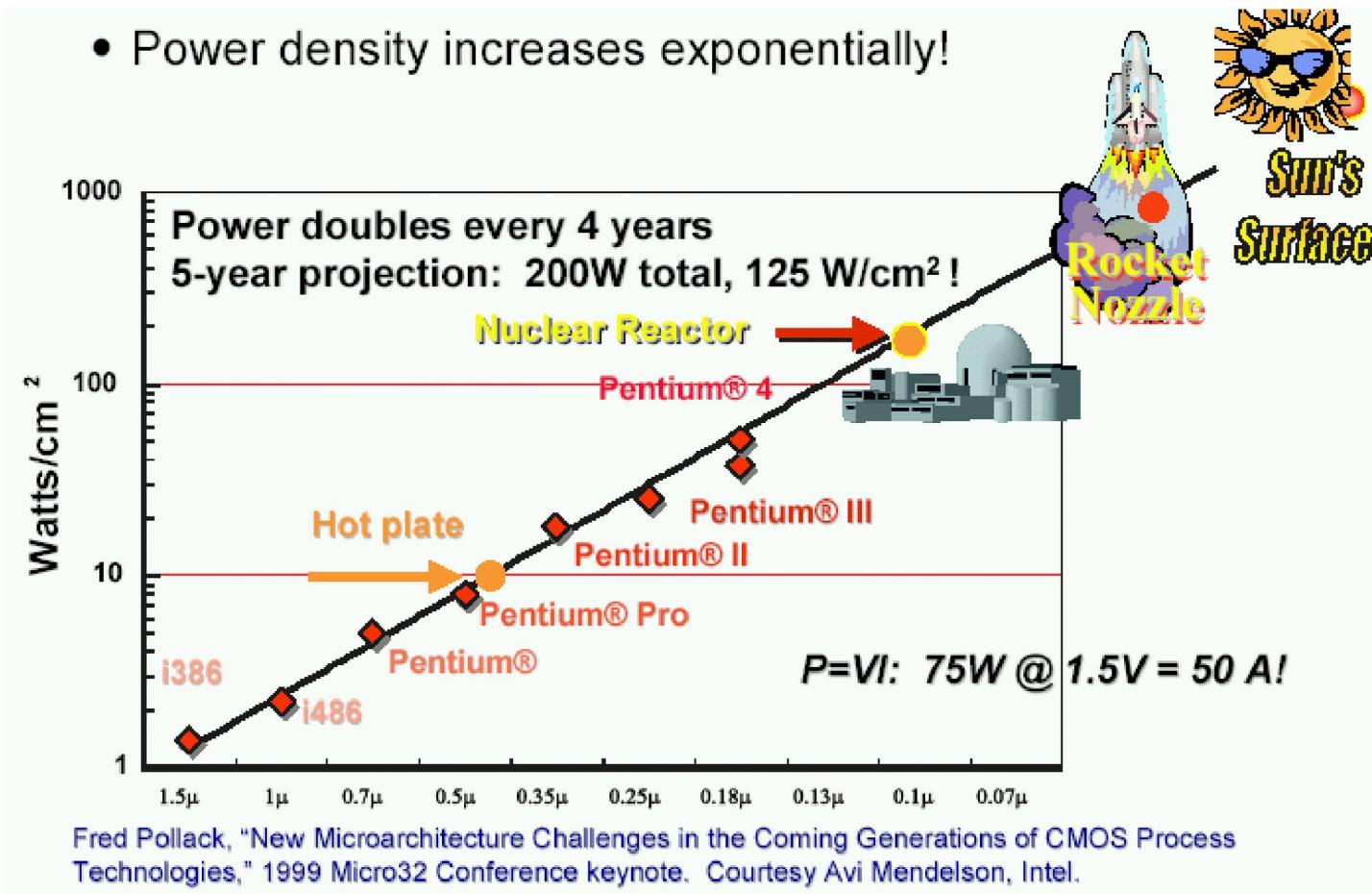


Límites impuestos por el consumo de energía



Límites por disipación de calor

La disipación de calor impone límites a la frecuencia máxima de reloj que se puede emplear

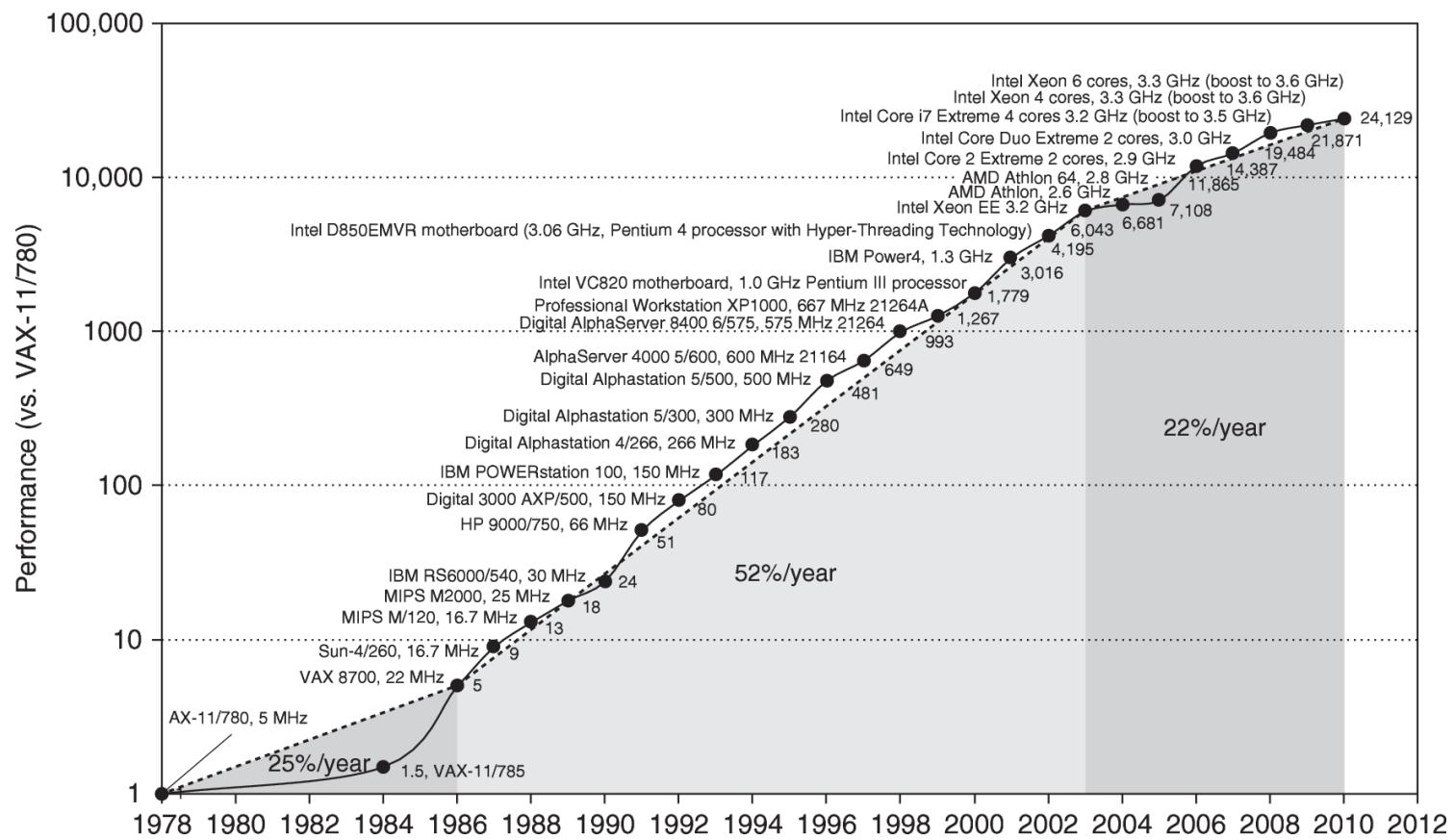


Límites por disipación de calor

En 2003 la frecuencia de reloj tocó techo.

Solución: procesadores multinúcleo.

Pero las mejoras cayeron de 52% a 22% por año.



Límites por disipación de calor

Se pasó de 1 a 8 núcleos en 7 años. Hoy, 7 años más tarde, los procesadores deberían tener 64 núcleos. Pero sólo tienen hasta 28 núcleos.



The screenshot shows the Intel Product Specifications page for Xeon Scalable Processors. The top navigation bar includes 'Products', 'Learn & Develop', 'Support', the Intel logo, 'USA (English)', and a search icon. Below the navigation, the breadcrumb path is 'Support Home > Product Specifications > Processors'. A search bar is also present. The main content area features a large circular icon of a microchip and the text 'Intel® Xeon® Scalable Processors'. A filter bar below the icon allows 'View All', 'Embedded', or 'Server' options. The main table lists five processor models:

Product Name	Status	Launch Date	# of Cores	Max Turbo Frequency	Processor Base Frequency	Cache	Compare All None
Intel® Xeon® Platinum 8180M Processor	Launched	Q3'17	28	3.80 GHz	2.50 GHz	38.5 MB L3	<input type="checkbox"/>
Intel® Xeon® Platinum 8180 Processor	Launched	Q3'17	28	3.80 GHz	2.50 GHz	38.5 MB L3	<input type="checkbox"/>
Intel® Xeon® Platinum 8176M Processor	Launched	Q3'17	28	3.80 GHz	2.10 GHz	38.5 MB L3	<input type="checkbox"/>
Intel® Xeon® Platinum 8176F Processor	Launched	Q3'17	28	3.80 GHz	2.10 GHz	38.5 MB L3	<input type="checkbox"/>
Intel® Xeon® Platinum 8176 Processor	Launched	Q3'17	28	3.80 GHz	2.10 GHz	38.5 MB L3	<input type="checkbox"/>

Límites por disipación de calor

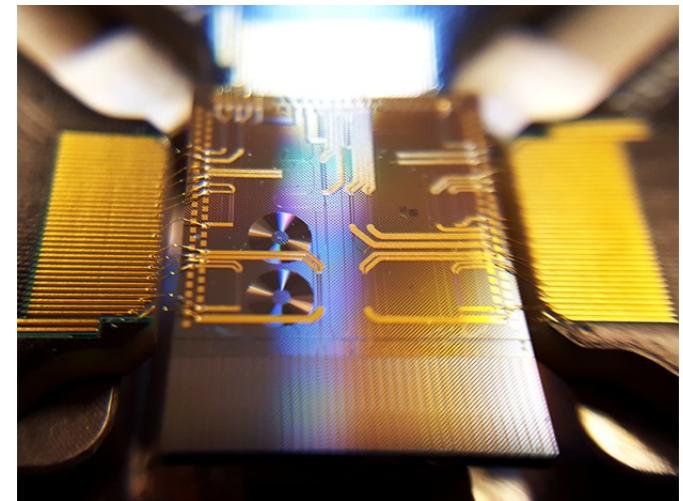


Según un famoso estudio publicado en 2011:

- Número de núcleos limitado por la disipación de calor
- Con 22 nm, 21% del chip debe mantenerse apagado
- Con 8 nm, esta cifra crece por encima del 50%
- Desde 2011 hasta 2024, solo se podrá conseguir una mejora en prestaciones de 8X en las aplicaciones paralelas habituales, lo cual está 24 veces por debajo del objetivo de duplicar las prestaciones cada nueva generación (dos años).

