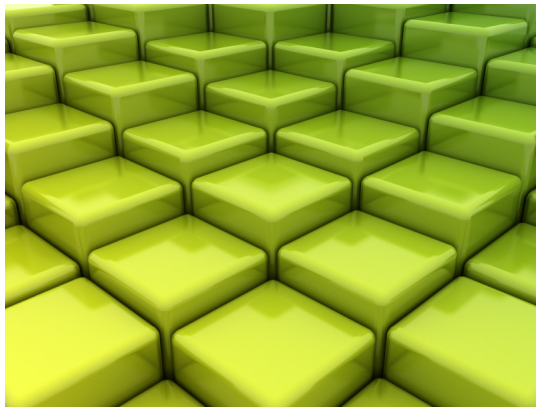
GPUs para HPC: Logros y perspectivas futuras

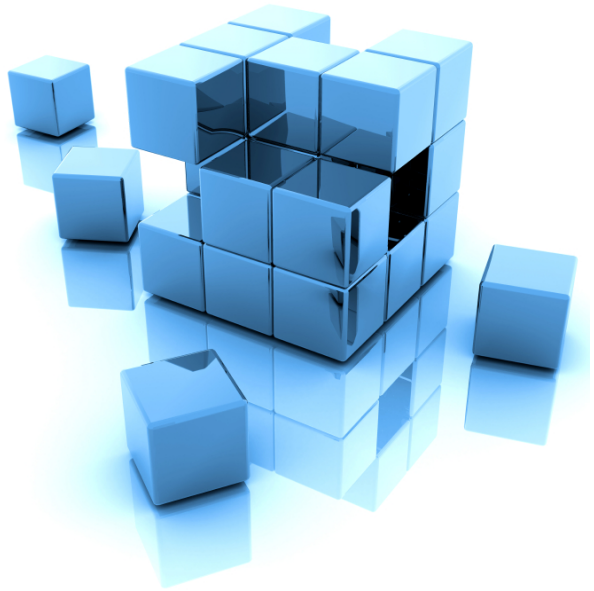


Manuel Ujaldón

Nvidia CUDA Fellow

Dpto. de Arquitectura de Computadores

Universidad de Málaga



Contenidos de la charla [40 diapositivas]

1. Evolución y logros [2]
2. El nuevo hardware [4]
3. Un desarrollo prometedor de Stacked DRAM: HMC [6]
4. Tridimensionalizando el sistema de memoria principal [15]
5. Mejoras del HMC 1.0 respecto a la DRAM actual [7]
6. Impacto sobre las GPUs [3]

I. Evolución

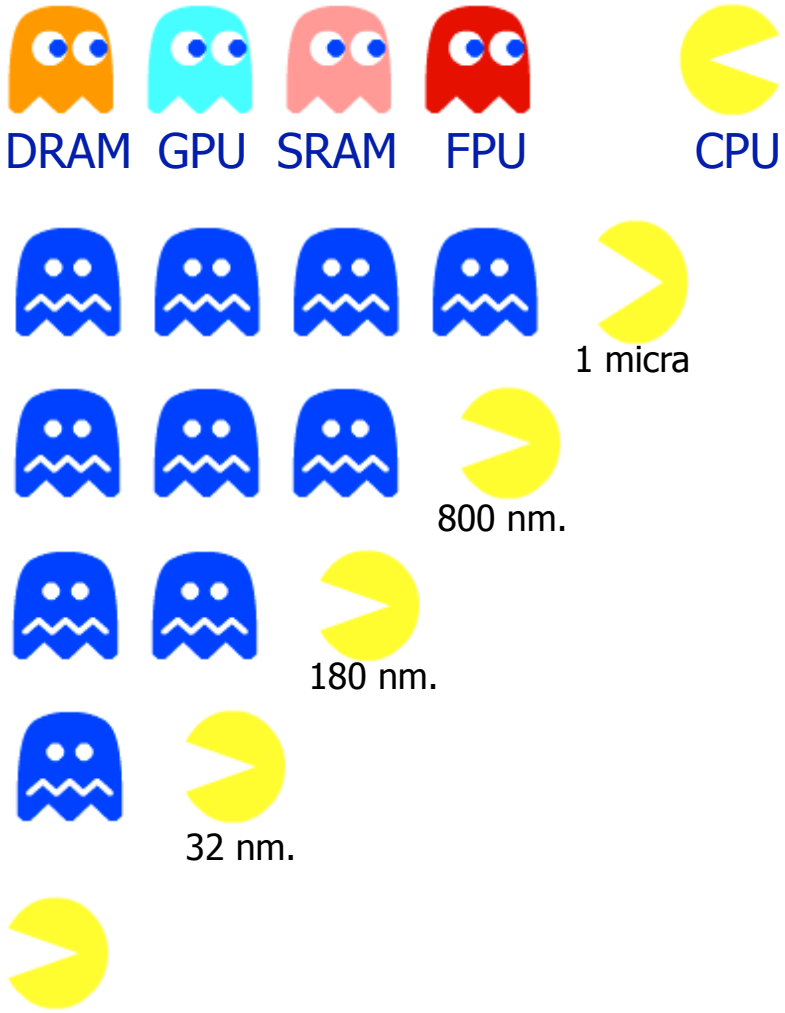
sarteco

Jornadas Sarteco

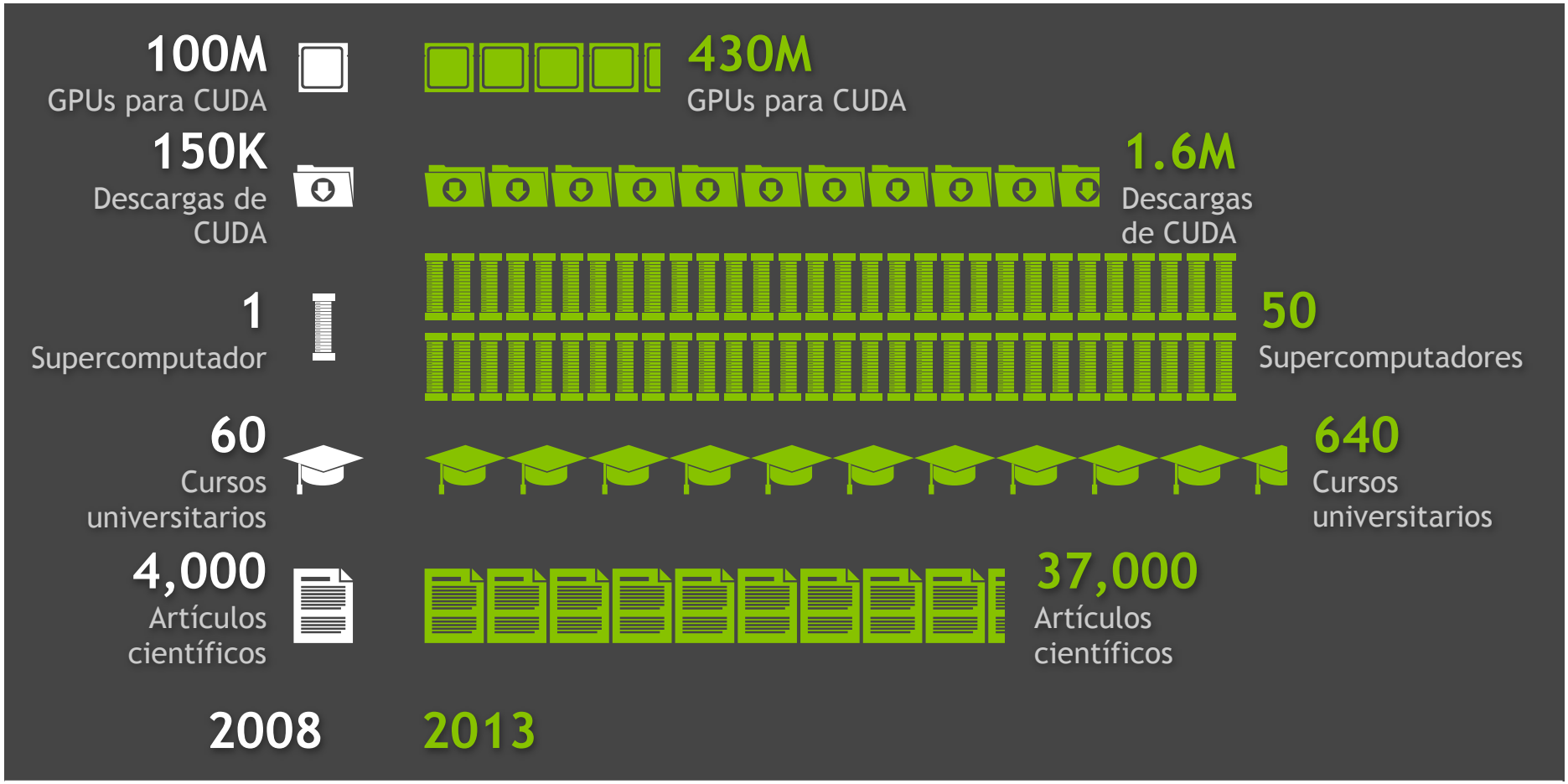
17-20 Septiembre, Madrid

Evolución

- En los inicios del PC, los componentes del procesador estaban descentralizados.
- Conforme se redujo la distancia de integración a menos de 1 micra, éstos se fueron integrando *on-die*:
  - 1989: FPU [Intel 80486DX].
  - 1999: SRAM [Intel Pentium III].
  - 2009: GPU [AMD Fusion].
  - 2016: DRAM [Nvidia Volta].
- El desenlace final es el concepto de SoC (System-on-Chip).



Logros: CUDA como motor propulsor de GPU



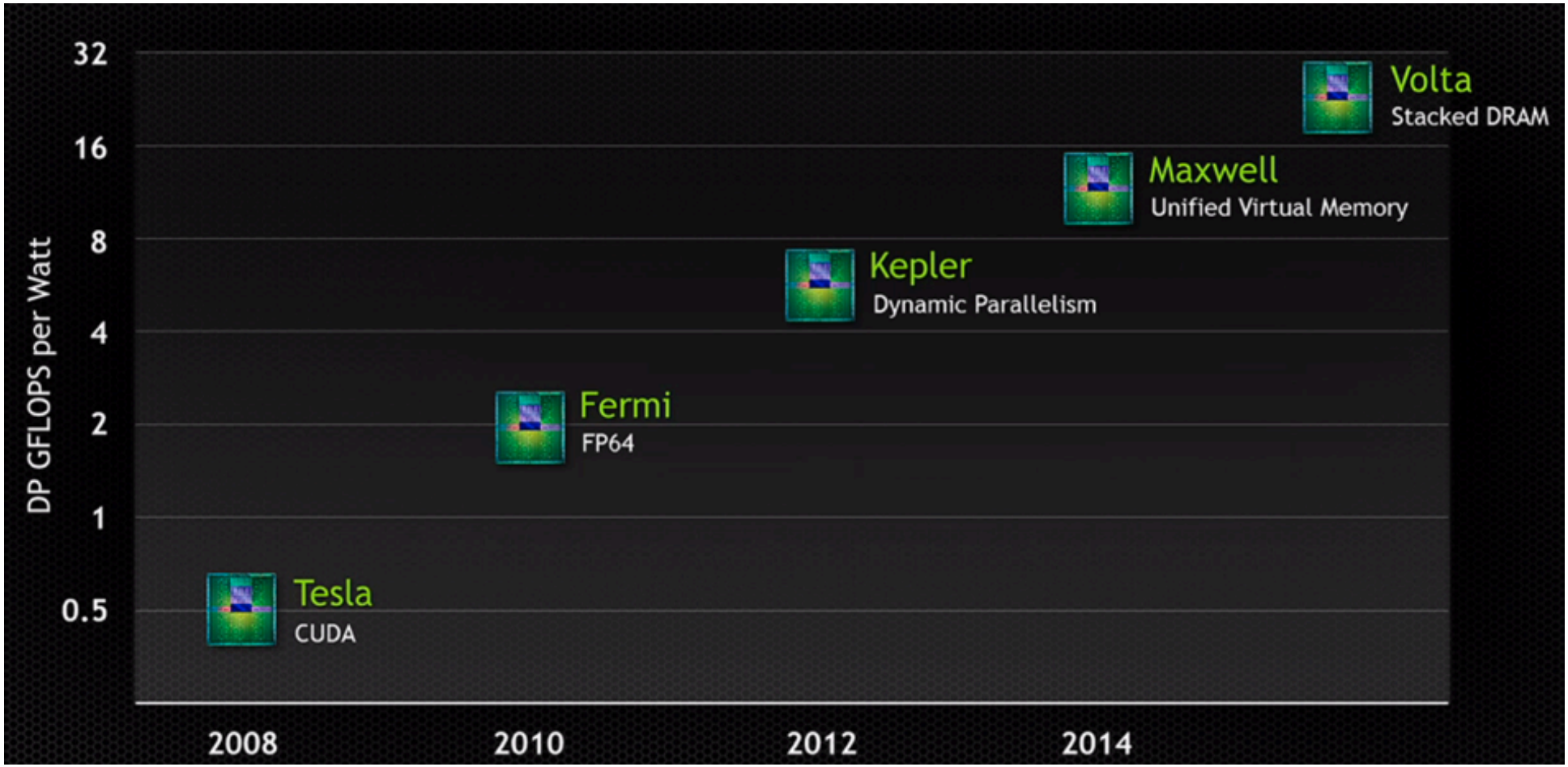
- Se produce una descarga del SW. de CUDA cada minuto.

