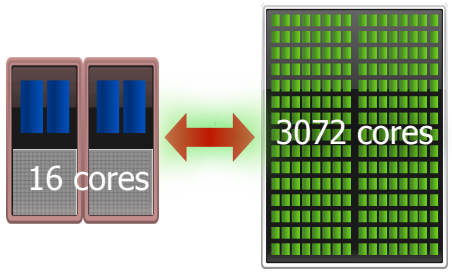
Computación heterogénea y su programación

Manuel Ujaldón
Nvidia CUDA Fellow
Departmento de Arquitectura de Computadores
Universidad de Málaga (España)

Indice de contenidos [38 diapositivas]

1. Introducción a la computación heterogénea

Computación heterogénea: El eje central de la próxima generación hardware



Utilizar tanto la CPU como la GPU
Cada procesador se encarga de ejecutar aquello en lo que es más eficiente

¿En qué aspectos es mejor cada procesador?

A favor de la CPU:

- Cachés muy rápidas.
- Buen manejo de las dependencias de datos y control.
- Muchos paradigmas para ejecutar hilos y procesos.
- Alto rendimiento sobre un único hilo de ejecución.
- Mejor cobertura de E/S.

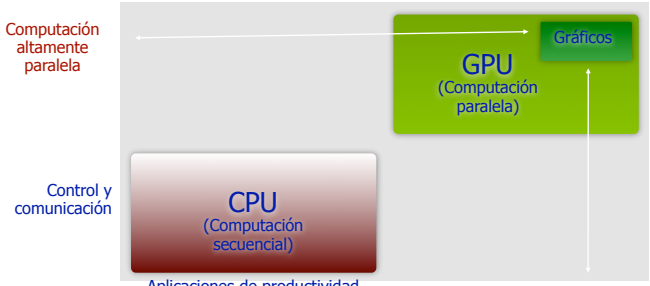
PARALELISMO DE TAREAS

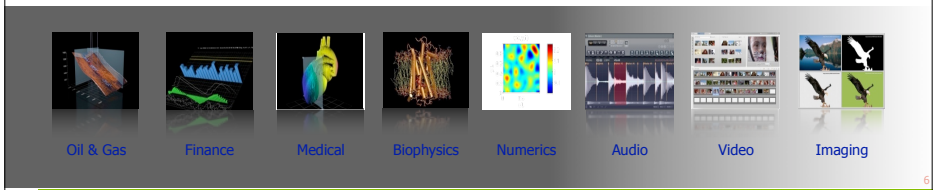
A favor de la GPU:

- Núcleos computacionales muy numerosos.
- Paralelismo masivo.
- Hardware dedicado para cálculos matemáticos.
- Alto rendimiento ejecutando tareas paralelas.
- DRAM muy veloz.