H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, D. Burger, Dark Silicon and the End of Multicore Scaling, ISCA'11

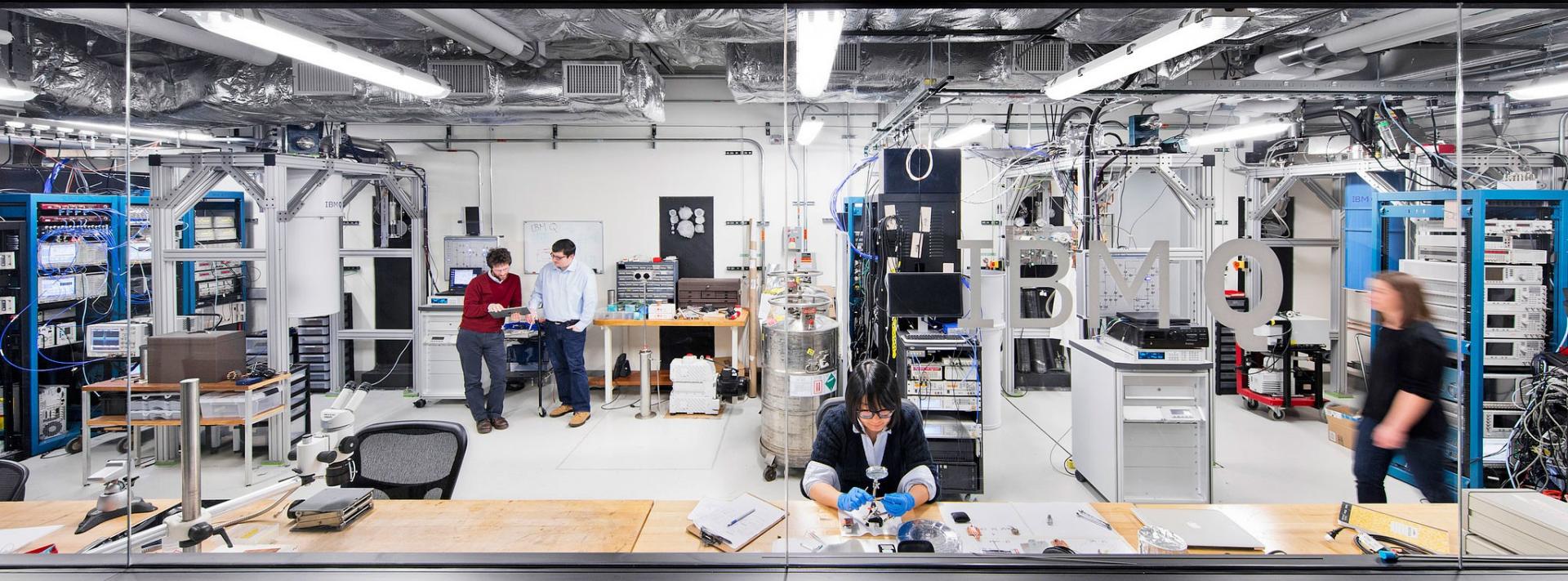
**¿Existe alguna
alternativa
tecnológica?**



Computadores cuánticos

En la actualidad, los computadores cuánticos no son todavía una alternativa viable

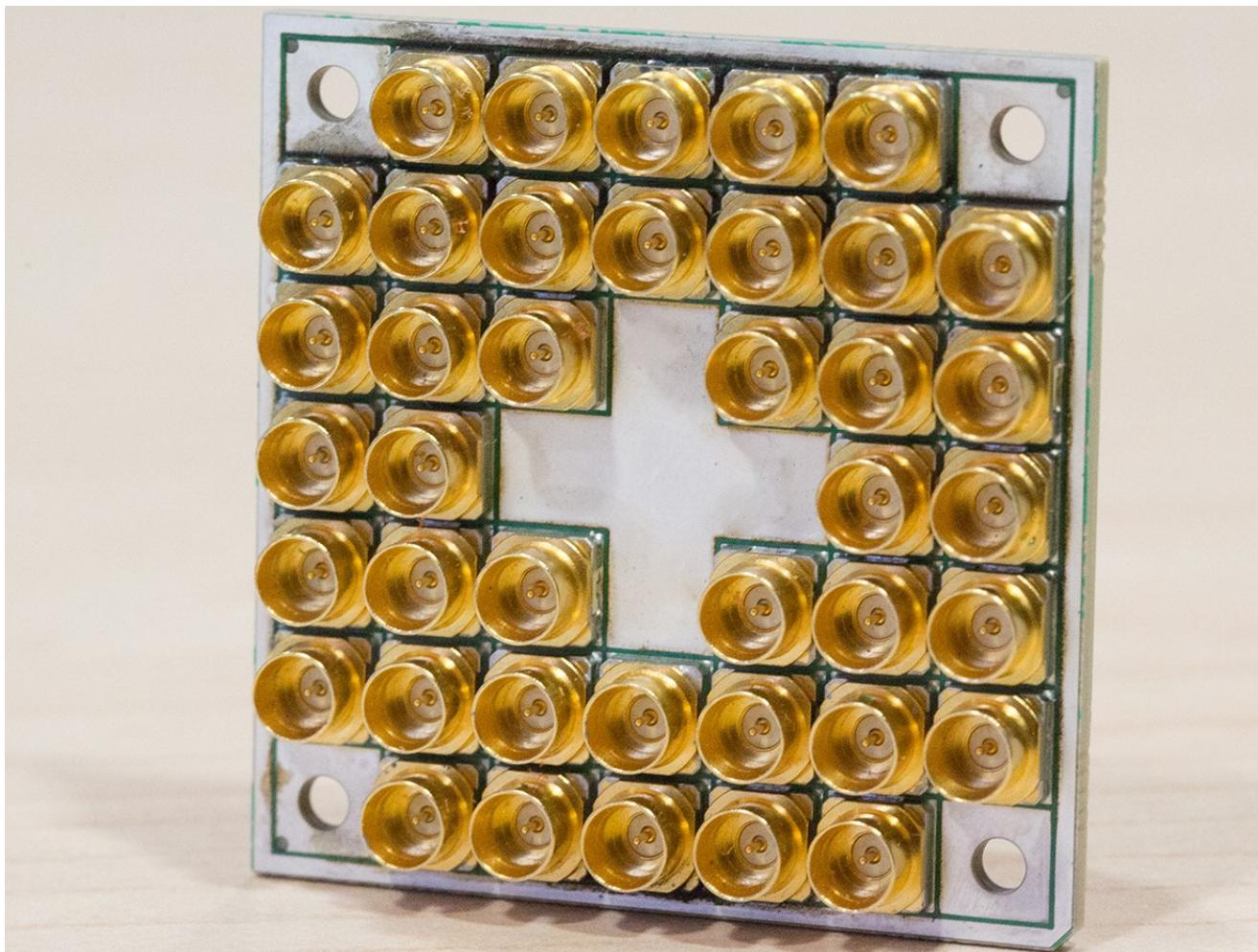
- IBM proporciona acceso a un computador cuántico con 5 qubits (IBM Q – Quantum Experience)
- Los qubits tienen una fiabilidad muy baja y hacen falta muchos qubits físicos para formar un qubit lógico suficientemente fiable (Surface code)
- Los dispositivos actuales tienen que trabajar a temperaturas cercanas a 0 K
- Sólo aportan grandes ventajas en algoritmos que aprovechan la superposición cuántica (alg. de Shor)



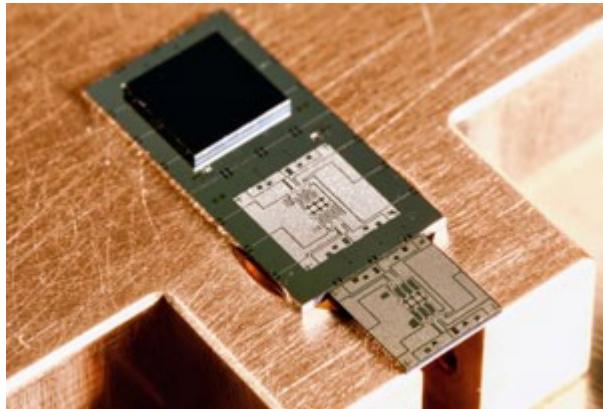
Los computadores cuánticos actuales no son la panacea. Tienen serias limitaciones:

- 1973: Alexander Holevo muestra que n qubits no pueden almacenar más de n bits de información
- 1981: Richard Feynman indica que parecía ser imposible simular la evolución de un sistema cuántico en un computador clásico de una manera eficiente
- 2017: Mikhail Lukin anuncia el simulador más potente de dispositivos cuánticos: puede simular 51 qubits
- 2017: IBM anuncia un computador cuántico con 17 qubits. Intel anuncia un chip con 17 qubits. Google anuncia un computador con 49 qubits