II. El nuevo hardware

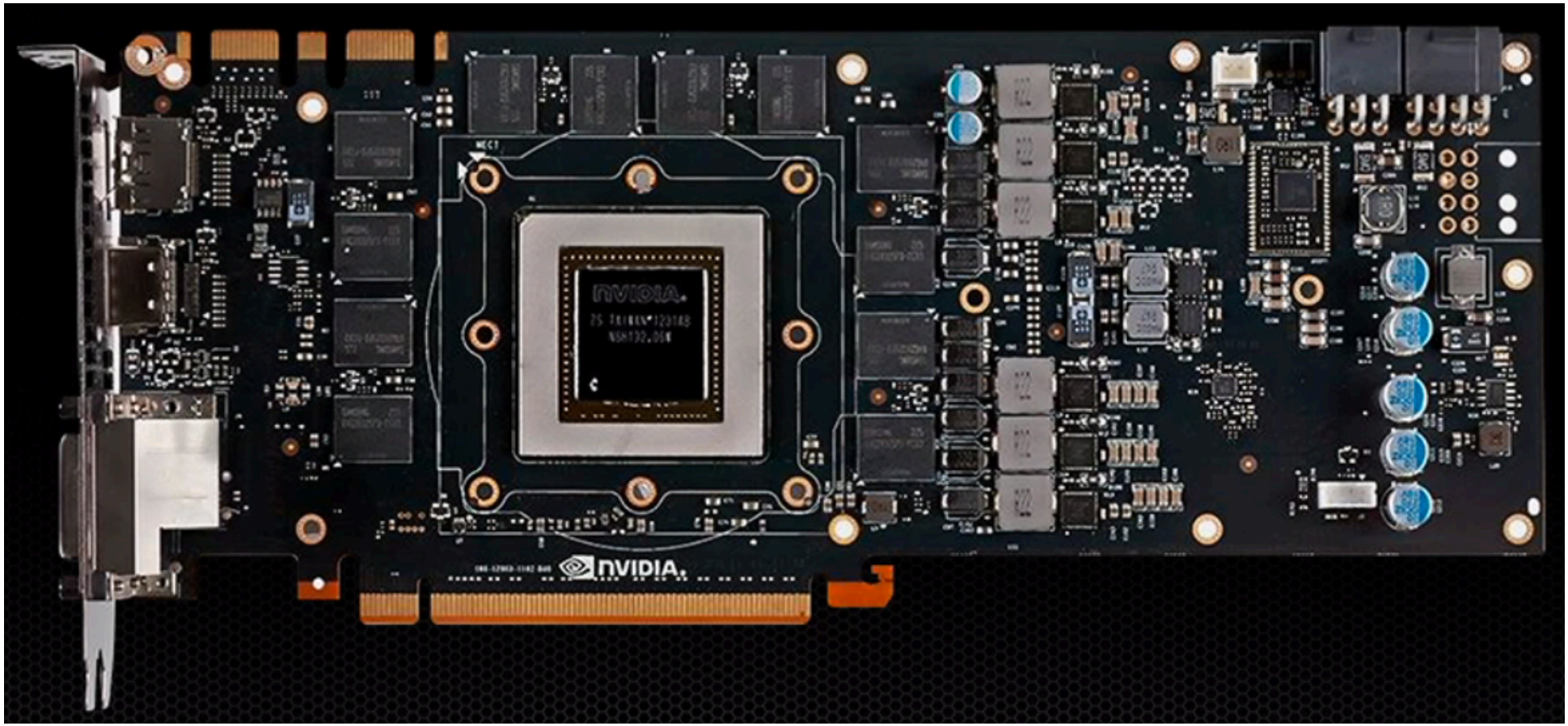
Jornadas Sarteco

17-20 Septiembre, Madrid

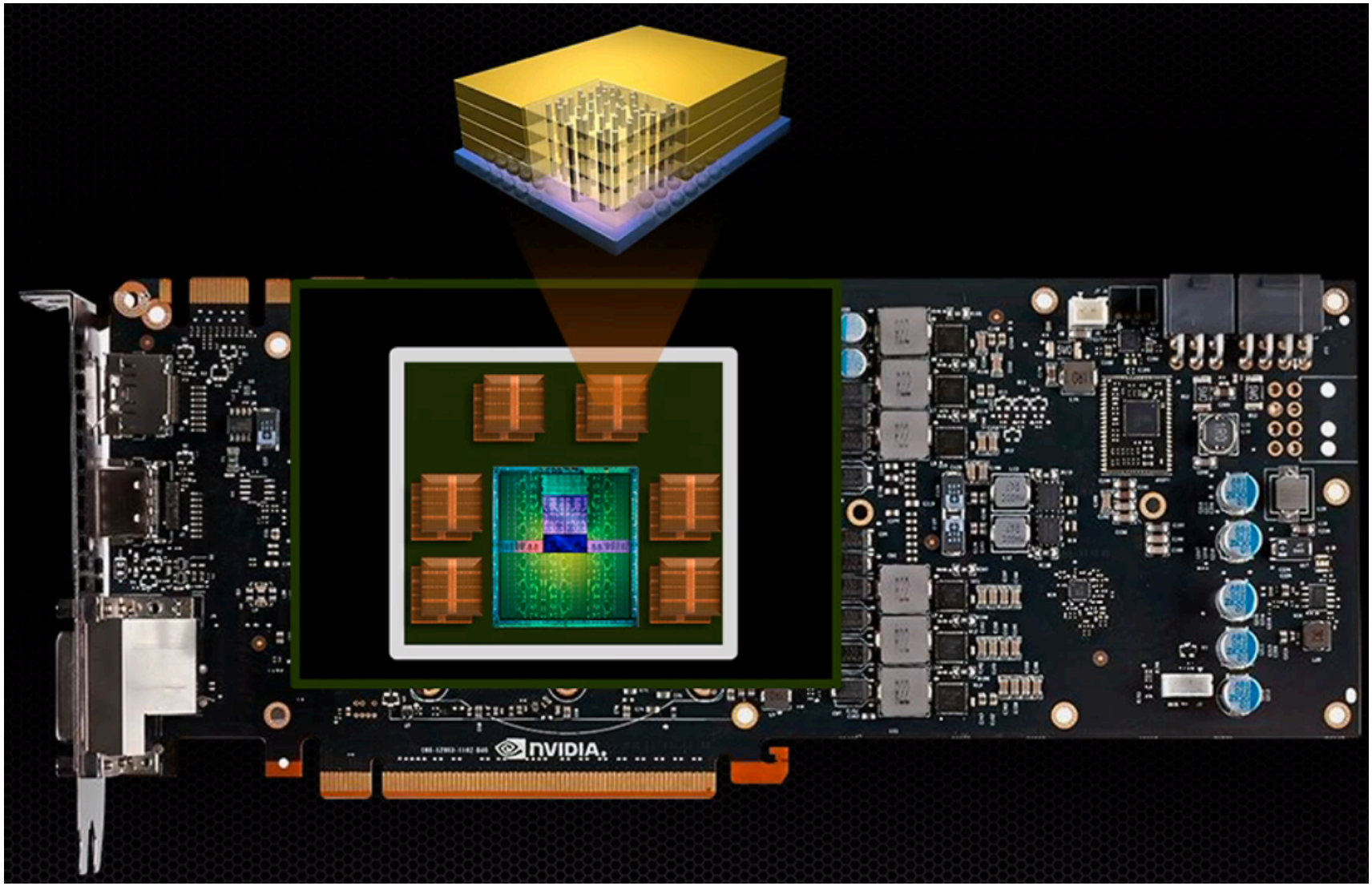
Hoja de ruta de las GPUs de Nvidia [GTC'13]



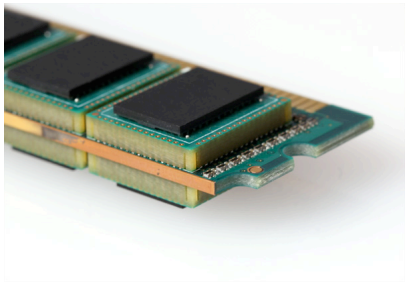
La tarjeta gráfica de 2013: GPU Kepler con memoria de vídeo GDDR5

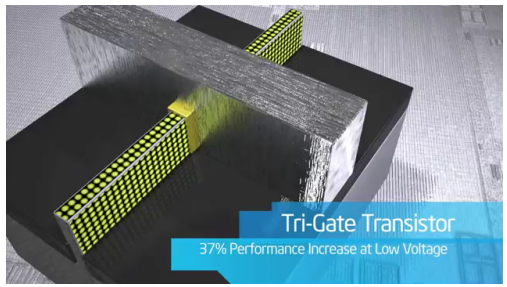


La tarjeta gráfica de 2016/17: GPU Volta con memoria Stacked (3D) DRAM



Aclaraciones previas





- Los desarrollos y análisis que veremos aquí NO guardan relación con los transistores 3D tri-gate de Intel de 22 nm. De hecho, ambas son mejoras compatibles y acumulables.
- La estructura 3D de los chips es también compatible con la integración 2D en la mayoría de sus variantes, por lo que no vamos a prescindir de nada de lo ya logrado.
- Aunque nos centramos en el procesador, y sobre todo la memoria, la fabricación 3D puede orientarse a cualquier tecnología: CPU-GPU, SRAM-DRAM, ASIC, DSP, empotrados...
- La evacuación térmica sigue produciéndose en las áreas periféricas, por lo que las capas centrales suelen dedicarse a la memoria DRAM.

III. Un desarrollo prometedor de 3D DRAM: HMC

sarteco

Jornadas Sarteco

17-20 Septiembre, Madrid

Hybrid Memory Cube Consortium (HMCC)

Logros y objetivos del HMCC	Fecha
Primeros artículos publicados sobre Stacked DRAM (basados en proyectos de investigación)	2003-2006
Primer desarrollo comercial de la tecnología, a cargo de Tezzaron Semiconductors	Enero de 2005
Micron Technologies y Samsung Electronics lanzan el HMCC	Octubre de 2011
Nvidia adopta Stacked DRAM para su GPU Volta	Marzo de 2013
Disponible la especificación HMC 1.0	Abril de 2013
Primeras muestras de producción basadas en el estándar	Mediados de 2014 (estimado)
Configuración 2.5 disponible	Finales de 2014 (estimado)

Miembros desarrolladores del HMCC (a fecha Julio de 2013)

Altera Corporation

ARM

IBM

Micron Technology, Inc

Open-Silicon, Inc.

Samsung Electronics Co., Ltd

SK hynix

Xilinx, Inc.

Fundadores del consorcio

Adopción por un amplio número de compañías del sector

- La iniciativa HMC se orientó inicialmente a HPC y redes, pero también tiene su utilidad para tecnología móvil y como alternativa a DDR.
- HMC se encuentra ligada a las CPUs, GPUs y ASICs en configuraciones punto a punto, donde el rendimiento de HMC ofrece el ancho de banda de la memoria óptica.