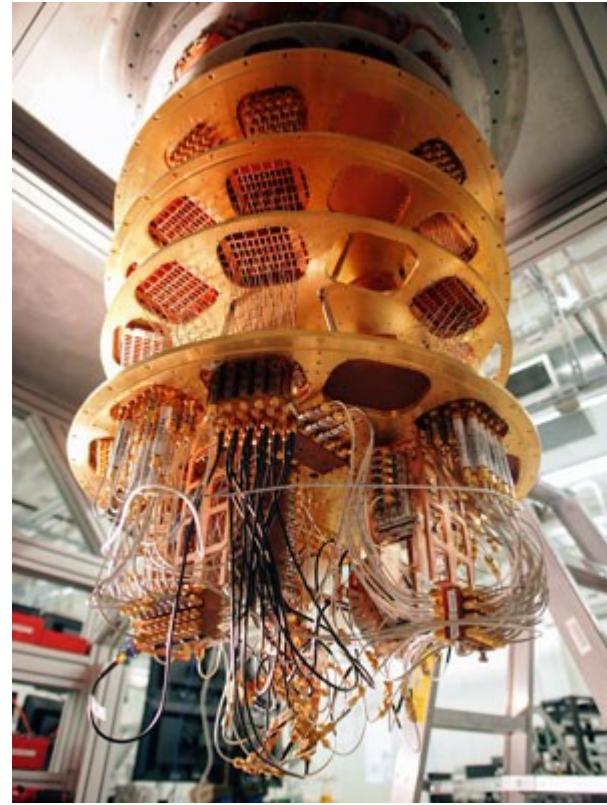
Chip de Intel con 17 qubits



Computador cuántico de Google

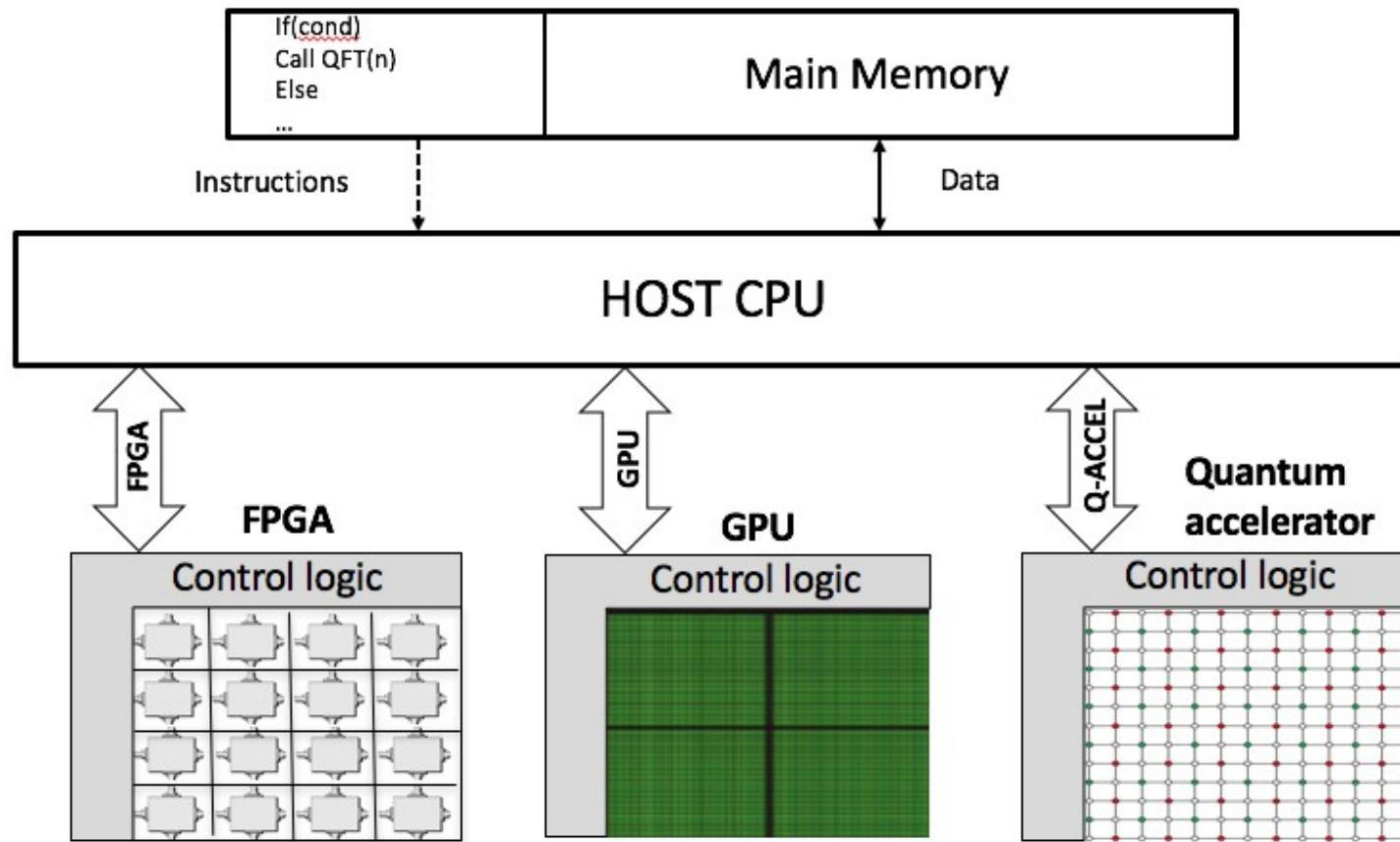


Chip con 2 x 3 qubits



Refrigerador a 10 miliKelvin

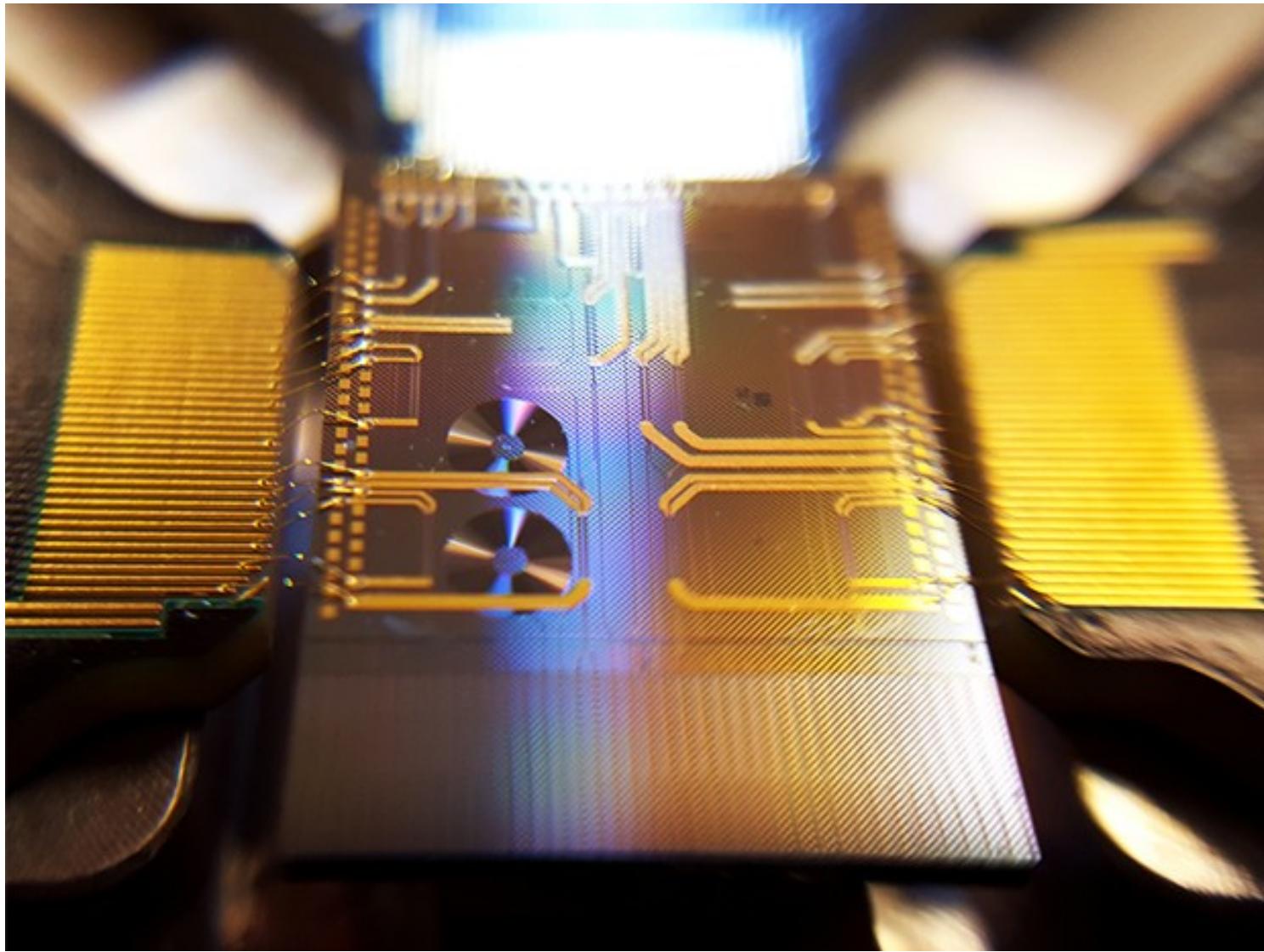
Acelerador accesible en la nube



Pero siempre aparece alguien que piensa de forma diferente, y puede cambiar el futuro:

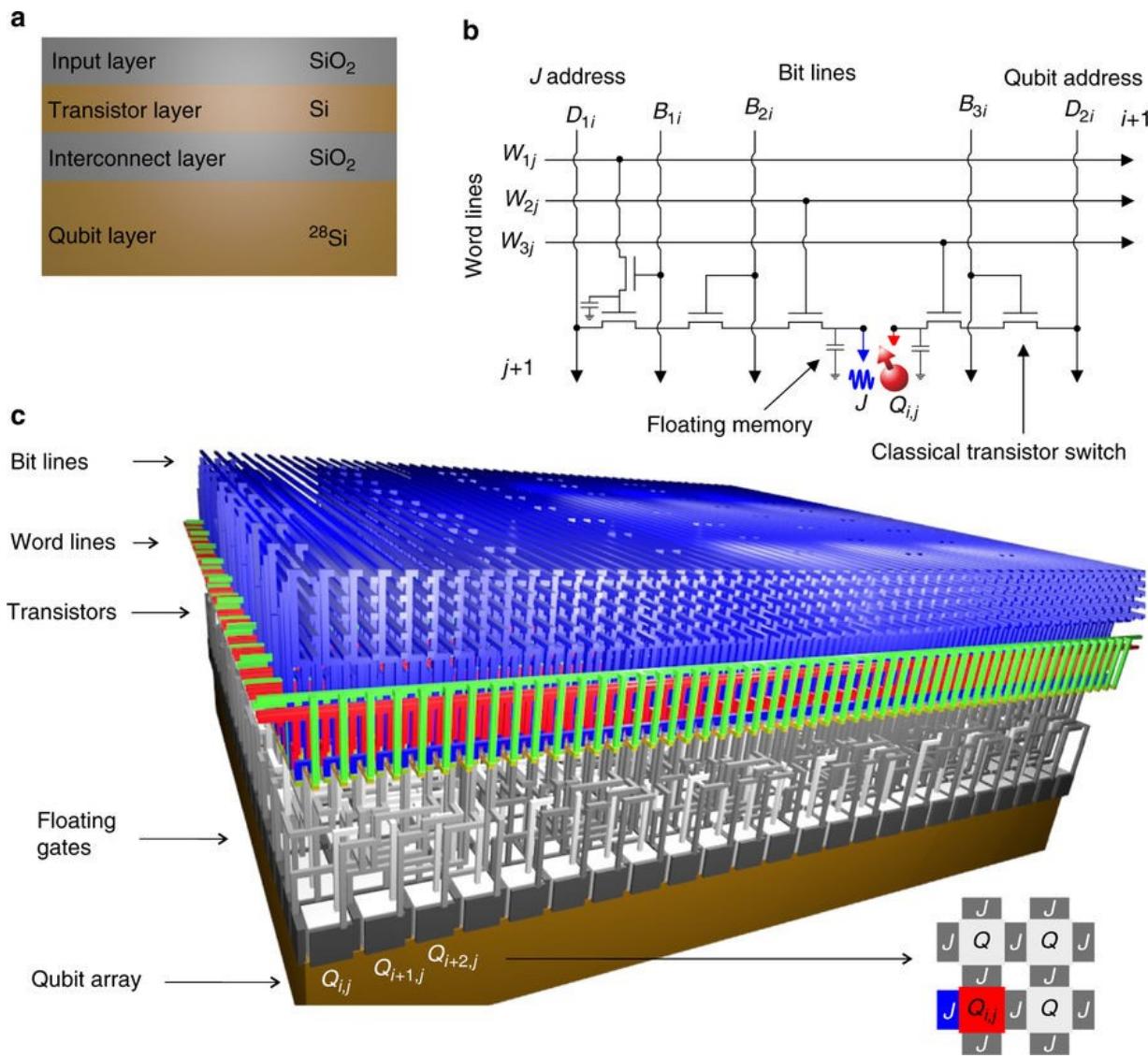
- Michael Kues (INRS, Canadá) ha desarrollado un microchip que puede generar dos qudits (fotones) entrelazados, cada uno con 10 estados
- Han utilizado componentes comerciales para telecomunicaciones. No necesitan trabajar a 0 K
- Han enviado pares de fotones entrelazados a través de 24 Kms de fibra óptica sin que se destruya el entrelazado, demostrando comunicación cuántica
- Creen que qudits con 96 estados son viables. Pero sólo han conseguido entrelazar dos qudits

Chip con dos qudits



- Tamaño
 - Un computador cuántico de iones atrapados en microondas con dos mil millones de qubits requiere un área de más de $100 \times 100 \text{ m}^2$. El mismo número de qubits superconductores requiere un área de $5 \times 5 \text{ m}^2$. Los qubits definidos por los estados de espín de los puntos cuánticos (QD) semiconductores podrían caber en un área de menos de $5 \times 5 \text{ mm}^2$.
- Control
 - Los prototipos actuales requieren acceso a todos y cada uno de los qubits físicos
 - Control por filas y columnas en qubits espín

Arquitectura CMOS para qubits espín



Computadores cuánticos

Así pues, la computación cuántica parece estar aún lejos en el tiempo:

We expect it will be five to seven years before the industry gets to tackling engineering-scale problems, and it will likely require 1 million or more qubits to achieve commercial relevance.

Mike Mayberry

Corporate Vice President
and Managing Director of Intel Labs

Computadores cuánticos

Así pues, la computación cuántica parece estar aún lejos en el tiempo:

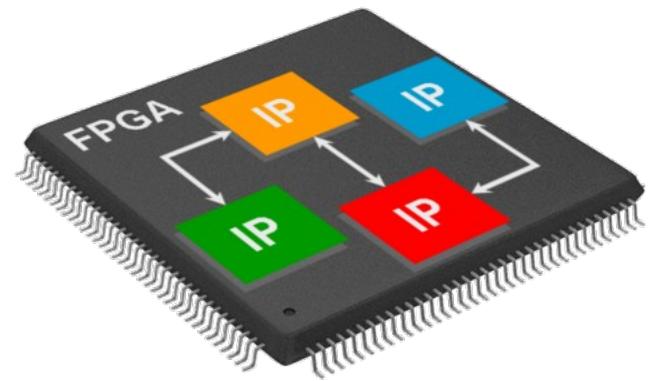
The most promising routes towards large-scale universal quantum computing all require quantum error correction (QEC), such as two dimensional surface code. These approaches also require a platform that can be scaled up to very large numbers of qubits, of order 10^8 .

M. Veldhorst, H.G.J. Eenink, C.H. Yang & A.S. Dzurak

Silicon CMOS architecture for a spin-based quantum computer

Nature Communications 8, Article number: 1766 (2017)

Mejorando la eficiencia energética: Aceleradores



- Existen procesadores especializados que ejecutan ciertas tareas repetitivas más rápido y son más eficientes desde el punto de vista energético que una CPU de última generación:
 - Las tarjetas gráficas controlan muchos operadores con cada unidad de control y se utilizan con frecuencia como aceleradores de cálculo
 - Las FPGAs permiten cambiar el diseño de sus circuitos en una fracción de segundo y se han propuesto muchas veces como aceleradores
 - Se están empezando a fabricar chips específicos para inteligencia artificial (redes neuronales)

Diagrama de la NVIDIA Tesla

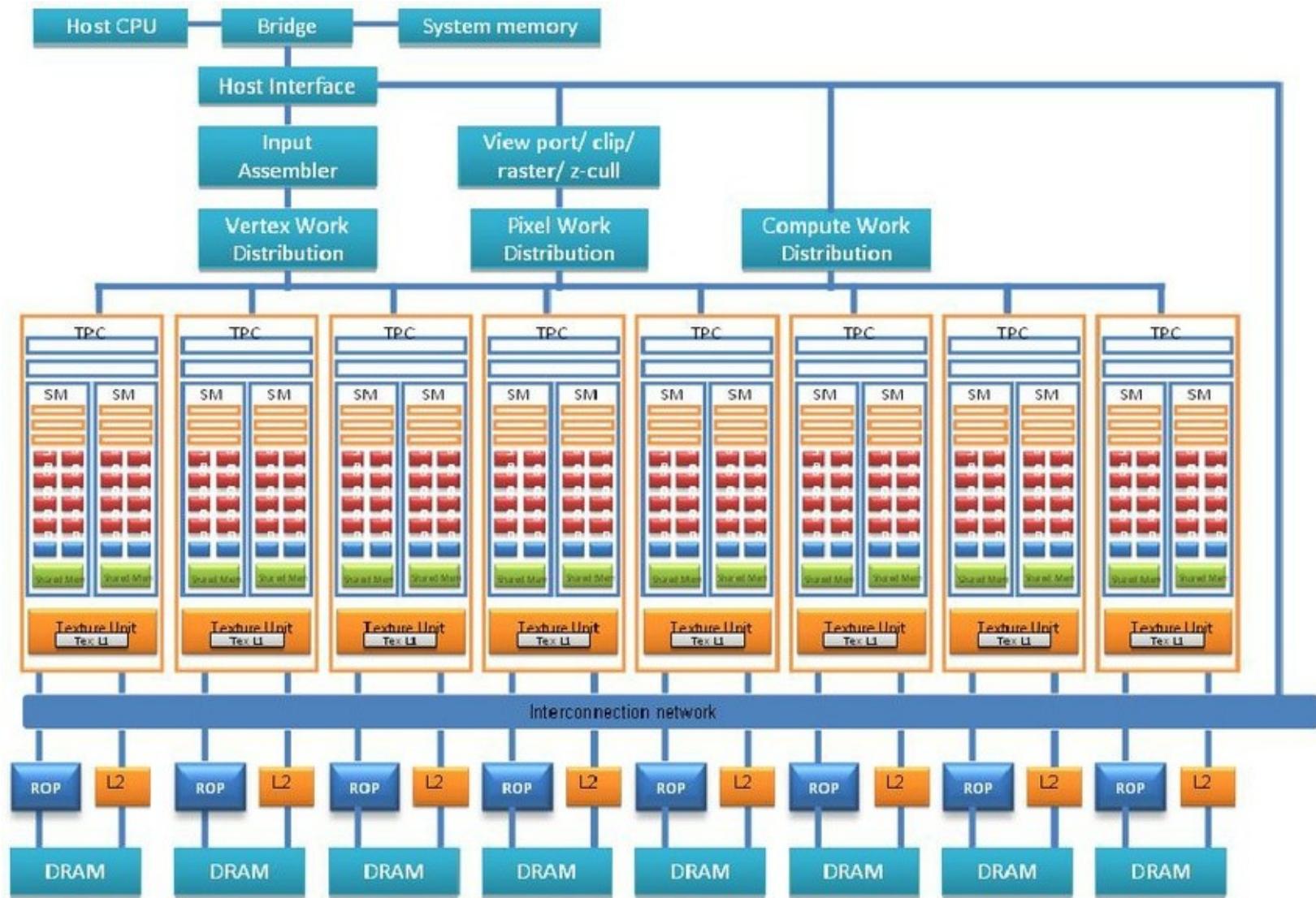
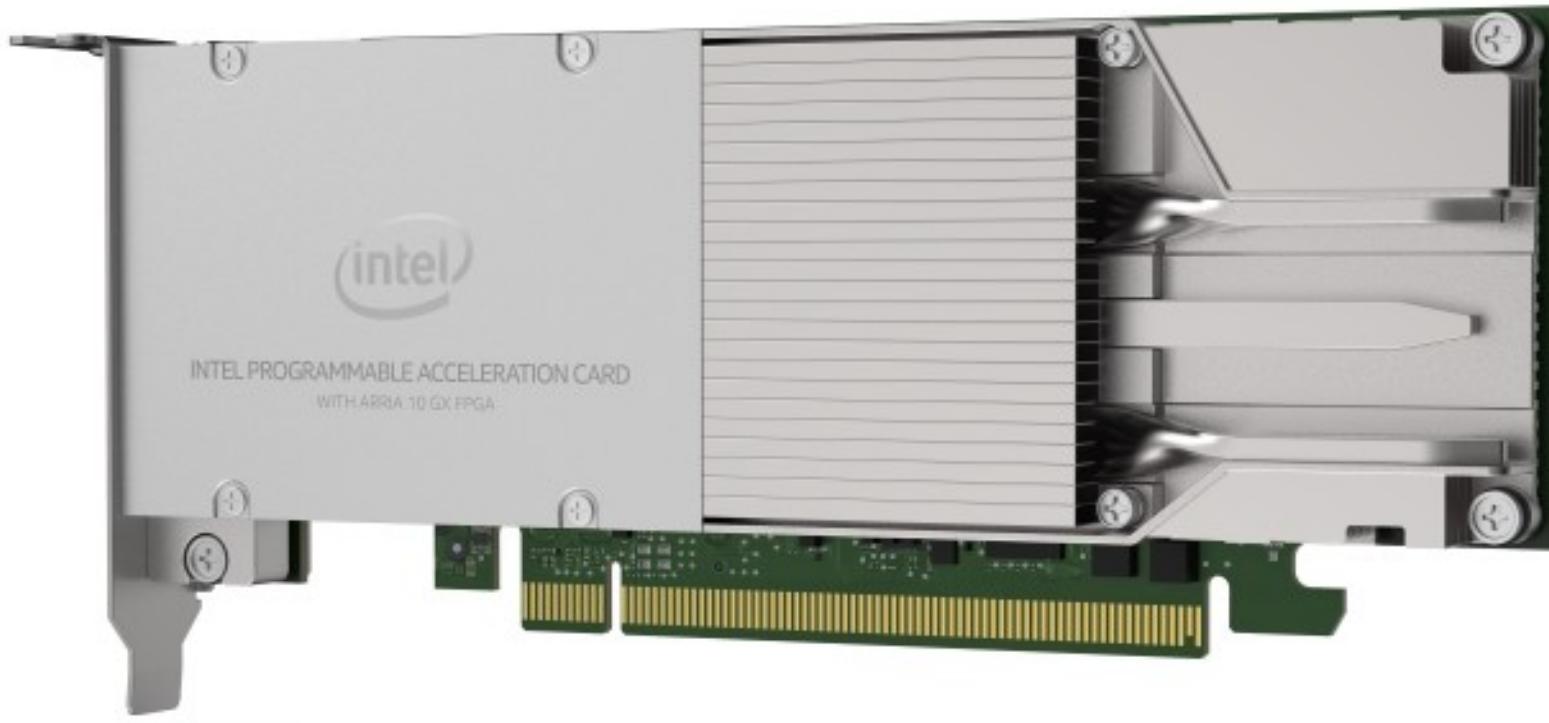


Diagrama de la NVIDIA GP100



- Las GPUs consiguen una potencia de cálculo mucho mayor mediante el uso de miles de unidades de coma flotante en cada chip
- Las GPUs ahorran energía eliminando los circuitos de ejecución de instrucciones fuera de orden y compartiendo la lógica de control entre múltiples unidades de coma flotante

Intel Arria 10 GX FPGA

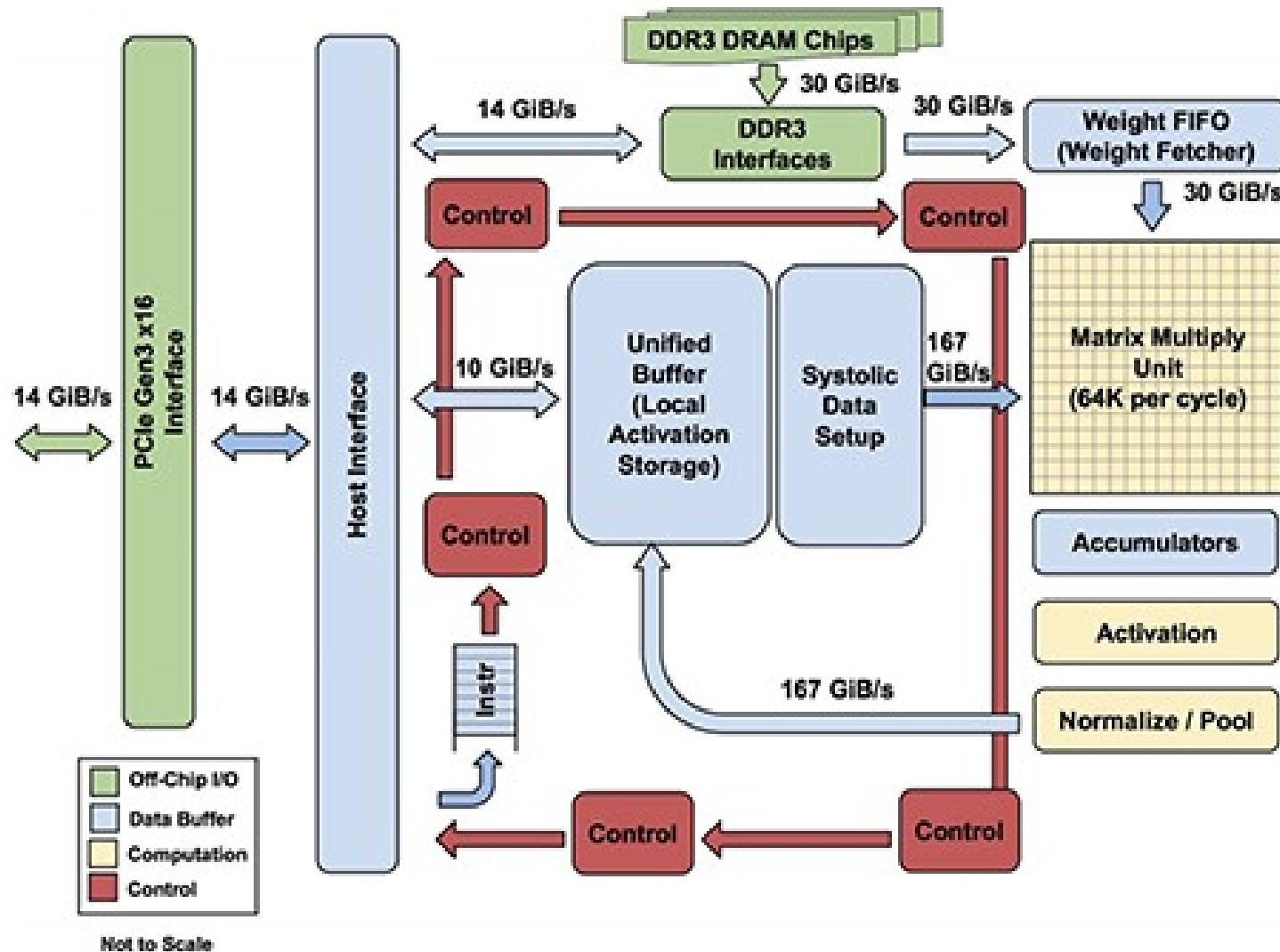


Intel FPGAs Power Acceleration-as-a-Service for Alibaba Cloud

<https://newsroom.intel.com/news/intel-fpgas-power-acceleration-as-a-service-alibaba-cloud/>