Adopter Members:

- Accel, Ltd.
- Achronix Semiconductor Corporation
- ADATA Technology Co., Ltd.
- AIRBUS
- Altior
- Analog Bits
- APIC Corporation
- Arira Design
- Arnold & Richter Cine Technik
- Atria Logic, Inc.
- BroadPak Corporation
- Cadence Design Systems, Inc.
- Cascade Microtech
- Convey Computer Corporation
- Cray Inc.
- DAVE Srl
- Design Magnitude Inc.
- Dream Chip Technologies GmbH
- eSilicon Corporation
- Exablade Corporation
- EZchip Semiconductor Ltd.
- FirstPass Engineering
- FormFactor, Inc.
- Fujitsu Advanced Technologies Ltd.
- Galaxy Computer System Co., Ltd.
- GDA Technologies
- GLOBALFOUNDRIES
- GraphStream Incorporated
- Green Wave Systems Inc.
- HGST, a Western Digital Company
- HiSilicon Technologies Co., Ltd.
- HOY Technologies
- Huawei Technologies
- Industrial Technology Research Institute (ITRI)
- Infinera Corporation
- Inphi Corporation
- Integrated Device Technology
- Ircona
- ISI / Nallatech
- Juniper Networks
- KALRAY
- Kool Chip Inc.
- Korea Advanced Institute of Science and Technology
- Lawrence Livermore National Laboratory
- LeCroy Corporation
- Liquid Logic, LLC
- LogicLink Design, Inc.
- Lomonosov Moscow State University
- Luxtera, Inc.
- Marvell
- Mattozetta Technologies
- Maxeler Technologies Ltd.
- MediaTek
- Memoir Systems Inc.
- Mentor Graphics
- Miranda Technologies Partnership
- Mobiveil, Inc.
- Montage Technology, Inc.
- Napatech A/S
- National Instruments
- NEC Corporation
- Netronome
- New Global Technology
- Northwest Logic
- Obsidian Research
- OmniPhy
- Oregon Synthesis
- Perfcraft
- Pico Computing
- Renesas Electronics Corporation
- Science & Technology Innovations
- SEAKR Engineering
- SIMMTECH Co., Ltd.
- Somerset Technology Services, Inc.
- STMicroelectronics
- Suitcase TV Ltd.
- T-Platforms
- Tabula
- Tech-Trek Ltd.
- Technion - Israel Institute of Technology
- Teledyne LeCroy
- Teradyne, Inc
- The Regents of the University of California
- Tilera Corporation
- Tongji University
- TU Kaiserslautern, Lehrstuhl Entwurf Mikroelektronischer Systeme
- UC Irvine
- United Microelectronics Corporation
- University of Heidelberg ZITI (Center for Computer Engineering
- University of Rochester
- University of Southern California
- Winbond Electronics Corporation
- Woodward McCoach, Inc.
- ZTE Corporation

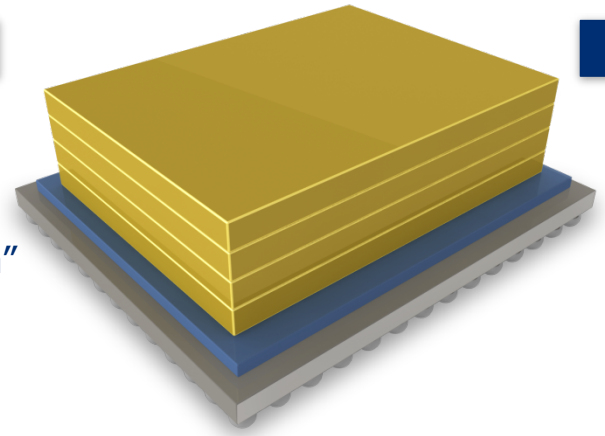
El Hybrid Memory Cube a grandes rasgos

Aproximación revolucionaria para flanquear el "Memory Wall"

- La hoja de ruta de la DRAM tiene limitaciones en ancho de banda y eficiencia energética.
- Micron introduce una nueva clase de memoria: Hybrid Memory Cube.
- Combinación peculiar de DRAMs sobre lógica adicional.

Aspectos clave

- Controlador de lógica diseñado por Micron.
- Enlace de alta velocidad con la CPU.
- Conexión a DRAM "Through Silicon Via" masivamente paralela.



Prototipos finalizados en silicio
HOY

Rendimiento sin precedentes

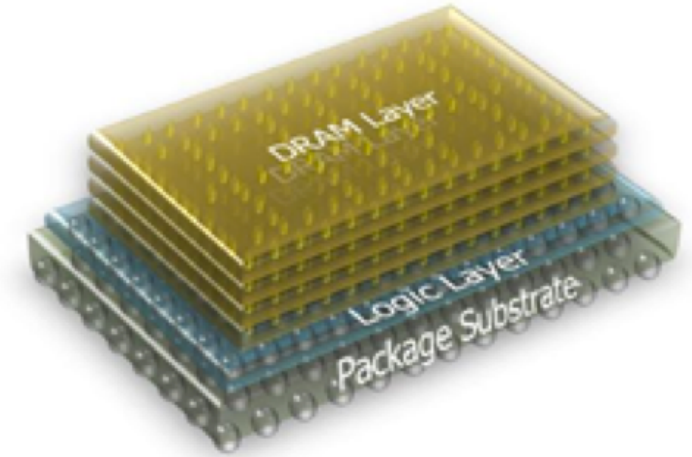
- Hasta 15 veces superior en ancho de banda a un módulo DDR3 [pero sólo 2x vs. GDDR5 unidireccionalmente].
- Consumo energético por bit inferior en un 70% a las tecnologías actuales [medido en número de señales activas, el ahorro energético sale del 50%].
- Ocupa un 90% menos de espacio que los módulos RDIMM actuales [se ahorra hasta el 95%, y midiendo el área de los módulos, superior al de los zócalos].

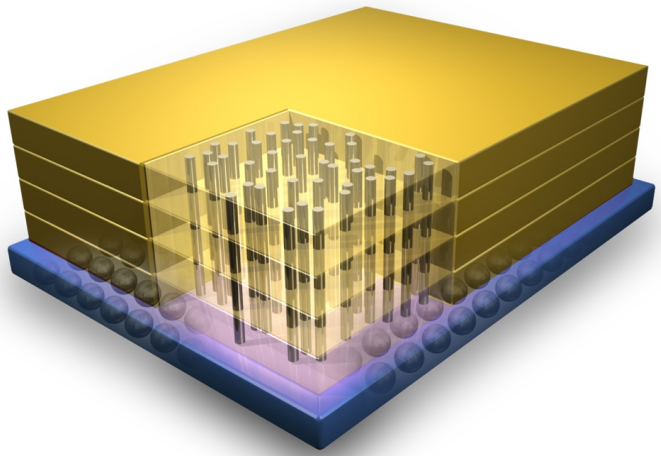
Eminentemente orientados a HPC y redes, aunque eventualmente pueden dirigirse productos a la computación doméstica

[según mi propio estudio, que presentaré más adelante]

Detalles de la integración en silicio

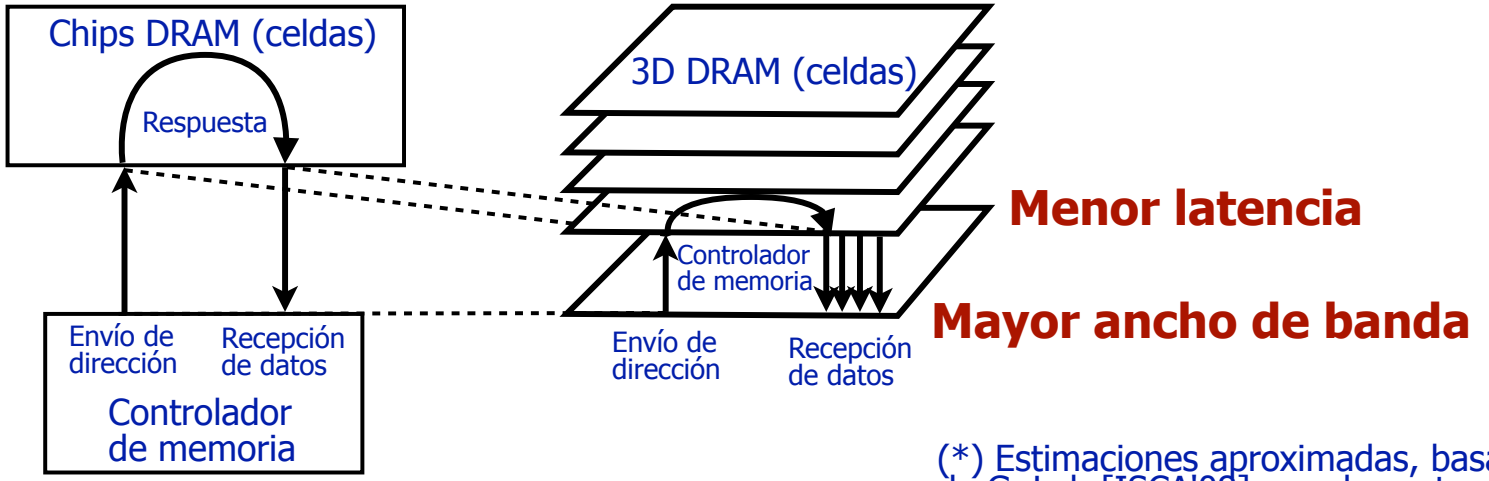
- Las celdas de DRAM se organizan en **torres** gemelas (vaults), que suscriben el entrelazado matricial de los chips de memoria DRAM coetánea.
- El controlador DRAM se sitúa en la **capa** inferior, y las matrices de celdas en las capas superiores.
- Controlador y datos se conectan con **vías** TSV (through-silicon vias) verticales, con surcos esculpidos en silicio de entre 4 y 50 um. según el fabricante.
  - Para surcos de 10 um., un bus de 1024 bits (16 canales de memoria) requiere un **área de integración** de 0.32 mm2, lo que representa un 0.2% del dado de una CPU (160 mm2).
  - La **latencia** vertical para recorrer la altura de una Stacked DRAM de 20 capas es de sólo 12 picosegs.
- El último paso es el ensamblaje conjunto de todos los elementos: Torres, capas y vías. Esto previene las capacitancias parásitas que reducen la velocidad de la señal y aumentan el consumo para conmutar.





¿Cómo revierte todo esto en un chip DRAM?

- Duplicando su velocidad (*), con 3 responsables básicos:
  - Las **conexiones más cortas** entre el controlador de memoria y las matrices de celdas DRAM mejora un tercio la velocidad.
  - La **expansión del bus hasta los 512 bits** gracias a la mayor densidad de las conexiones revierte en otro tercio de mejora.
  - La **menor latencia** gracias a las conexiones TSV más rápidas y los **mayores entrelazados** en una geometría 3D mejoran el otro tercio.



(*) Estimaciones aproximadas, basadas en simulaciones de G. Loh [ISCA'08] que demostraron mejoras de 2.17x.

IV. Tridimensionalizando el sistema de memoria principal

Jornadas Sarteco

17-20 Septiembre, Madrid

Cómo funciona el sistema de memoria DRAM

Zócalos: 3,2,1,0 4,5,6,7

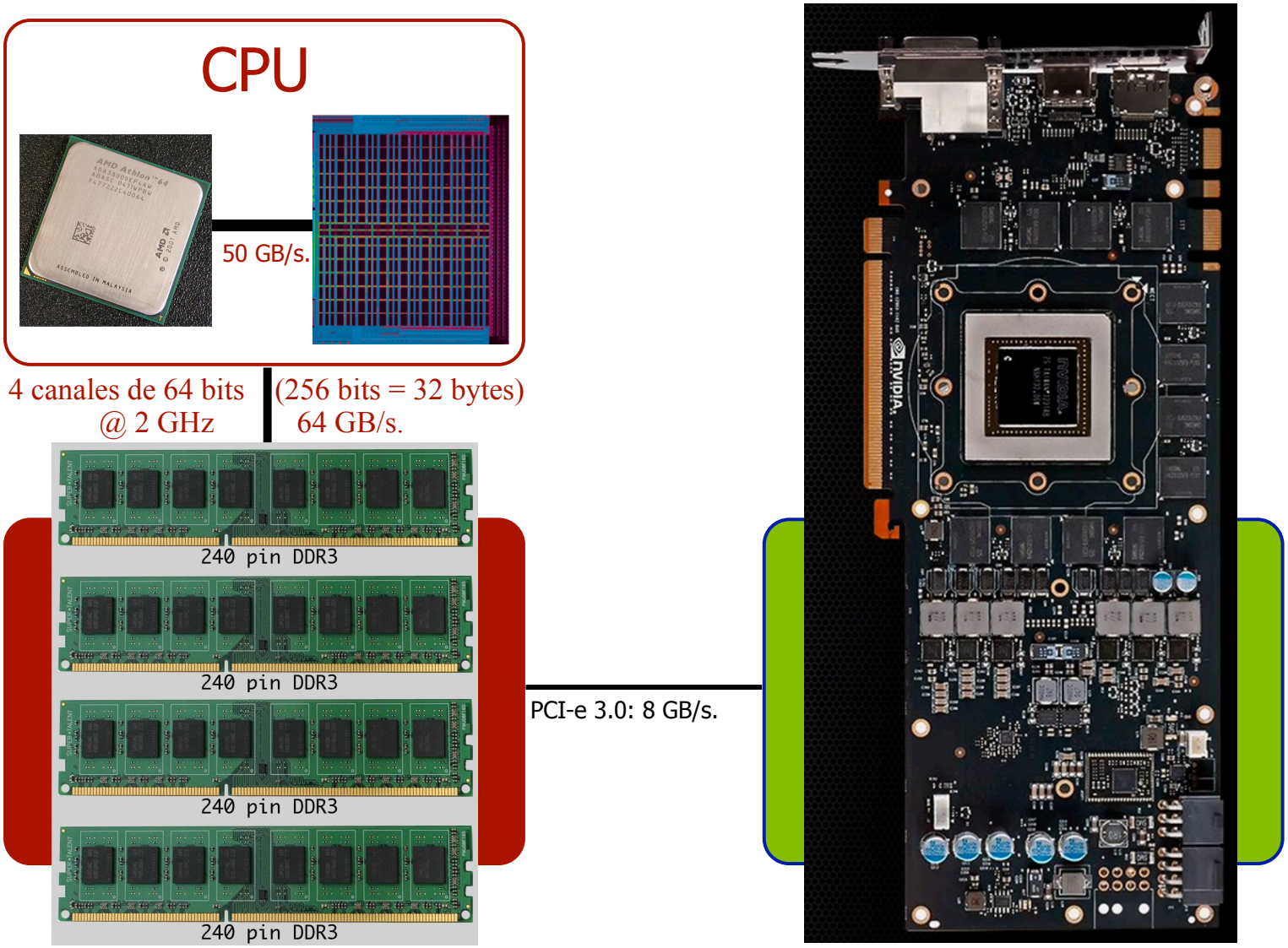
- Un ejemplo: La placa base Asus Rampage IV Extreme.
  - 8 zócalos de memoria DDR3.
  - Configuración quad-channel.
  - Dos bancos independientes.
  - Los módulos deben llenar todos los zócalos del mismo color.

Banco 0 (obligatorio): Zócalo 0 + zócalo 2 + zócalo 4 + zócalo 6 = 256 bits de anchura.

Banco 1 (opcional): Zócalo 1 + zócalo 3 + zócalo 5 + zócalo 7 = 256 bits de anchura.
Utilizar el banco 1 permite:
- Adquirir un sistema de memoria que no sea potencia de 2 (ej: 8 GB + 4 GB = 12 GB).
- Hacer de cortafuegos en caso de avería.
- Ampliar memoria el día de mañana si lo dejamos inicialmente vacío.

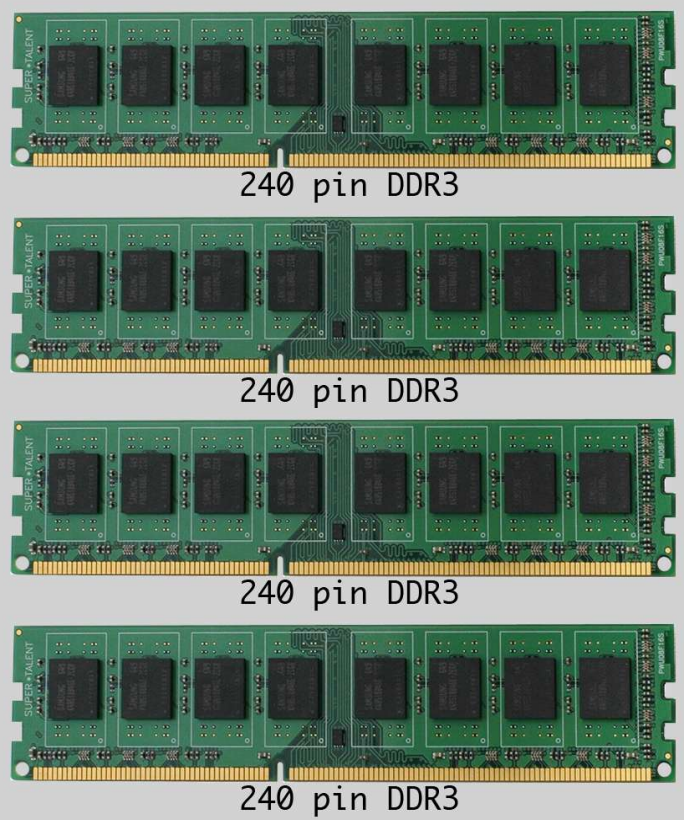
Pero con la memoria 3D diremos adiós a toda esta modularidad en favor de la velocidad, como ya sucede con el sistema de memoria de la GPU. Nos centramos pues en el banco 0.

Típico sistema DRAM conformado con 4 módulos de memoria DDR3

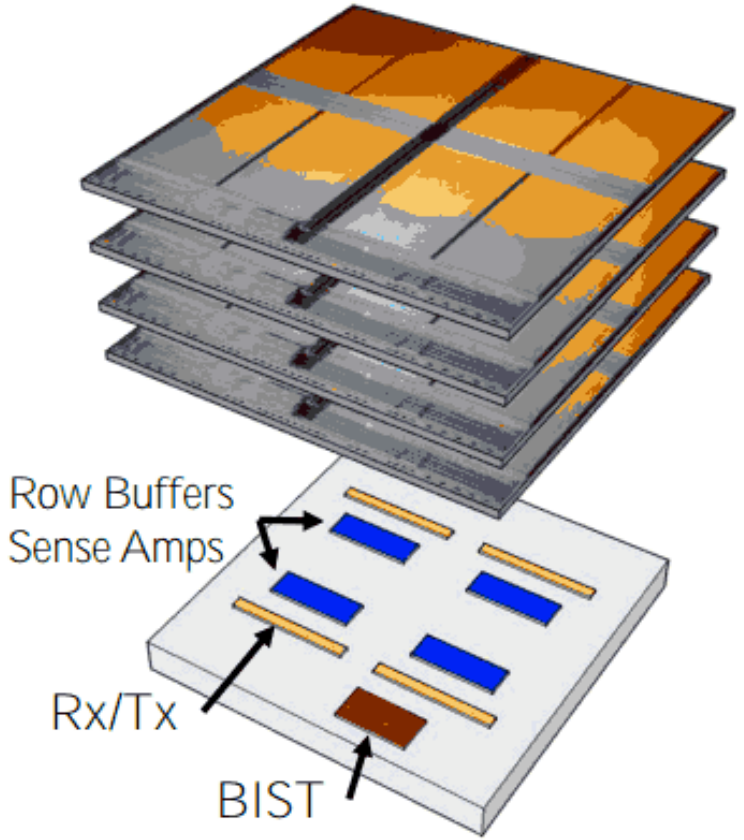


Tridimensionalizando la memoria DRAM

- Así es la DRAM de 2013:



- Y así será la DRAM de 2015:



- ¿Cómo se organiza la vieja arquitectura en este nuevo silicio?

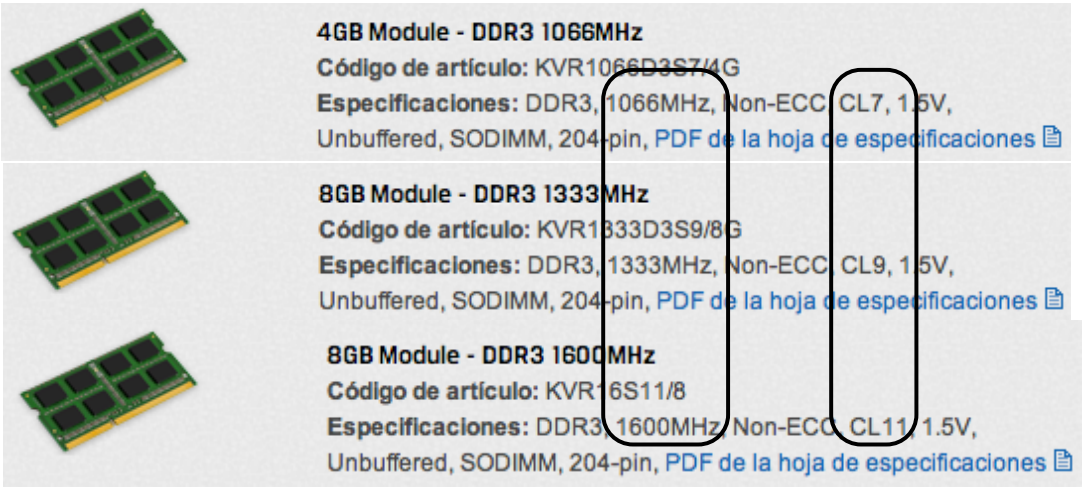
Tridimensionalizando la memoria DRAM (2)

- La memoria 3D debe aportar ideas, además de consolidar todo lo ya logrado con FPM, EDO, BEDO, SDRAM, DDR.
- Aunque hay cosas que quedan obsoletas, como la división del módulo en chips para abaratar costes de fabricación.
- Ahora, zócalo = canal = módulo = chip de memoria DDR.
- Y la arquitectura del chip debe respetarse, ya que 3D aspira a implementar mejor sus pilares de optimización:
  - Direccionamientos bidimensionales, para multiplexar coordenadas de fila y columna minimizando las líneas del bus y los pines del chip.
  - Ráfagas, para amortizar la latencia de fila al llenar líneas de caché.
  - Entrelazados, para aprovechar localidad espacial y prebúsquedas.

Lo que cuesta mejorar latencia y ancho de banda en los chips de memoria DRAM

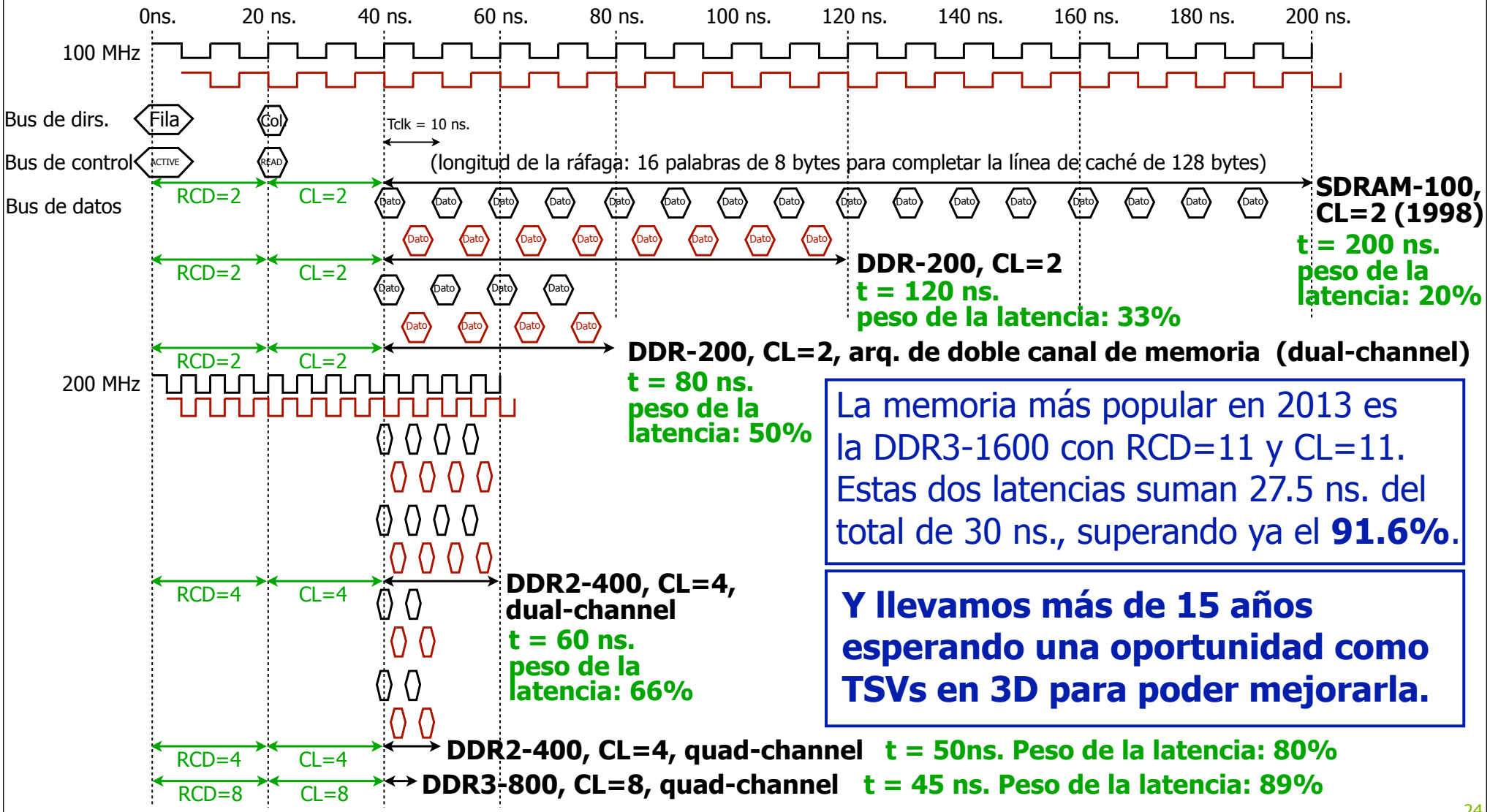
Año	Tipo de memoria	Frecuencia	Latencia CAS (en CL)	Latencia CAS (en ns.)	Ancho de banda del módulo (MB/s.)
1998	SDRAM-100	100 MHz	2	2 ciclos de 10 ns. = 20 ns.	800
2002	DDR-200	2x 100 MHz	2	2 ciclos de 10 ns. = 20 ns.	1600
2005	DDR2-400	2x 200 MHz	4	4 ciclos de 5 ns. = 20 ns.	3200
2008	DDR3-800	2x 400 MHz	8	8 ciclos de 2.5 ns. = 20 ns.	6400