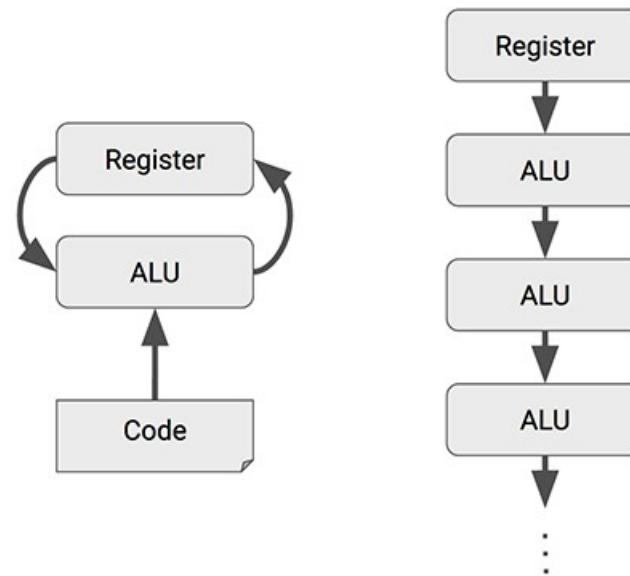
- Las TPUs (Tensor Processing Units) y las DPUs (Dataflow Processing Units) consiguen una potencia de cálculo mucho mayor mediante el uso de decenas de miles de unidades aritméticas de baja precisión en cada chip
- Las TPUs y DPUs ahorran energía eliminando los circuitos de ejecución de instrucciones fuera de orden, compartiendo la lógica de control entre todas las unidades aritméticas y ejecutando cientos de operaciones sobre cada valor leído de memoria

Diagrama de la TPU de Google



Funcionamiento de la TPU de Google

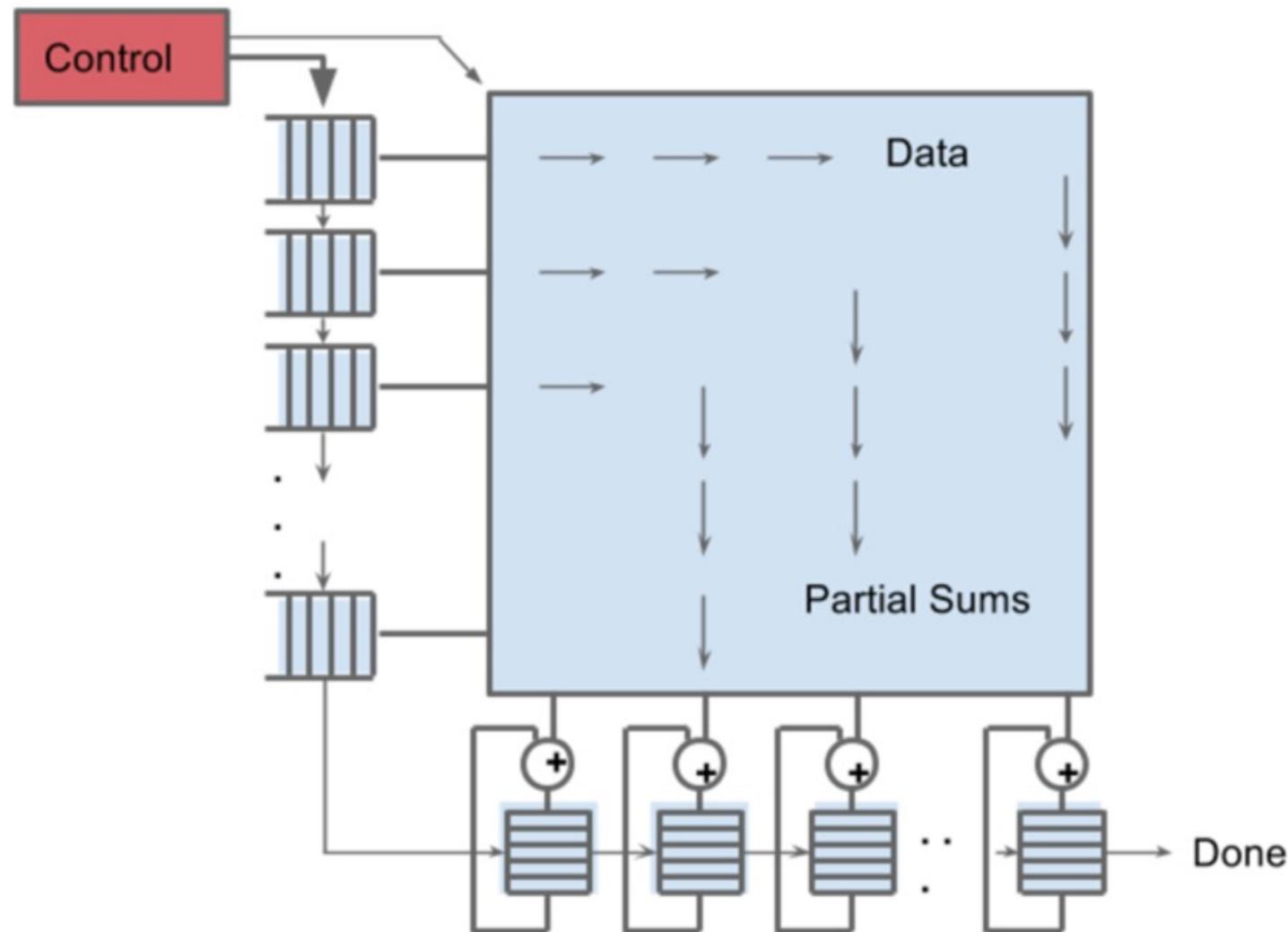
- Cada valor leído de memoria es enviado a cientos de unidades aritméticas, una tras otra, en las que es procesado, ahorrando muchos accesos a memoria



Funcionamiento de la TPU de Google

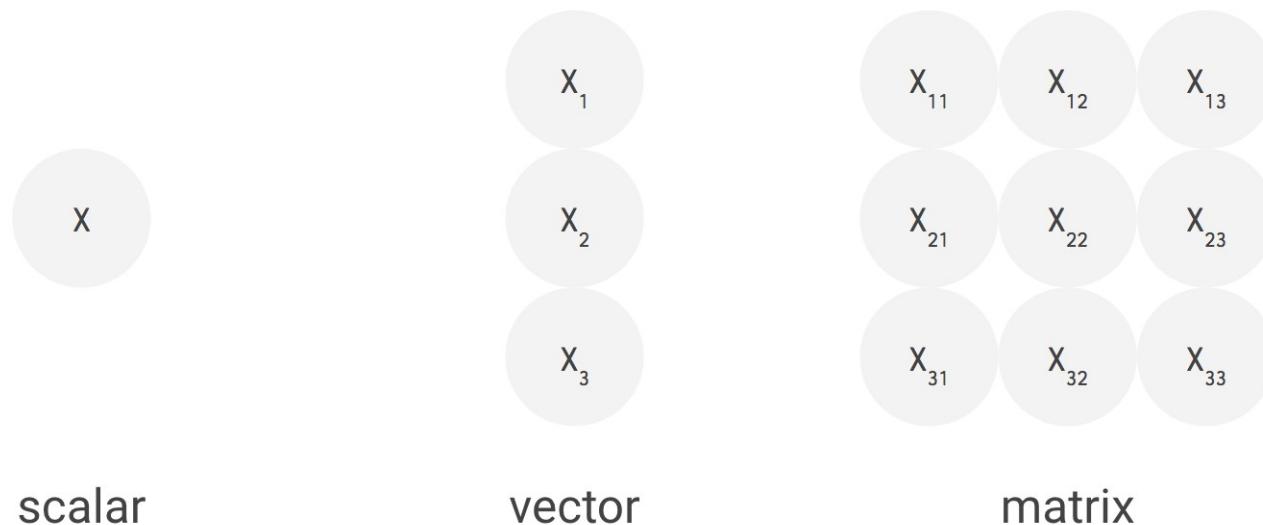
- Se leen cientos de valores de memoria por ciclo, cada uno de los cuales será procesado cientos de veces en sucesivos ciclos
- Las filas y columnas de las matrices a multiplicar se envían en paralelo a la matriz de unidades aritméticas, siendo procesadas a medida que avanzan
- Cada ciclo se inicia una nueva multiplicación de matrices