En 10 años no se ha podido bajar la latencia de la saga DDR.
Y ahora, en 2013, vamos a peor:



Latencia CAS (en ns.)	A. banda (MB/s.)
7 x (1/533MHz)=13.13 ns.	8528
9 x (1/666MHz) = 13.50 ns.	10664
11 x (1/800MHz) = 13.75 ns.	12800

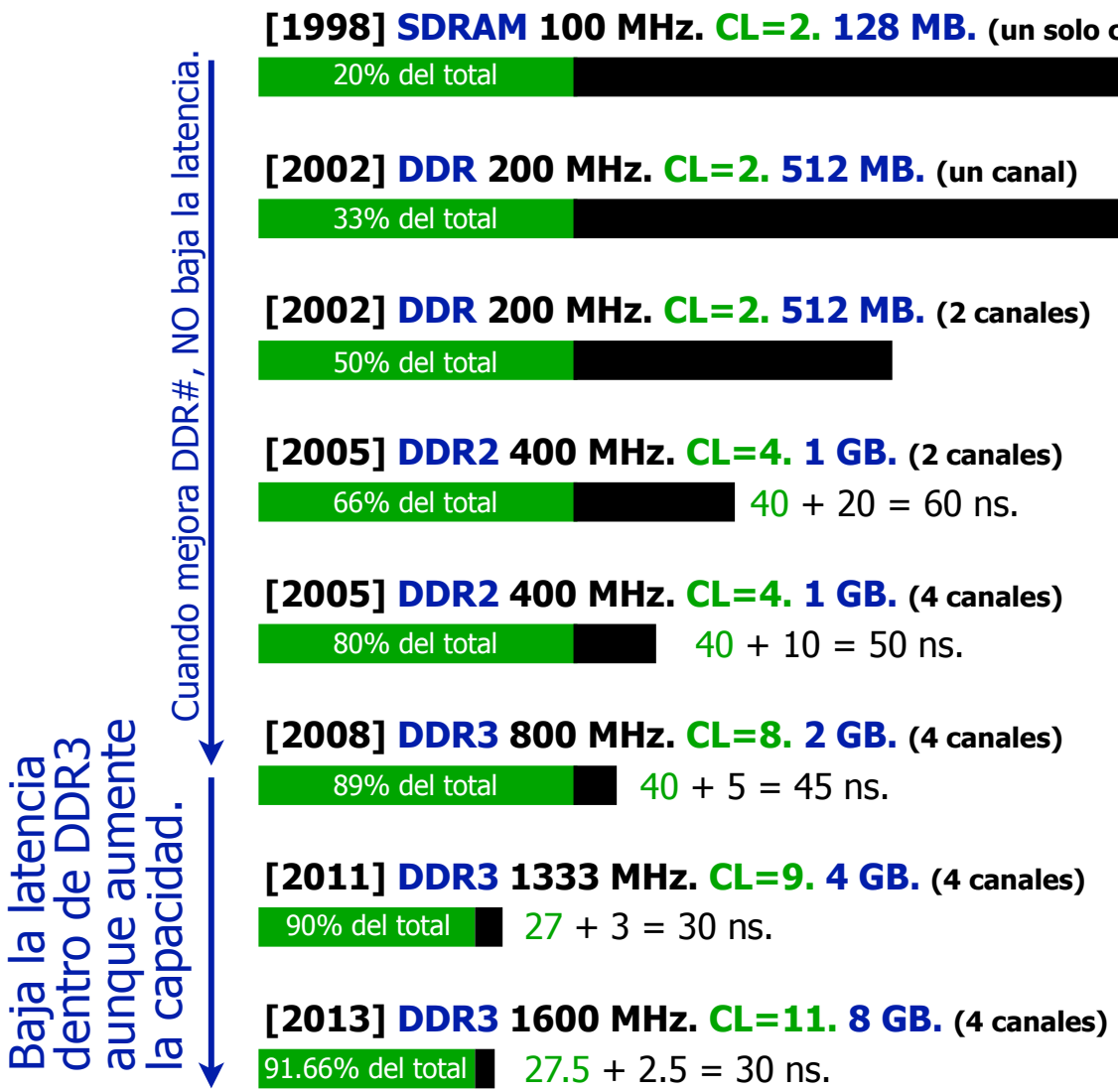
Funcionamiento de la saga de memoria DDR para servir una línea de caché de 128 bytes



La memoria más popular en 2013 es la DDR3-1600 con RCD=11 y CL=11. Estas dos latencias suman 27.5 ns. del total de 30 ns., superando ya el **91.6%**.

Y llevamos más de 15 años esperando una oportunidad como TSVs en 3D para poder mejorarla.

La capacidad (GB.) y sobre todo el interfaz (DDR#) perjudican a la latencia



Si la latencia baja históricamente 2x/década. ¿Por qué no lo hace aquí?

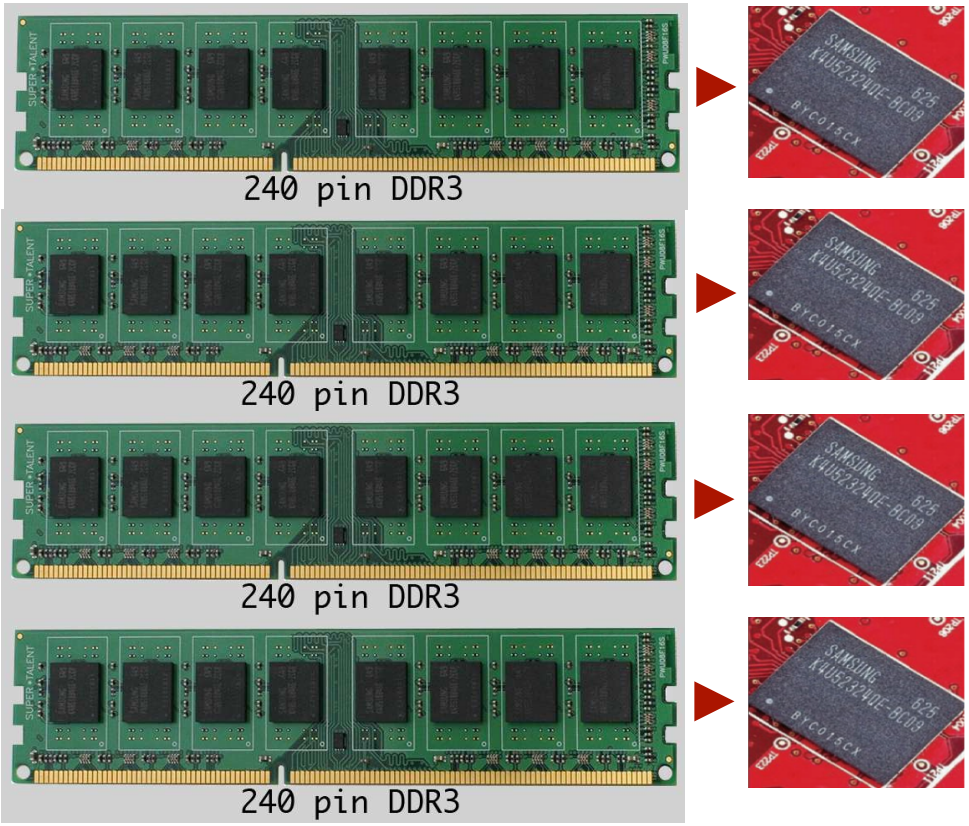
Porque "cuanto más grande, más lento".

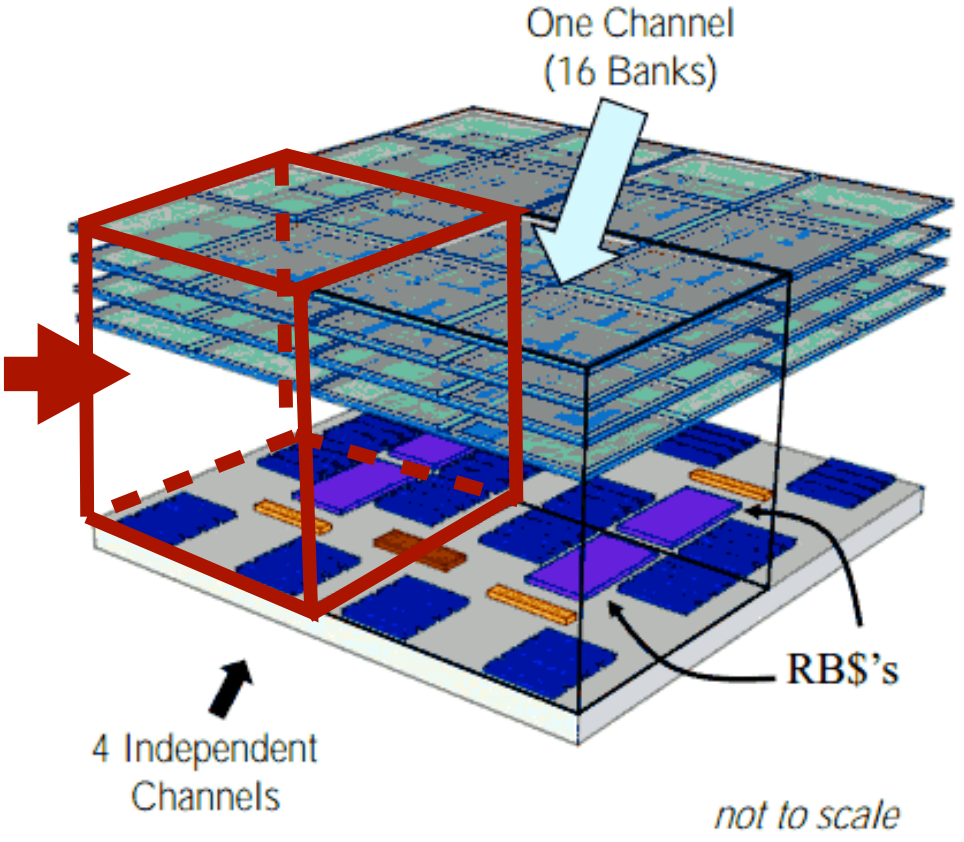
Pero... ¿qué hace al chip de memoria más grande?

Hay una respuesta lógica (# GB.) y otra tapada (DDR#).

La base del chip 3D se dedica al interfaz. Las capas, a la capacidad.