Funcionamiento de la TPU de Google

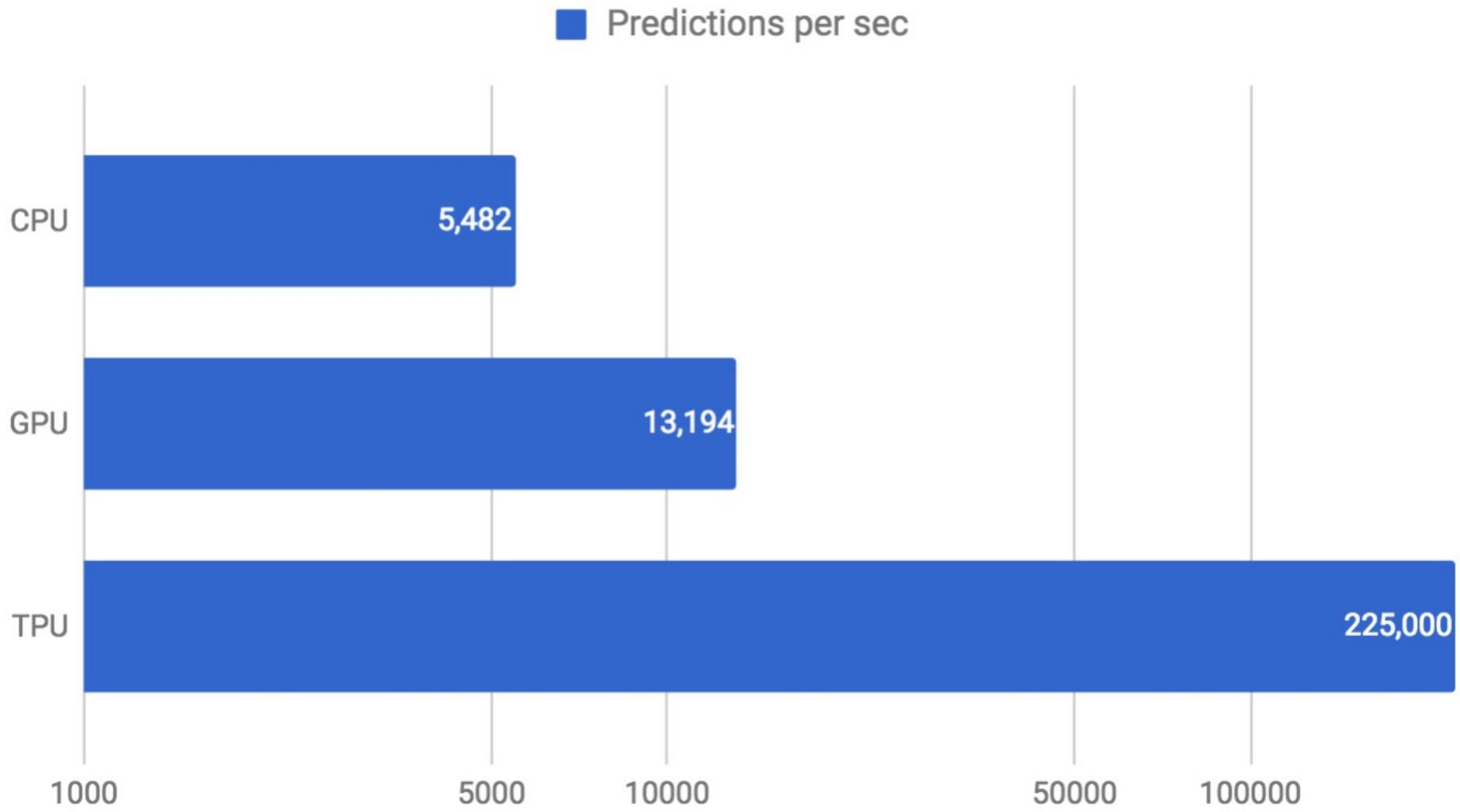


Comparación entre CPU, GPU y TPU

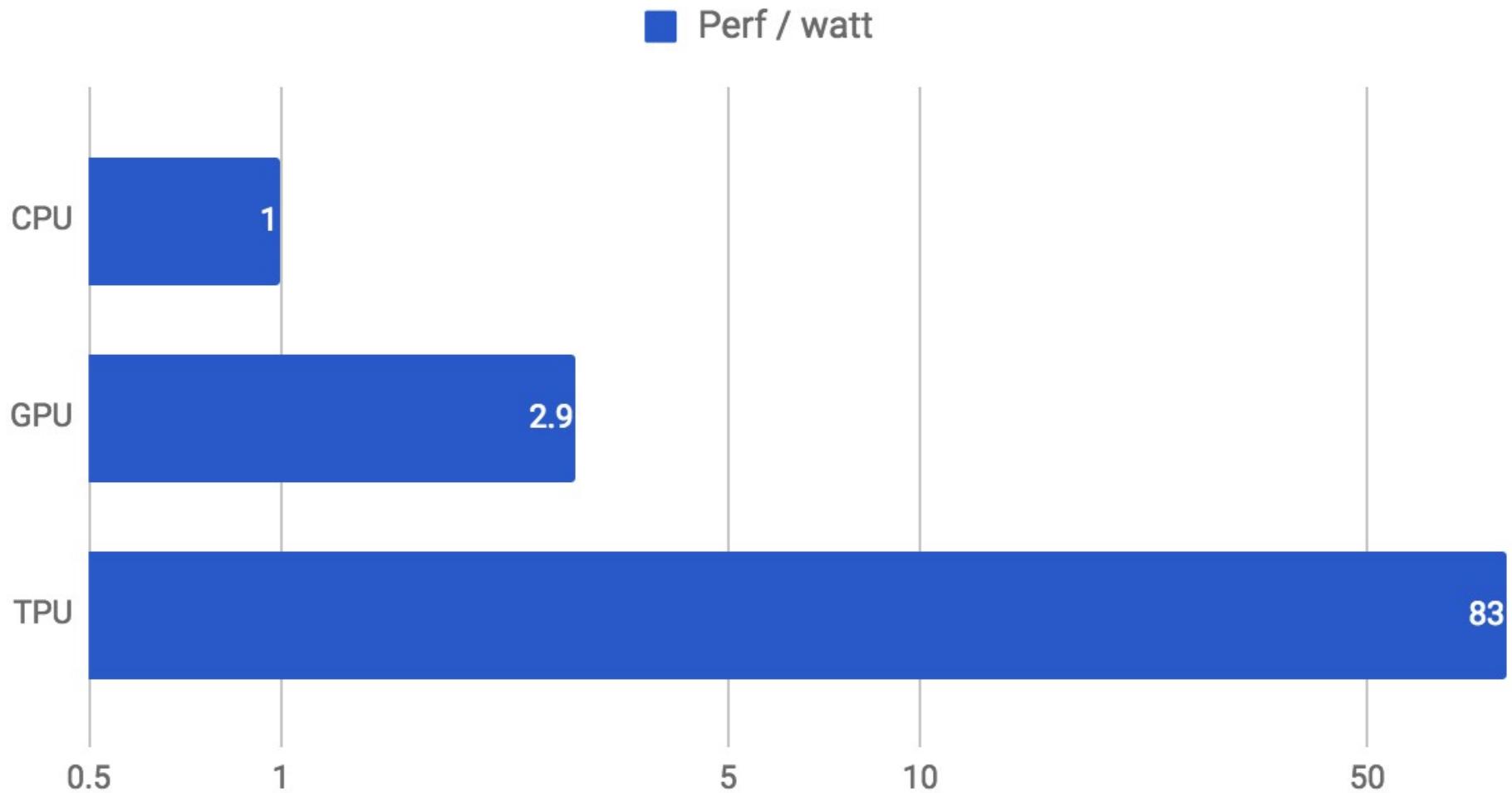
- El resultado es que mientras un procesador convencional procesa un escalar, una GPU procesa un vector, y una TPU procesa una matriz entera



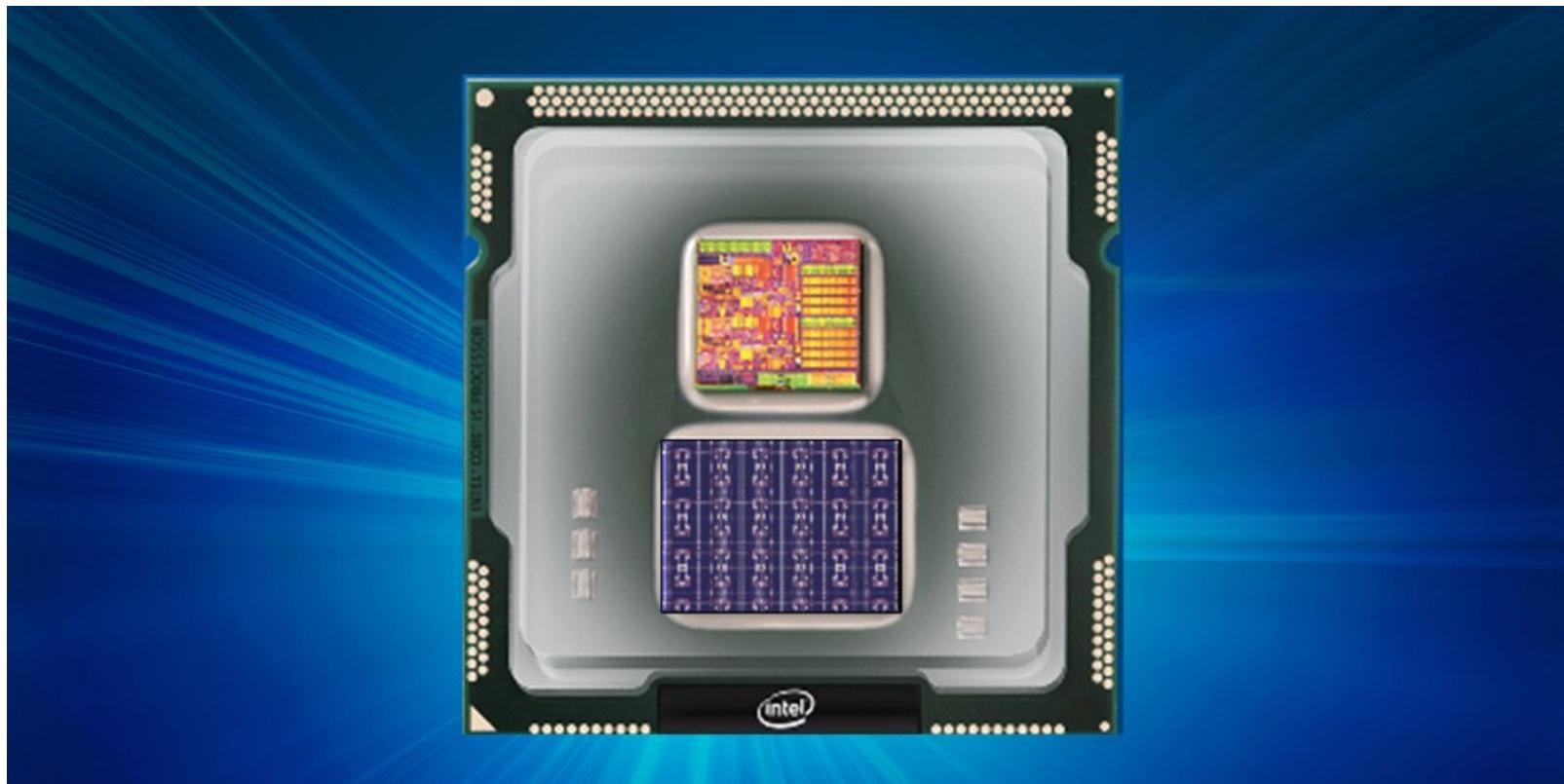
Comparación entre CPU, GPU y TPU



Comparación entre CPU, GPU y TPU

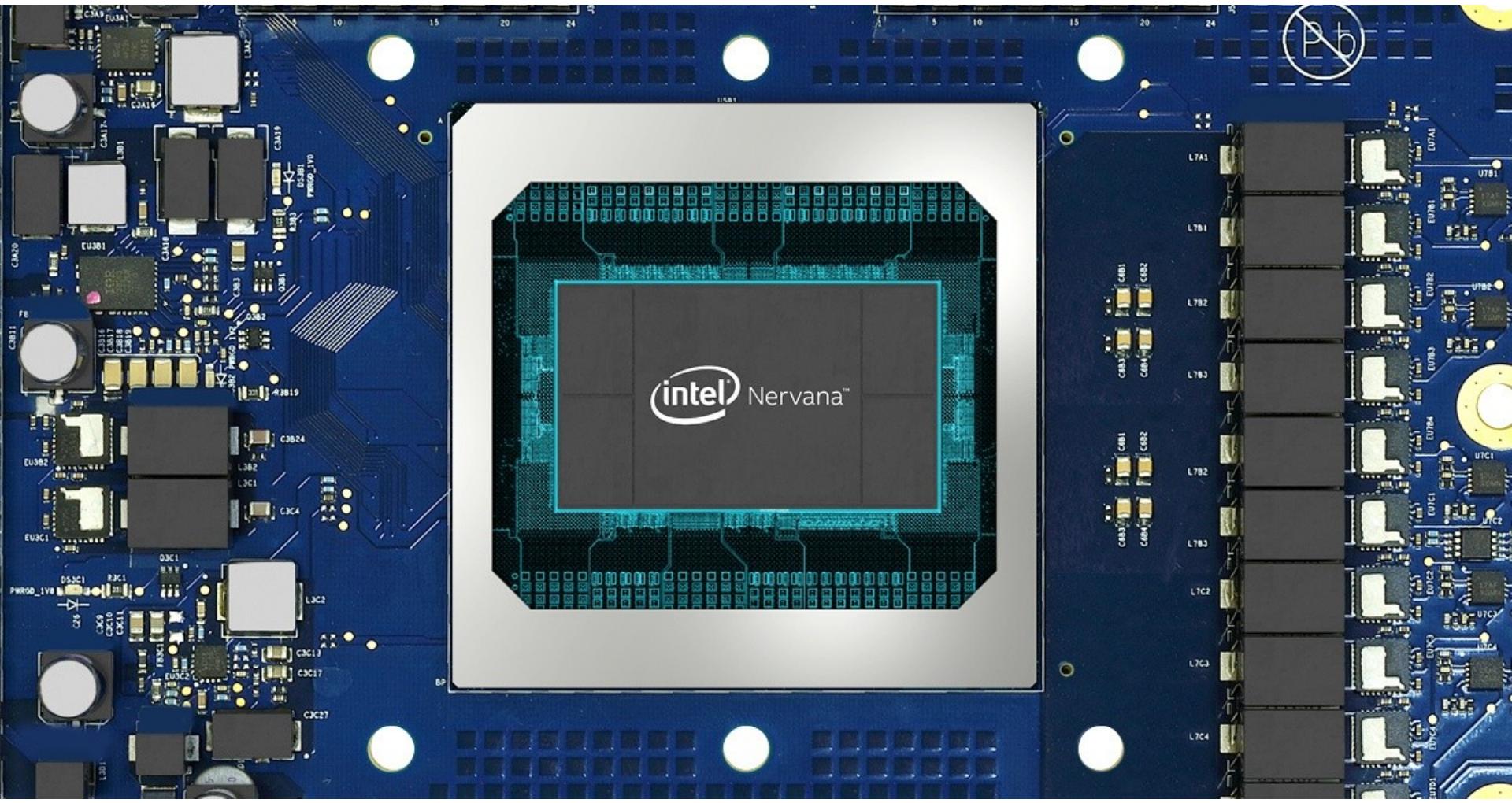


Chips de Intel para inteligencia artificial



**Intel's New Self-Learning Chip Promises to Accelerate Artificial Intelligence
Intel Introduces First-of-Its-Kind Self-Learning Chip Codenamed Loihi**
<https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/>