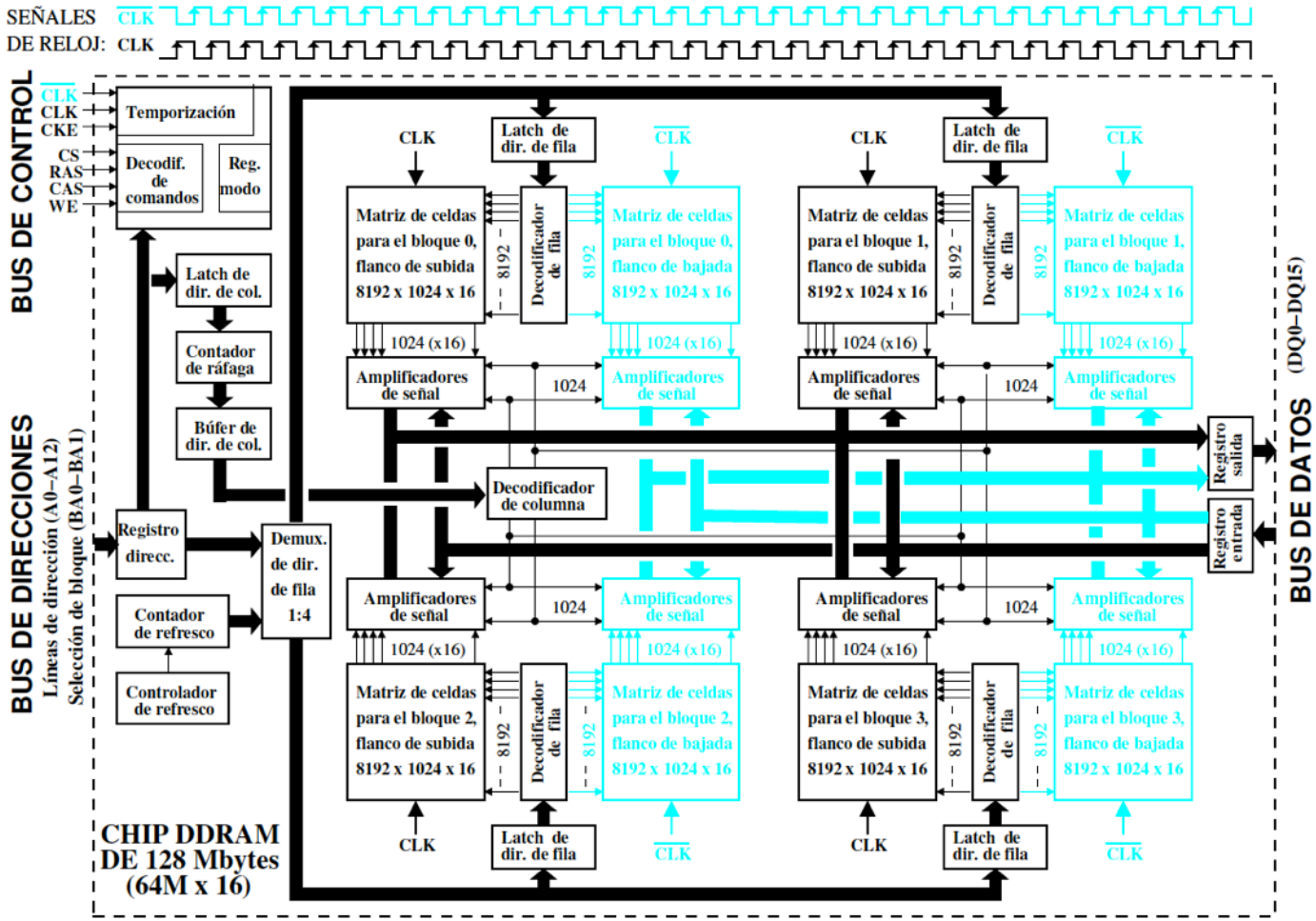
4 canales, 4 módulos, 32 chips. → 4 canales, 4 chips. → 4 canales, 1 chip 3D.



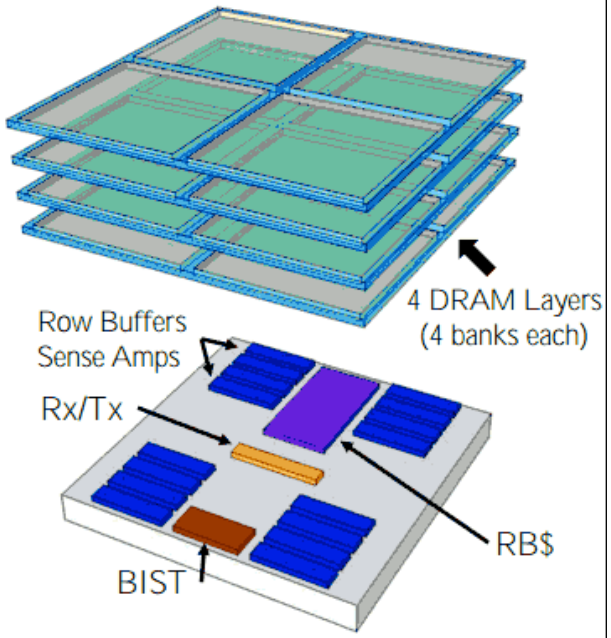


Más beneficios: Comunicaciones más cortas en 3D (y rápidas usando TSVs)

- Arquitectura de un chip de memoria (DDR#):



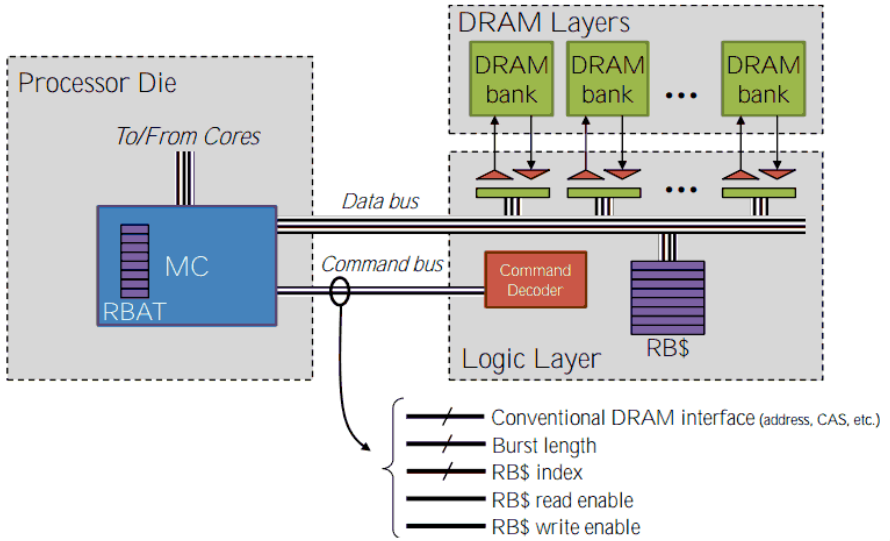
- Propuesta para su implementación 3D:



TSVs aceleran el llenado de los buffers de fila: Bajará la latencia RCD.

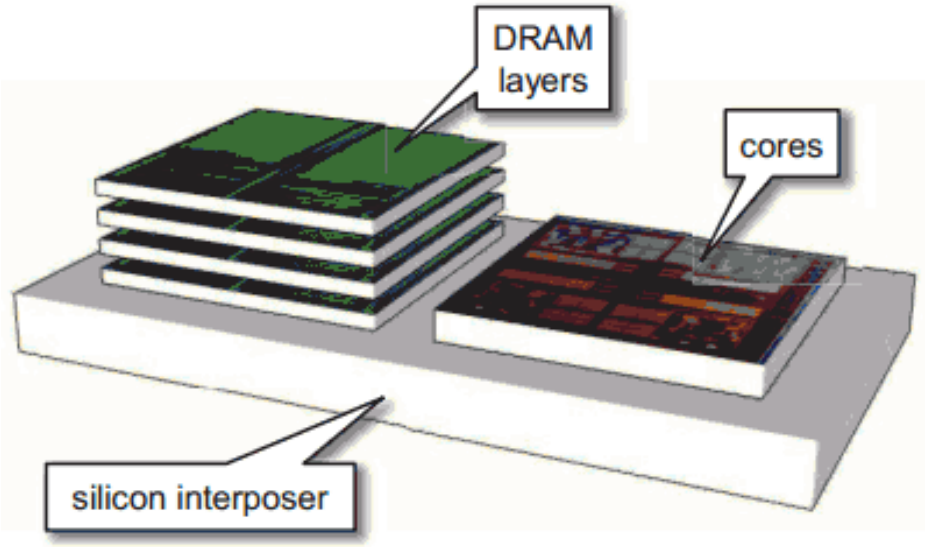
Más beneficios: Mayores entrelazados

- El entrelazado ha sido la gran baza utilizada para mejorar los anchos de banda sin que la latencia ayudase:
  - En anchura, respondiendo desde todos los chips del módulo.
  - En frecuencia, respondiendo en flanco de subida y bajada (DDR).
  - En longitud, descomponiendo la matriz de celdas en subunidades que actúan simultáneamente para sostener relojes más agresivos, mayores líneas de caché y precargas.
- Con los nuevos chips 3D:
  - En anchura, ya no es necesario.
  - En frecuencia, todo depende de la velocidad de conexión con los cores.
  - En longitud es donde más y mejor vamos a poder aprovecharlo (DDR6).

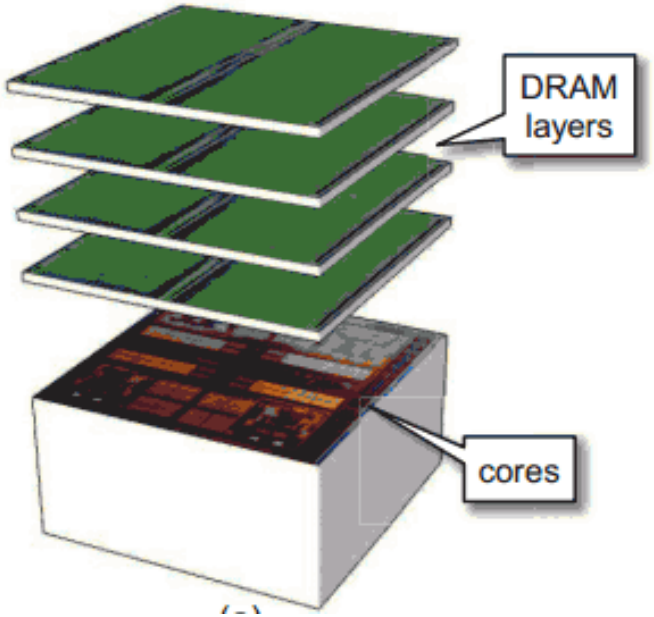


Integración junto al chip del procesador. Hoja de ruta y anuncios comerciales

- 2.5D Stacked DRAM:
  Dos chips adosados, con silicio en la frontera que los une.
  Disponible en 2014.
  - HMCC 1.0.
  - PS4 de Sony (bus de 512 líneas).



- 3D Stacked DRAM:
  Un único chip con capas heteregéneas (estim. 2015).
  - HMCC 2.0.
  - AMD + GlobalFoundries.



Construcción de los nuevos chips DRAM 3D (basada en la iniciativa HMC)

1. La DRAM se particiona en 16 torres, de forma similar a como ya hiciera la saga SDR-DDR con los bancos o matrices de celdas para explotar localidad espacial (legado de los diseños FPM de los 80).
2. Se extrae la lógica común a estas 16 torres y se sitúa en la capa base del área de silicio.
3. La DRAM se apila en configuraciones de 4 u 8 capas.
4. Se construyen las torres, y las TSVs taladrando agujeros en silicio a través de sus capas. Las TSVs serán los buses internos, y las torres los canales de la saga DDR, con gran entrelazado y escalabilidad.
5. Un bus de alta velocidad (link) conecta procesador y DRAM, que irá mejorando al pasar de la fase de transición 2.5D a 3D. Está dotado de:
    1. Conmutación avanzada.
    2. Control de memoria optimizado.
    3. Interfaz simple.
    4. 16 carriles para transmisión y recepción simultáneas. El carril aporta 10 Gb/s. (el link, 20 GB/s.).

Ambos buses resultan esenciales para aprovechar el legado de la saga DDR.

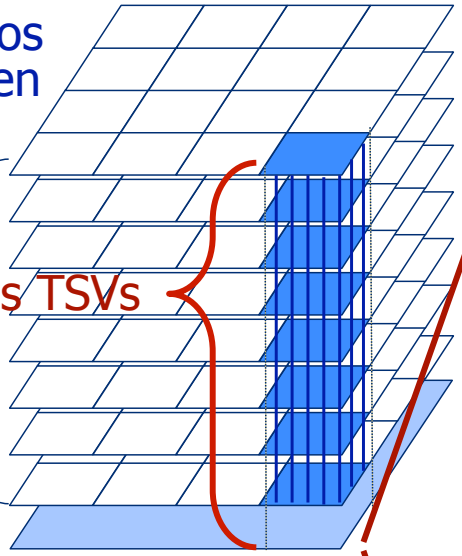
Integración junto al procesador en estructuras 3D (HMC 1.0)

Paso 1: Particionamos la matriz de celdas en 16 (futuras torres)

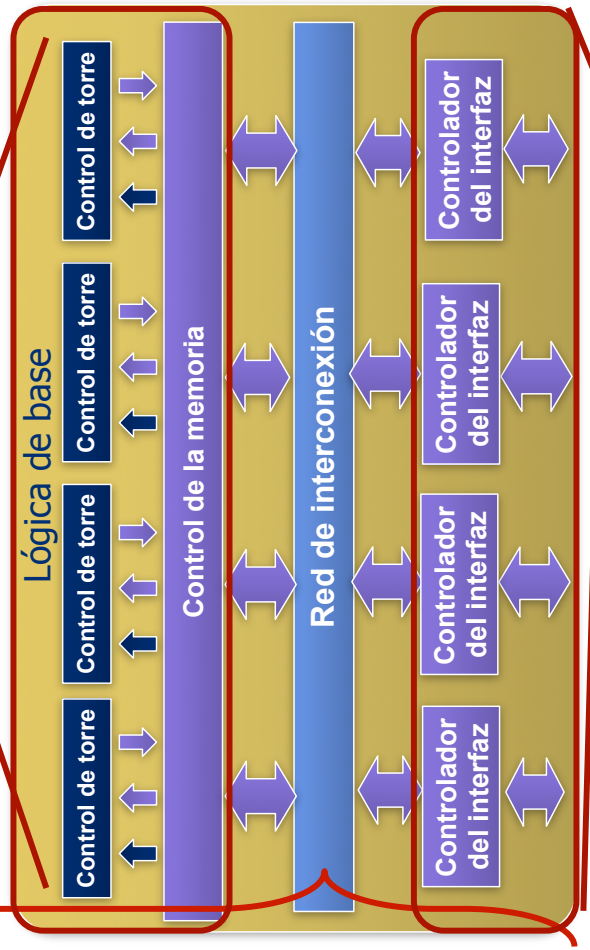
Paso 4: Trazamos las TSVs

Paso 3: Apilamos las capas DRAM

Paso 2: Situamos la lógica común en la base



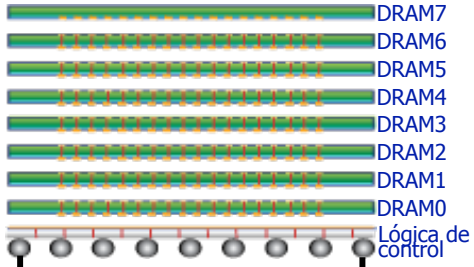
Paso 5: Se incorporan los buses que conectan el chip de memoria 3D con el procesador.



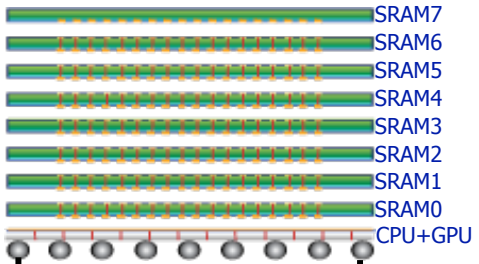
Enlaces al procesador, que puede ser otro chip 3D, pero más heterogéneo:
- Base: CPU y GPU.
- Capas: Caché (SRAM).

Los transistores SRAM conmutan más lentos a menor voltaje, por lo que la caché basará su velocidad en matrices apiladas y entrelazadas, como ya hace la DRAM.

Tecnología 3D para la memoria DRAM



Tecnología 3D para el procesador



Construcción 3D de un PC basado en Haswell

- Tenemos CPU, GPU y SRAM en diferentes proporciones para 8 modelos diferentes, con áreas de integr. siguientes:

- Y además queremos incorporar capas de DRAM.

Intel ya publicó un estudio sobre las mejores opciones (*)

- Axioma: Una DRAM es 8 veces más densa que una SRAM.
- Hipótesis: Un core ocupa un área similar al de 2 MB de L3 (se cumple en los 22nm. de Ivy Bridge si no contamos L2).
- Evaluación: 2 capas, con las siguientes alternativas (todas alcanzaron temperaturas similares):

Capa 1	Capa 2	Área	Latencia	A. banda	Consumo
2 cores + 4 MB L3	Vacía	1+0 = 1	Alta	Alto	92 W.
2 cores + 4 MB L3	8 MB L3	1+1 = 2	Media	Medio	106 W.
2 cores	32 MB. DRAM	1/2+1/2=1	Baja	Bajo	88 W.
2 cores + 4 MB L3	64 MB. DRAM	1+1 = 2	Muy baja	Muy bajo	98 W.

- Dado el peso de la latencia, damos por ganadora a la última opción. La DRAM es la mayor beneficiaria de 3D.

(*) B. Black y 14 autores más. "Die Stacking (3D) Microarchitecture", publicado en MICRO'06.

V. Mejoras del HMC 1.0 respecto a la DRAM actual

Jornadas Sarteco

17-20 Septiembre, Madrid

Velocidades de transferencia entre el controlador de memoria y el procesador

Ancho de banda en 2013 entre el controlador de memoria y el procesador (en cada dirección)	Short reach (para cables de alcance entre 20 y 25 cm)	Ultra Short reach (para líneas de circuito impreso de 5 a 8 cm)
Por cada pin	15 Gbits/s.	10 Gbits/s.
Por cada enlace HMC (16 bits)	30 GBytes/s.	20 GBytes/s.
Por cada canal de memoria (64 bits)	120 GBytes/s.	80 GBytes/s.
Para una CPU de 4 canales DRAM	No se aplica	320 GByte/s.
Para una GPU con bus de 384 bits	No se aplica	480 GByte/s.

Ancho de banda objetivo entre el controlador de memoria y el procesador (en cada dirección)	Short reach (para cables de alcance entre 20 y 25 cm)	Ultra Short reach (para líneas de circuito impreso de 5 a 8 cm)
Por cada pin	28 Gbits/s.	15 Gbits/s.
Por cada enlace HMC (16 bits)	56 GBytes/s.	30 GBytes/s.
Por cada canal de memoria (64 bits)	224 GBytes/s.	120 GBytes/s.
Para una CPU de 4 canales DRAM	No se aplica	480 GBytes/s.
Para una GPU con bus de 384 bits	No se aplica	720 GBytes/s.

Una comparativa en ancho de banda con las tecnologías existentes en la actualidad

- En un PC con memoria de cuádruple canal (256 bits):
  - [2013] Una CPU con DDR3 a 4 GHz (2x 2000 MHz): 128 Gbytes/s.
  - [2014] Una CPU con HMC 1.0 (primera generación): 320 Gbytes/s.
  - [2015] Una CPU con HMC 2.0 (segunda generación): 448 Gbytes/s.
- En una tarjeta gráfica con un bus de 384 bits:
  - [2013] Una GPU con GDDR5 a 7 GHz (2x 3500 MHz): 336 Gbytes/s.
  - [2014] Una GPU con 12 chips de 32 bits integrados con near-mem. HMC 1.0 llegaría a **480 Gbytes/s.** (6 canales HMC 1.0 de 80 GB/s.).
  - [2015] Una GPU con HMC 2.0 (112 GB/s.) lograría **672 Gbytes/s., lo que duplica el ancho de banda frente a la tecnología GDDR5 más avanzada de finales de 2013.**

(*) Tomando las estimaciones de ancho de banda dadas por el HMCC 1.0 y 2.0 (20 y 28 GB/s., respectivamente, por cada enlace de 16 bits en cada sentido de la transmisión). Nvidia ya anunció en GTC'13 anchos de banda de 1 TB/s. para Volta.

Lo que cuesta a cada tecnología alcanzar 640 GB/s.

Circuitería necesaria	DDR3L-1600	DDR4-3200	Stacked DRAM HMC 1.0
Ancho de banda unidirecc. (GB/s.)	12.8 por módulo	25.6 por módulo	20 por enlace de 16 bits
Items necesarios para 640 GB/s.	50 módulos	25 módulos	32 enlaces (8 chips 3D)

Actividad eléctrica	DDR3L-1600	DDR4-3200	Stacked DRAM HMC 1.0
Líneas eléctricas activas necesarias	143 por módulo	148 por módulo	270 por chip
Número total de líneas activas	7150	3700	2160 **(ahorro el 70%)**

Consumo energético	DDR3L-1600	DDR4-3200	Stacked DRAM HMC 1.0
Vatios (W.)	6.2 por módulo	8.4 por módulo	5 por enlace
Consumo total para 640 GB/s.	310 W.	210 W.	160 W. **(ahorro el 50%)**

Espacio ocupado en placa base	DDR3L-1600	DDR4-3200	Stacked DRAM HMC 1.0
Área del módulo (ancho x alto)	165 mm. x 10 mm. = 1650 mm2		1089 mm2 por chip
Area total ocupada para 640 GB/s.	825 cm2	412.5 cm2	43.5 cm2 **(ahorro 95%)**

Mejoras publicadas por otros fabricantes

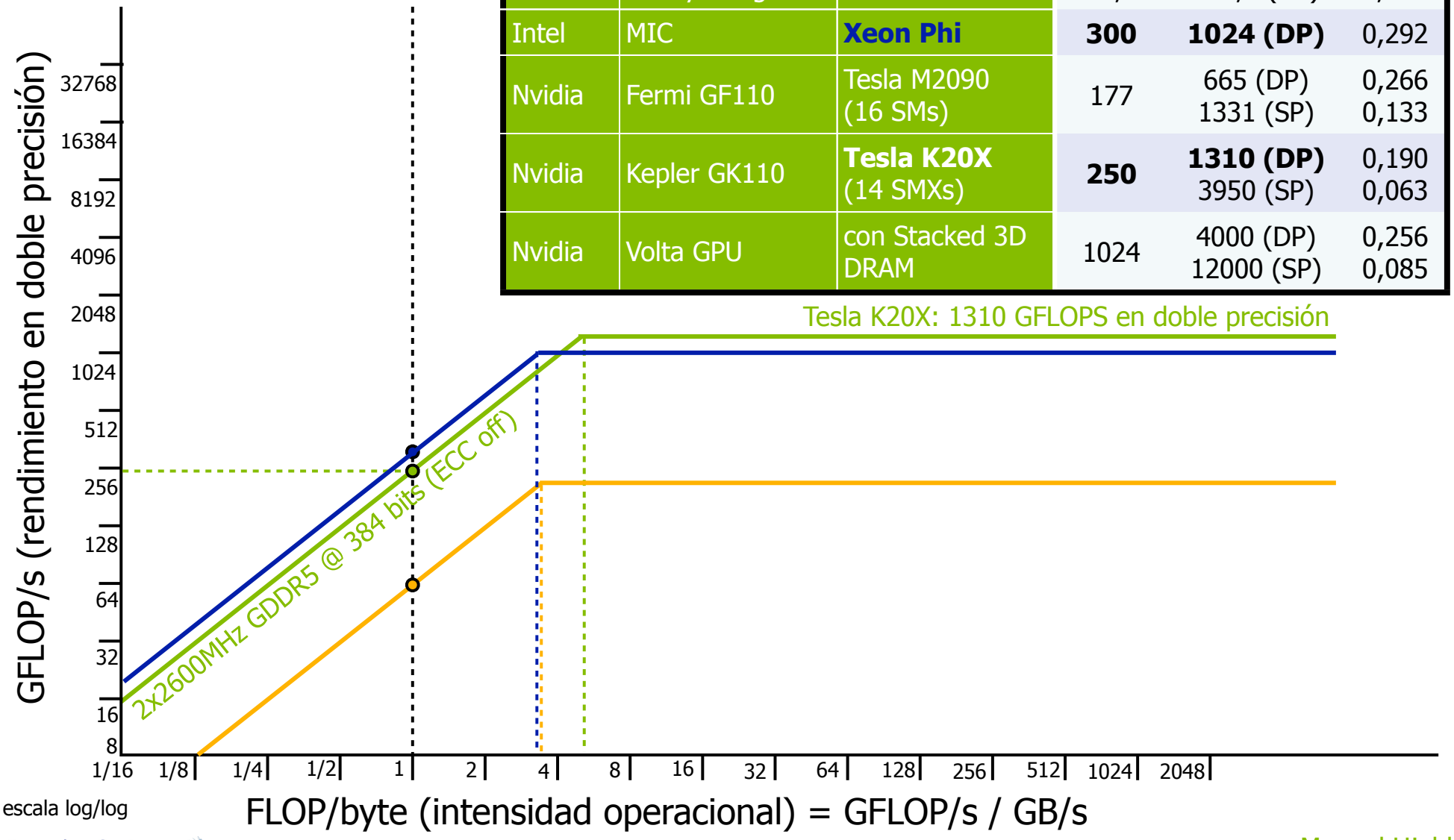
- [2008] El prototipo de Tezzaron Semiconductors reduce la latencia CAS, la latencia RAS y el RAS to CAS Delay un 32.5% respecto a la tecnología 2D para esa misma memoria.
- [2009] El chip de 8 GB. 3D DDR3 de Samsung aumenta el ancho de banda desde los 1066 MB/s. hasta los 1600 MB/s., un 50% más. El consumo pasivo se reduce un 50%, y el activo un 25%.
- [2012] La implementación de IBM para Micron dentro del estándar HMC 1.0 alcanza anchos de banda de 128 GB/s. consumiendo 10 vatios (comparado con los 82 vatios que consumen los 15 DIMMs de DDR3-1333 equivalentes).