Chips de Intel para inteligencia artificial

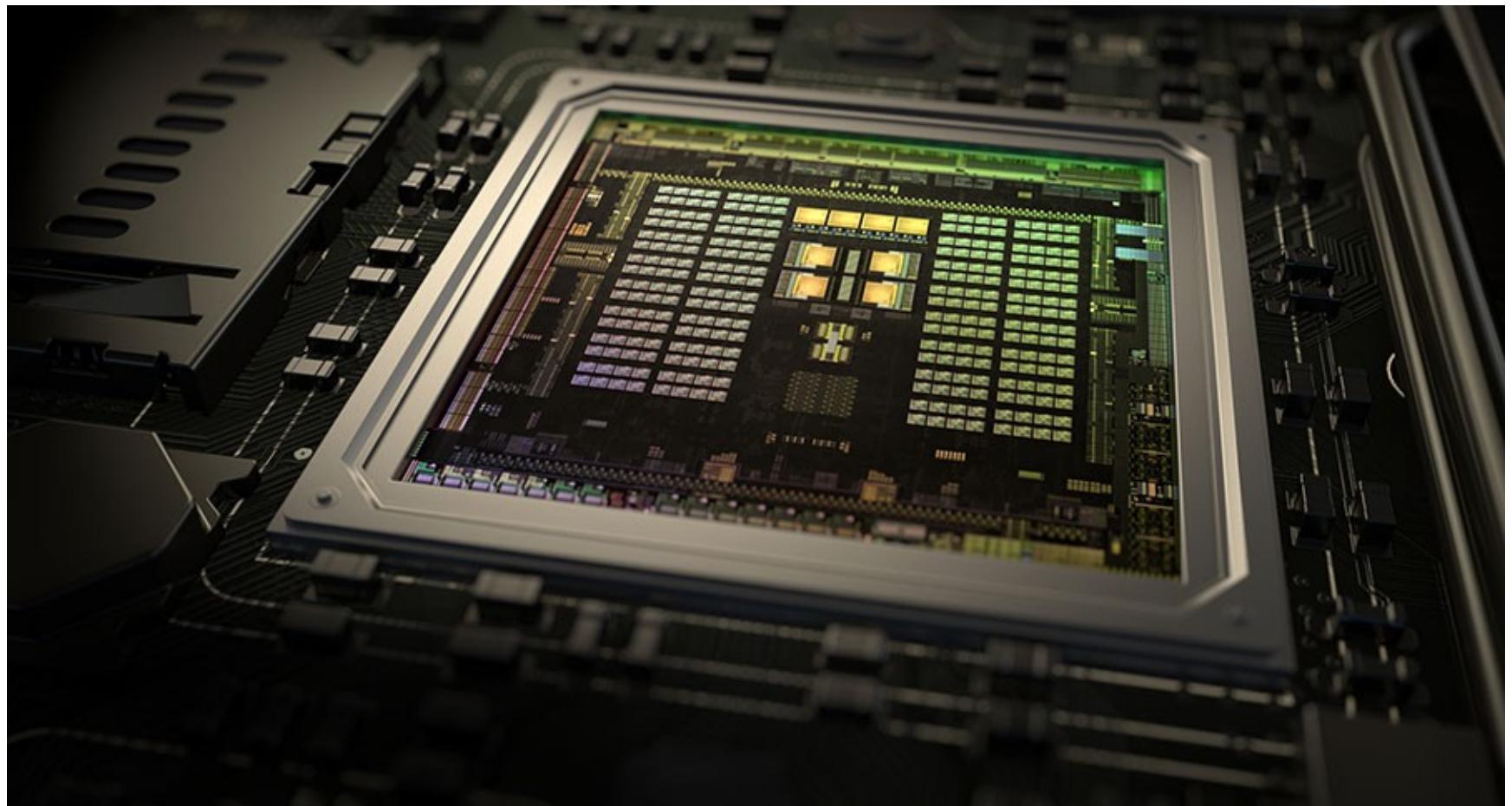


Atacando los problemas térmicos

- Las GPUs y FPGAs mejoran significativamente la eficiencia energética a nivel de clúster, pero los problemas de disipación de calor en el nivel de chip permanecen (hasta 300W por chip)
- ¿Qué podemos hacer a nivel de chip para evitar tener que apagar parte del mismo en el futuro?
 - Integrar los núcleos de la CPU y una variedad de aceleradores en el mismo chip, para que no estén todos funcionando a la vez
 - Reducir el tamaño de los chips
 - Integrar varios chips en cada encapsulado

Núcleos CPU y aceleradores en un chip

NVIDIA Tegra X1

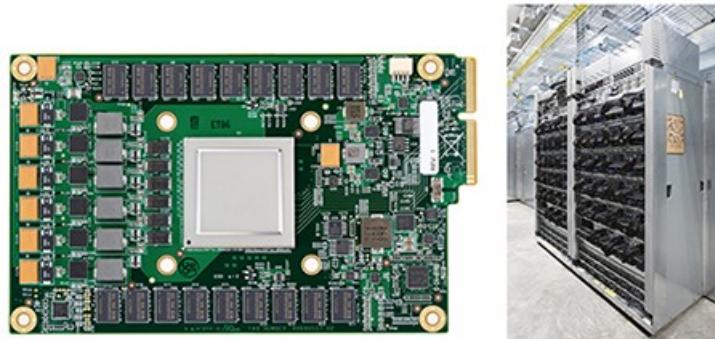


Núcleos CPU y aceleradores en un chip

- Tegra X1
 - GPU: NVIDIA Maxwell 256 núcleos
 - CPU: 4 núcleos de 64 bit ARM® A57, 2MB L2
- Tegra X2
 - GPU: NVIDIA Pascal 256 núcleos
 - CPU: 2 núcleos de 64 bit NVIDIA Denver2 ARMv8
- Adecuado para aplicaciones móviles (7,5 -15W)

Reducir el tamaño de los chips

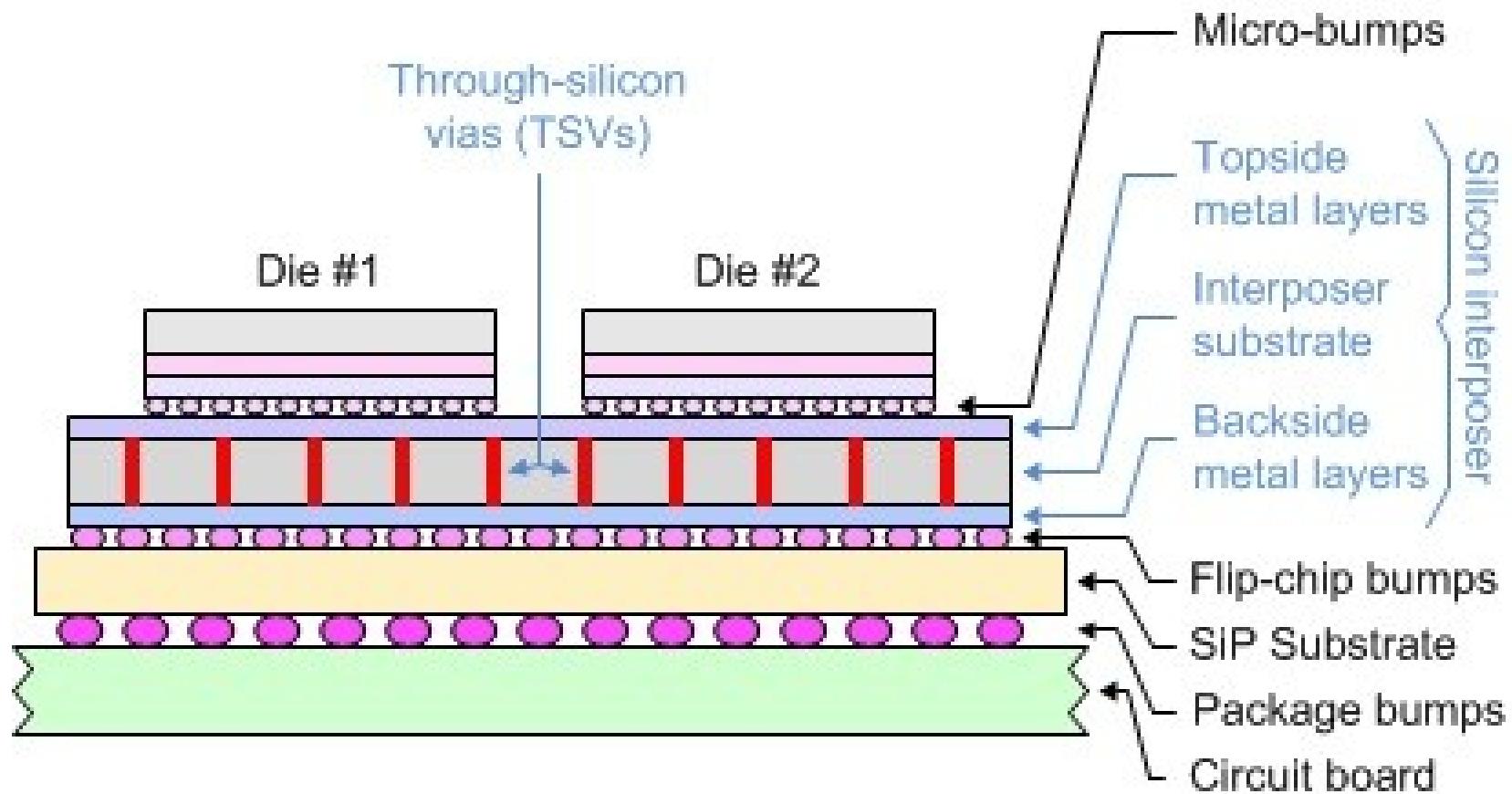
- TPU de Google
 - 64k unidades aritméticas enteras de 8 bits por chip
 - Mitad de área respecto a una CPU o GPU actual
 - Importante mejora en la productividad de las plantas de fabricación (bajo porcentaje de chips defectuosos)
- Adecuado sólo para redes neuronales



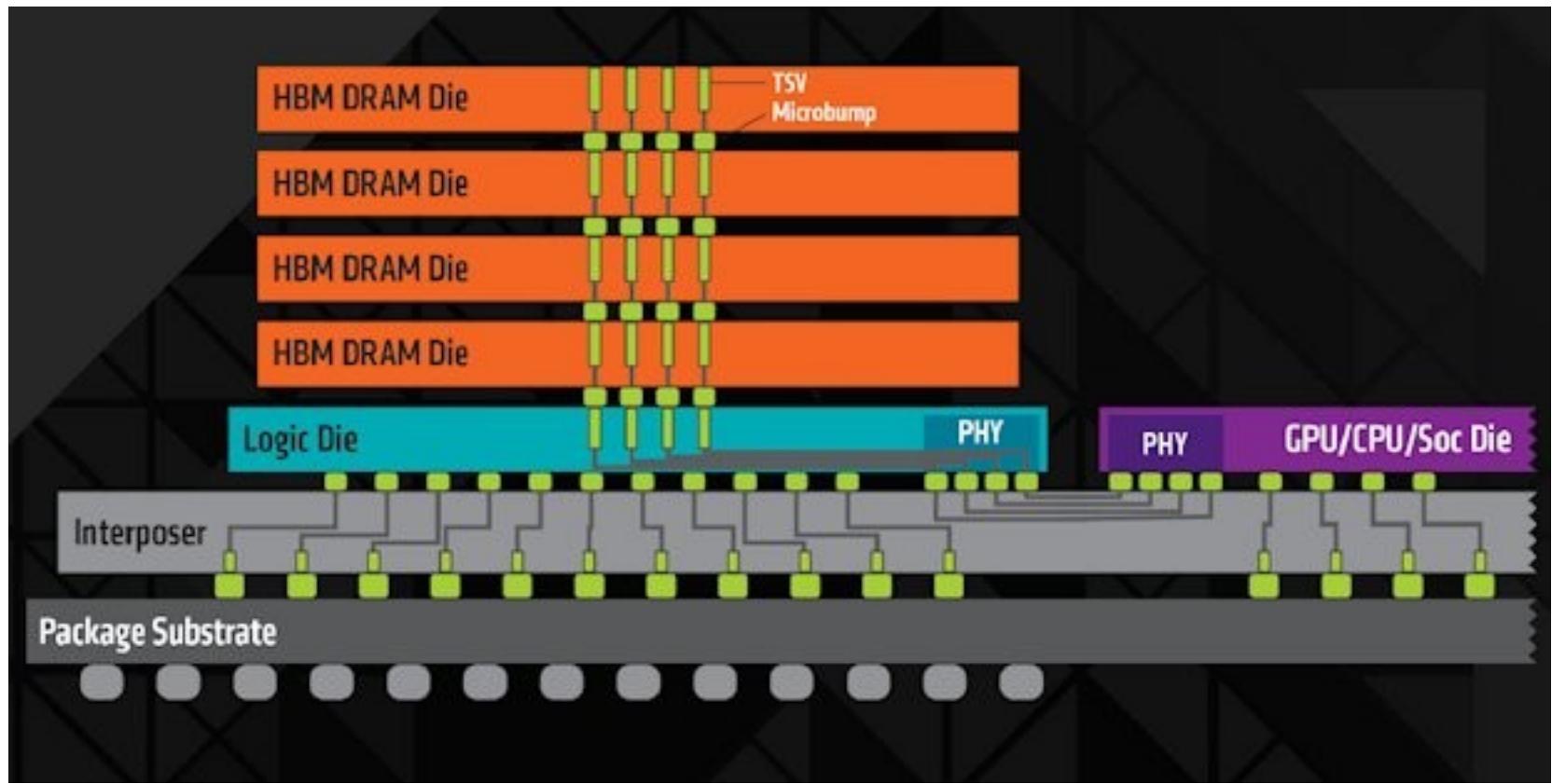
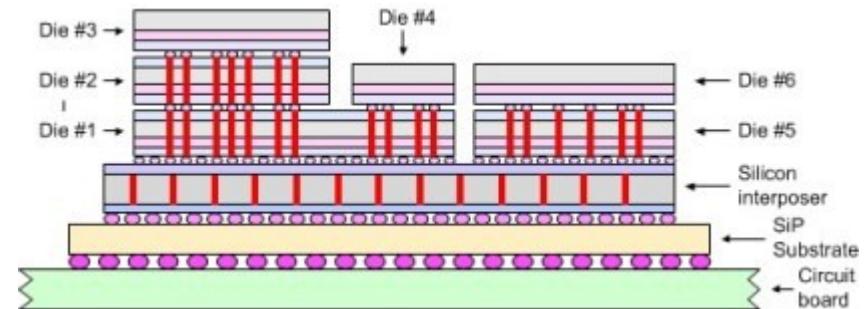
Varios chips en un encapsulado