VI. Impacto sobre las GPUs

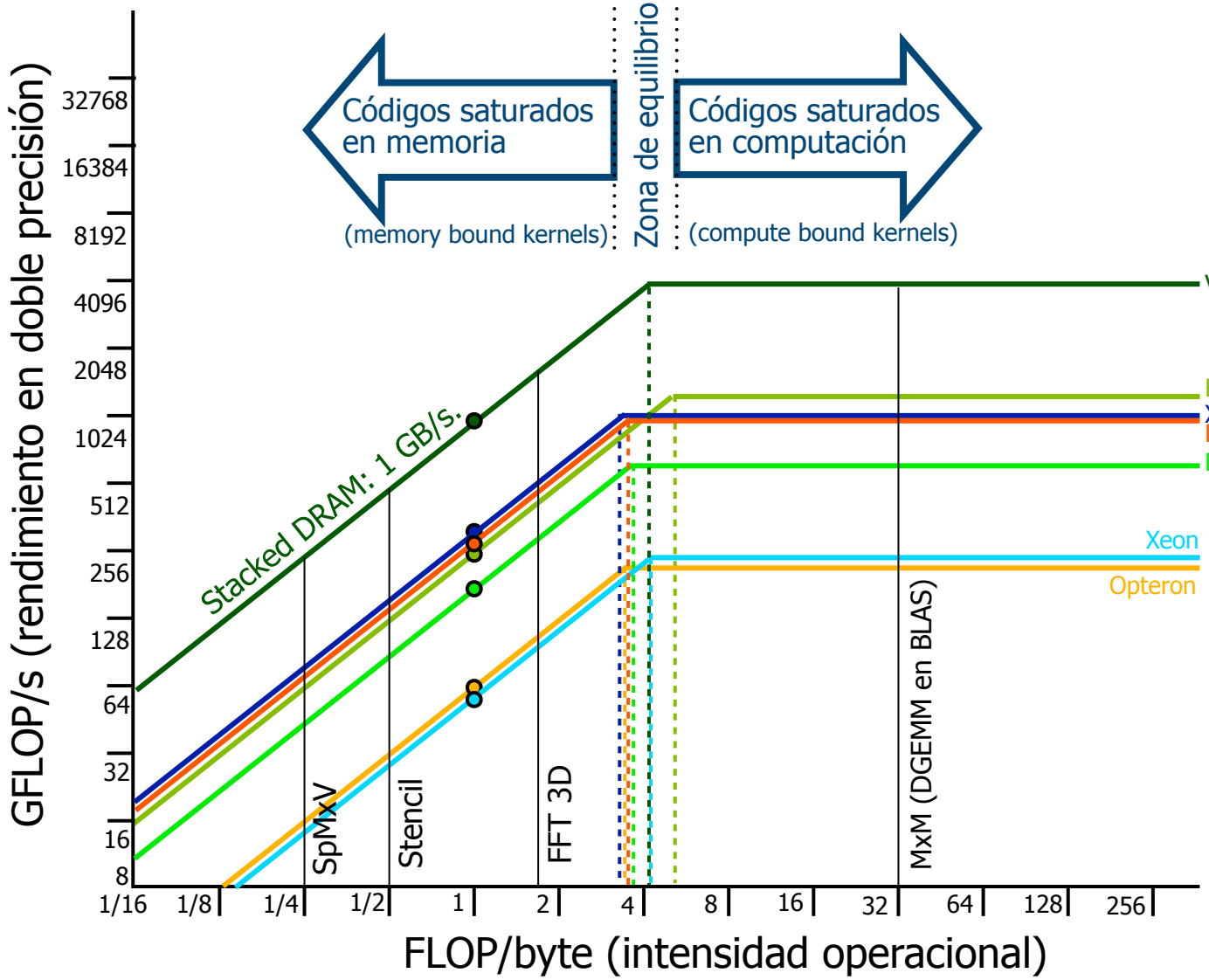
Jornadas Sarteco
17-20 Septiembre, Madrid

Plataformas a comparar

Marca	Microarquit.	Modelo	GB/s.	GFLOP/s.	Byte/ FLOP
AMD	Bulldozer	**Opteron 6284**	**59,7**	**217,6 (DP)**	0,235
AMD	Souther Islands	Radeon HD7970	288	1010 (DP)	0,285
Intel	Sandy Bridge	Xeon E5-2690	51,2	243,2 (DP)	0,211
Intel	MIC	**Xeon Phi**	**300**	**1024 (DP)**	0,292
Nvidia	Fermi GF110	Tesla M2090 (16 SMs)	177	665 (DP) 1331 (SP)	0,266 0,133
Nvidia	Kepler GK110	**Tesla K20X** (14 SMXs)	**250**	**1310 (DP)** 3950 (SP)	0,190 0,063
Nvidia	Volta GPU	con Stacked 3D DRAM	1024	4000 (DP) 12000 (SP)	0,256 0,085

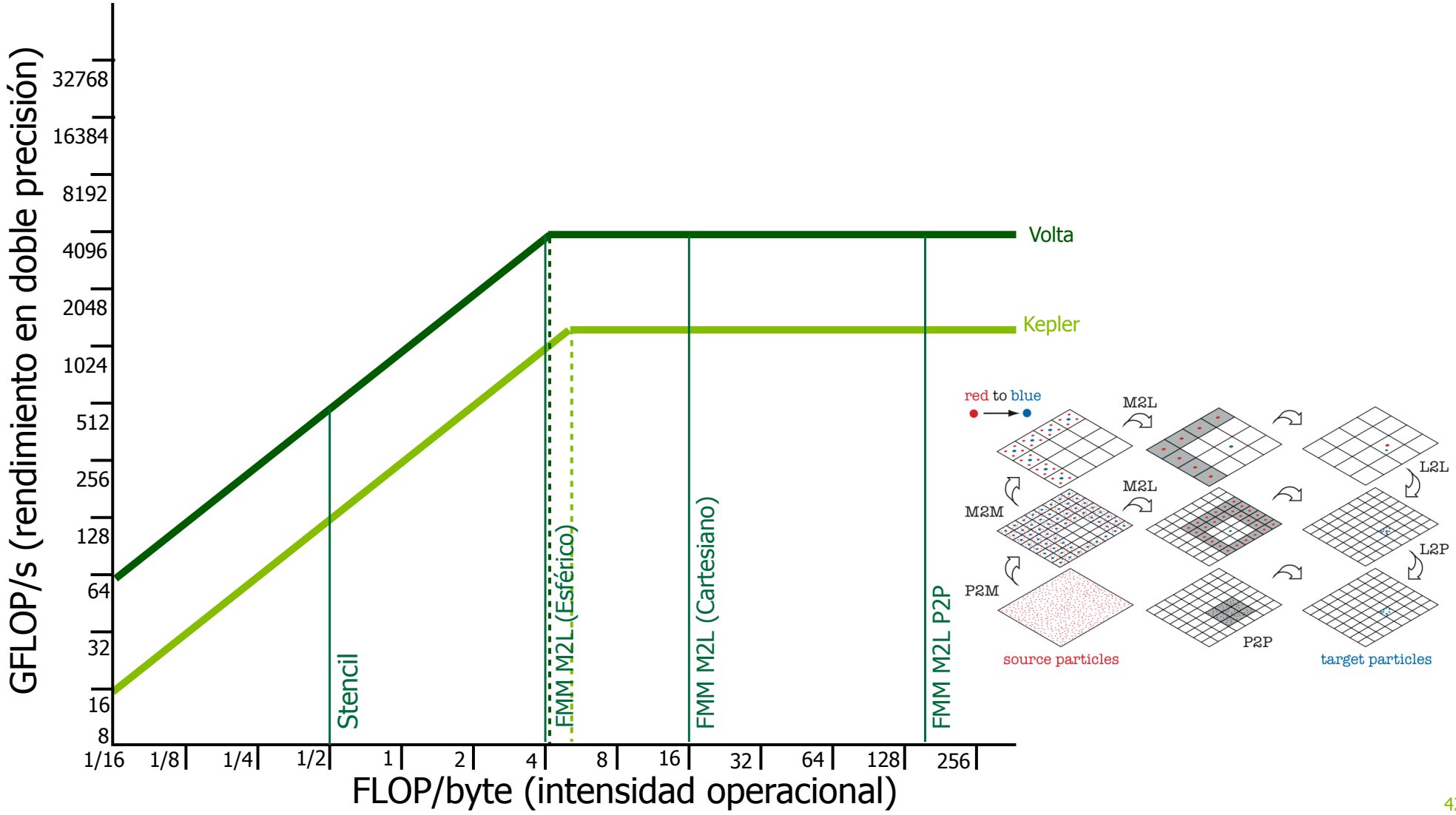


El modelo Roofline: Análisis de códigos. Hardware vs. Software



Procesador	GB/s.	GFLOP/s.	B/FLOP
Opteron	60	217 (DP)	0,235
Radeon	288	1010 (DP)	0,285
Xeon	51	243 (DP)	0,211
Xeon Phi	300	1024 (DP)	0,292
Fermi	177	665 (DP) 1331 (SP)	0,266 0,133
Kepler	250	1310 (DP) 3950 (SP)	0,190 0,063
Volta	1024	4000 (DP) 12000 (SP)	0,256 0,085

El modelo Roofline: Evolución del software. Caso estudio: FMM (Fast Multipole Method)



Conclusiones

- La mayoría de códigos acelerados en GPU tienen su limitador de rendimiento en el acceso a memoria y han evolucionado protegiéndose frente a ella.
- Aparece la integración 3D que favorece sobre todo a la memoria, permitiéndole:
  - Reducir la latencia de su estructura de celdas, que llevaba dos décadas estancada.
  - Mejorar la saga DDR#, aumentando: Capacidad (3D), comunicaciones (TSVs) y entrelazado (torres).
- La próxima generación de chips 3D será más heterogénea, para migrar hacia el SoC (System-on-Chip) integrando ya juntos todos los protagonistas: CPU, GPU, SRAM y DRAM.

Para más información

- El consorcio Hybrid Memory Cube:
  - http://www.hybridmemorycube.org (especificación 1.0 disponible).
- Un completo índice de empresas ligadas a la memoria 3D (productos, materiales, equipos, investigación, etc):
  - http://www.tezzaron.com/technology/3D_IC_Summary.html
- Las especificaciones del módulo de memoria 3D DRAM de Viking Technology, ya disponible comercialmente:
  - http://www.vikingtechnology.com/uploads/dram_stacking_technology.pdf
- El modelo Roofline, una delicia de artículo:
  - S. Williams, A. Waterman, D. Patterson. "Roofline: An Insightful Visual Performance Model for Multicore Architectures". Communications of the ACM, Abril, 2009.

Agradecimientos

- A los ingenieros de Nvidia, por compartir ideas, material, diagramas, presentaciones. Y a la firma, por su financiación.
- A Lorena Barba (CUDA Fellow), por su gran aportación al modelo Roofline y su aplicación al algoritmo FMM.
- A Scott Stevens y Susan Platt (Micron) por el acceso a material técnico del HMCC, incorporado a esta presentación bajo su explícito permiso.
- A Gabriel Loh (GaTech, ahora en AMD) por prestarme las figuras de sus artículos sobre tecnologías 3D Stacked DRAM precursoras.
- Y sobre todo, a los organizadores de las Jornadas SARTECO por invitarme a estar aquí.

Muchas gracias por vuestra atención

- Siempre a vuestra disposición en el Dpto. de Arquitectura de Computadores de la Universidad de Málaga:
  - e-mail: ujaldon@uma.es
  - Teléfono: +34 952 13 28 24
  - Página Web: http://manuel.ujaldon.es
- Actividades como CUDA Fellow:
  - 7 en Europa.
  - 11 en América.
  - 4 en Sudáfrica.
  - 4 en Australia.
  - 2 en Nueva Zelanda.