- Fragmentar un chip en varios más pequeños:
 - Descomponer un chip grande en chips más pequeños con idénticas o diferentes funciones.
 - Pila 3D de memoria HBM DRAM, núcleos de CPU, núcleos de GPU, otros aceleradores
 - Importante mejora en la productividad de las plantas de fabricación (bajo porcentaje de chips defectuosos)
 - Integración de los circuitos de cálculo con la memoria
 - Ancho de banda de memoria muy elevado (HBM DRAM)

2.5D interposer



3D interposer

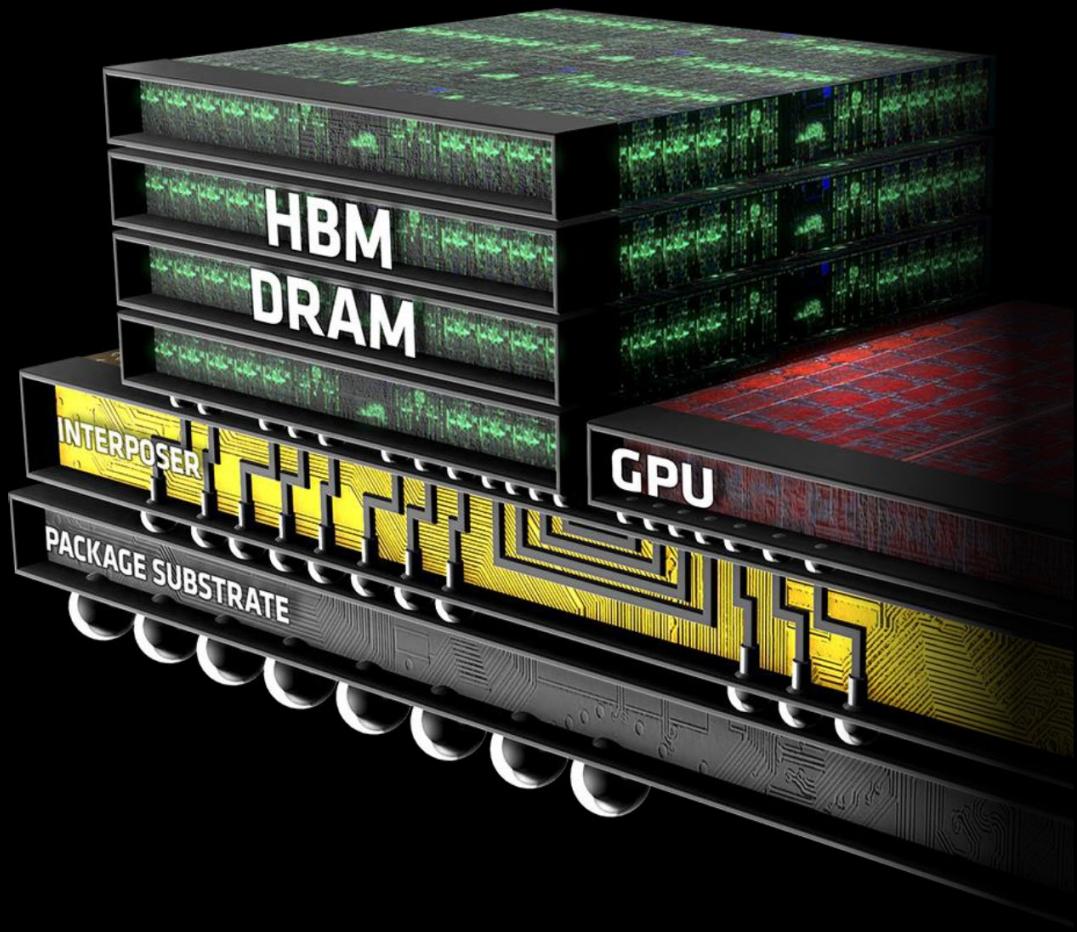


GRAPHICS TECHNOLOGY LEADERSHIP



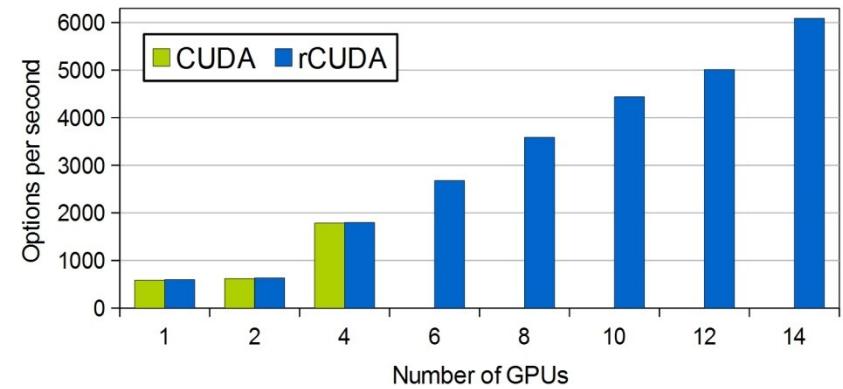
► HIGH BANDWIDTH MEMORY

- ▲ First in the Industry with High Bandwidth Memory (HBM) Technology
- ▲ 3D HBM DRAM Die Stack on Silicon Interposer
- ▲ >3X Performance/Watt Compared to GDDR5³
- ▲ >50% Power Savings Versus GDDR5⁴



- La tecnología de interposición de silicio es muy prometedora para la integración a gran escala de memoria y múltiples aceleradores específicos en el encapsulado del procesador
- Se pueden integrar múltiples pilas de memoria en el encapsulado de un procesador, lo que aumenta tanto la capacidad como el ancho de banda proporcionado por la memoria 3D
- Interconexiones especializadas para conectar núcleos de CPU, aceleradores y memoria están actualmente en desarrollo

Maximizando los beneficios del uso de aceleradores



El ancho de banda de los enlaces de comunicaciones es similar al de la memoria

- ConnectX-6 con Virtual Protocol Interconnect® soporta dos puertos InfiniBand o Ethernet de 200Gb/s, con latencia inferior a 600 nanosegundos, y 200 millones de mensajes por segundo
- Intel OmniPath (OP) HFI soporta 100 Gb/s por puerto, llegando hasta 160 millones de mensajes por segundo. Intel fabrica procesadores que integran OP

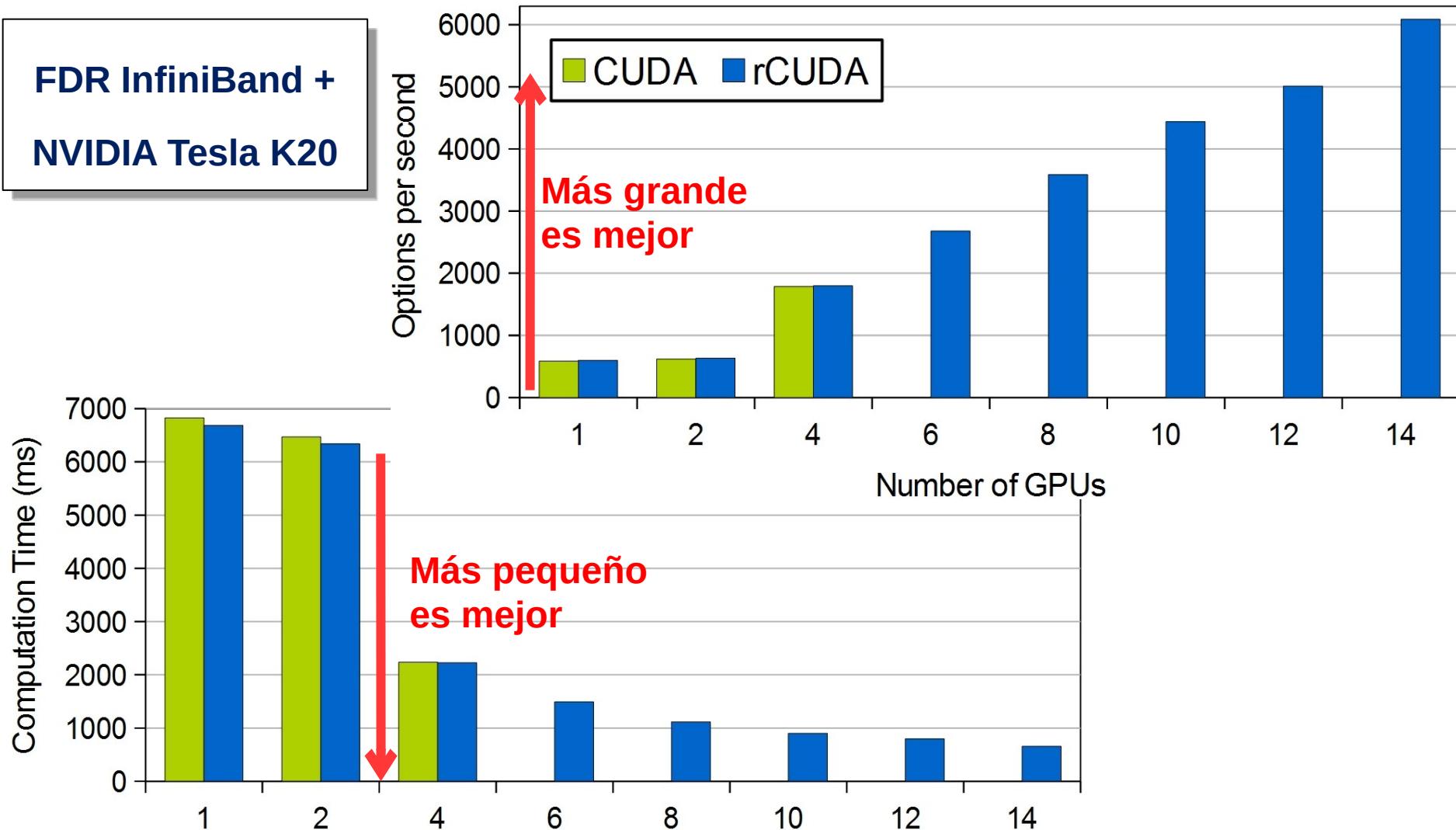
Maximizando los beneficios

- Los aceleradores son muy especializados y solo se pueden usar en ciertos fragmentos de código
- La utilización de un acelerador suele ser inferior a lo permitido por los límites de calentamiento
- Nuestra solución: Un ***grupo compartido de aceleradores virtualizados***
- Tenemos los ingredientes: aceleradores en un clúster e interconexión de alta velocidad
- Nuestra aportación:  . Proporciona acceso remoto a GPUs y virtualiza su uso

Maximizando los beneficios

- MonteCarlo Multi-GPU

FDR InfiniBand +
NVIDIA Tesla K20



- Estamos ante una revolución sin precedentes en las tecnologías de la información
- Las limitaciones que se avecinan en escala de integración (y las actuales en calentamiento) se van a suplir con un uso masivo de aceleradores
 - Ya existen dispositivos móviles con aceleradores integrados y aparecerán dispositivos más potentes
 - Habrá una oferta masiva de aceleradores diversos en la nube (incluyendo los cuánticos, en un futuro)
- Las aplicaciones de inteligencia artificial van a ser el principal motor del cambio