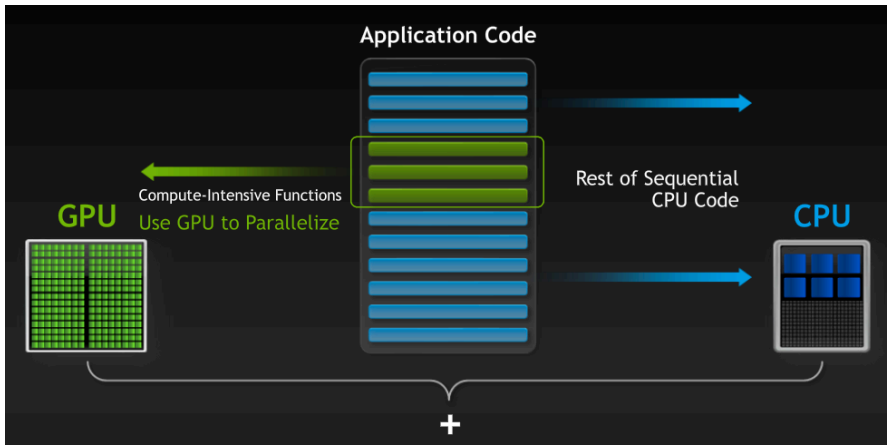
PARALELISMO DE DATOS

La escalabilidad es la principal diferencia

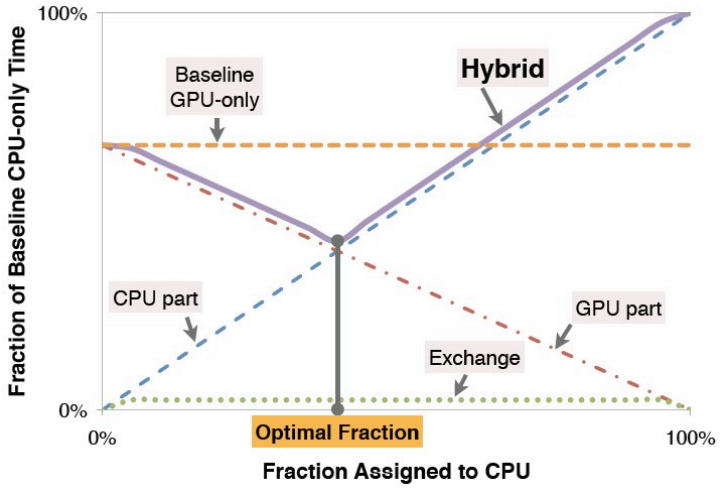




La mejor estrategia consiste en ver la CPU y la GPU como mundos **complementarios**



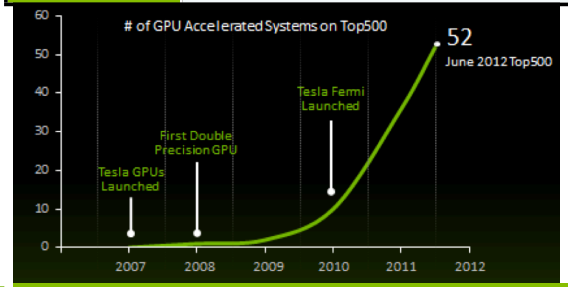
Y recuerda mantener ambos procesadores ocupados, pero no **demasiado** ocupados



2. Supercomputadores con arquitectura heterogénea

Los supercomputadores basados en GPUs van cobrando relevancia según el Top500.org

	Noviembre, 2010	Junio, 2011	Noviembre, 2011	Junio 2012
Nvidia	**10**	**12**	**35**	**52**
ATI Radeon	1	2	2	2
Cell	6	5	2	2
Intel Xeon Phi	0	0	0	1
Total	**17**	**19**	**39**	**57**

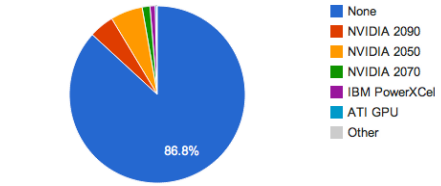


A finales de 2010 había sólo 10 supercomputadores basados en GPUs, y Tesla Fermi fue entonces un punto de inflexión significativo. - Ahora nos aprestamos a ver el efecto que produce Kepler durante 2013.

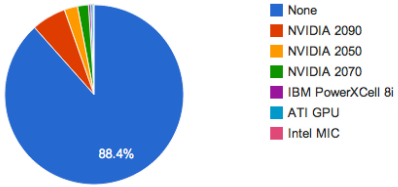
Contribución de las GPUs al Top 500 por modelo comercial y su peso en el rendimiento

Accelerator/Co-Processor	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
None	442	88.4	107103147.87	140616821.21	12004293
NVIDIA 2090	31	6.2	5626155.43	11318582.04	572312
NVIDIA 2050	12	2.4	7262060	14304324.8	532298
NVIDIA 2070	10	2	1775173.41	3265356.28	146174
IBM PowerXCell 8i	2	0.4	1168500	1537632	136800
ATI GPU	2	0.4	360710	647429.2	26268
Intel MIC	1	0.2	118600	180992	9800

Accelerator/Co-Processor Performance Share



Accelerator/Co-Processor System Share



Los supercomputadores más potentes se construyen mayoritariamente con GPUs

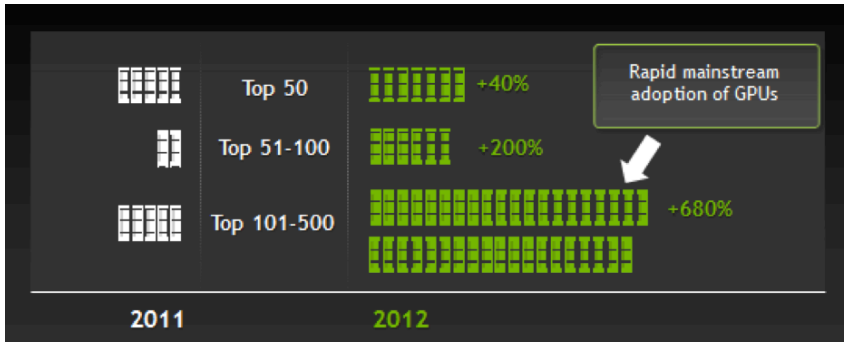
18 de Junio de 2012:

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom / 2011 IBM	1572864	16324.75	20132.66	7890.0
2	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
3	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	786432	8162.38	10066.33	3945.0
4	Leibniz Rechenzentrum Germany	**SuperMUC** - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR / 2012 IBM	147456	2897.00	3185.05	3422.7
5	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
6	DOE/SC/Oak Ridge National Laboratory United States	**Jaguar** - Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090 / 2009 Cray Inc.	298592	1941.00	2627.61	5142.0
7	CINECA Italy	**Fermi** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	163840	1725.49	2097.15	821.9
8	Forschungszentrum Juelich (FZJ) Germany	**JuQUEEN** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	131072	1380.39	1677.72	657.5
9	CEA/TGCC-GENCI France	**Curie thin nodes** - Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR / 2012 Bull	77184	1359.00	1667.17	2251.0
10	National Supercomputing Centre in Shenzhen (NSCS) China	**Nebulae** - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0

9 de Noviembre de 2011:

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.0
4	National Supercomputing Centre in Shenzhen (NSCS) China	Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0
5	GSIC Center, Tokyo Institute of Technology Japan	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.6
6	DOE/NNSA/LANL/SNL United States	Cray XE6, Opteron 6136 8C 2.40GHz, Custom / 2011 Cray Inc.	142272	1110.00	1365.81	3980.0
7	NASA/Ames Research Center/NAS United States	SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband / 2011 SGI	111104	1088.00	1315.33	4102.0
8	DOE/SC/LBNL/NERSC United States	Cray XE6, Opteron 6172 12C 2.10GHz, Custom / 2010 Cray Inc.	153408	1054.00	1288.63	2910.0
9	Commissariat a l'Energie Atomique (CEA) France	Bull bullx super-node S6010/S6030 / 2010 Bull	138368	1050.00	1254.55	4590.0
10	DOE/NNSA/LANL United States	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.0

Crecimiento de los supercomputadores basados en GPUs durante el año 2011



- Cuanto menos caro y sofisticado sea un supercomputador, más probable resulta encontrar GPUs en su interior.
- Las GPUs democratizan el segmento HPC.

3. Arquitecturas heterogeneas de bajo coste

Variantes arquitecturales: GPU, IGP, EPG, HPU

Dificultades para la integración conjunta entre CPU y GPU

- Las GPUs presentan un alto grado de complejidad, con necesidades de consumo y refrigeración más exigentes.
- El uso de la memoria es radicalmente diferente en CPU y GPU, y su gestión en la capa software también lo ha sido. Se vislumbra un cuello de botella en el ancho de banda.
- El ciclo de desarrollo de una GPU es más rápido (2 años), que el de una CPU (4 años o más). En la práctica, esto supone apostar por diseños microprogramados o cableados.
- La GPU ha crecido más en Mt. y en área de integración, y todo ello a pesar de que se fabrica con anchura de puerta de 28 nm. vs 22 nm. en CPU. Una parte proporcional de este patrimonio deberá delegarse a la CPU.

Las GPUs son minoría en el contexto global de los chips gráficos

- La lista de completa de personajes muestra un camino inequívoco hacia la integración conjunta de CPU y GPU:
  - GPU: Graphics Processing Unit. El procesador autónomo, fabricado casi en exclusiva por Nvidia (GeForce) y AMD/ATI (Radeon).
  - IGP: Integrated Graphics Processor. El procesador integrado en el puente norte del juego de chips de la placa base. Como Intel es líder en la fabricación de juegos de chips, también lo es de IGPs. Constituyen una fase de transición y caminan hacia la extinción.
  - EPG: Embedded Processor Graphics. El camino hacia la integración conjunta, que aún no es posible en 32-22 nm. Es el líder de ventas en gama baja.
  - HPU: Heterogeneous Processing Unit. Integración final de una CPU y una GPU en un mismo chip. El destino final de nuestra historia.

Relevancia económica de los procesadores gráficos

- Cada trimestre se venden entre 15 y 18 millones de tarjetas gráficas. Anualmente, las ventas superan los 60 millones, siendo uno de los mercados más lucrativos en la industria informática.
- El gasto medio es de 300 dólares por unidad, entre:
  - Gama alta para servidores y estaciones: Unos 1.500 dólares.
  - Gama media para PC: Entre 100 y 250 dólares.
  - Gama diversa para multitud de instrumentos científicos.
- No contabilizamos en estos números la amplia gama baja de soluciones integradas (chips gráficos), que elevan las ventas a más del doble (ver siguiente diapositiva).

El mercado de los chips gráficos y su peso en el contexto global de la circuitería para PC

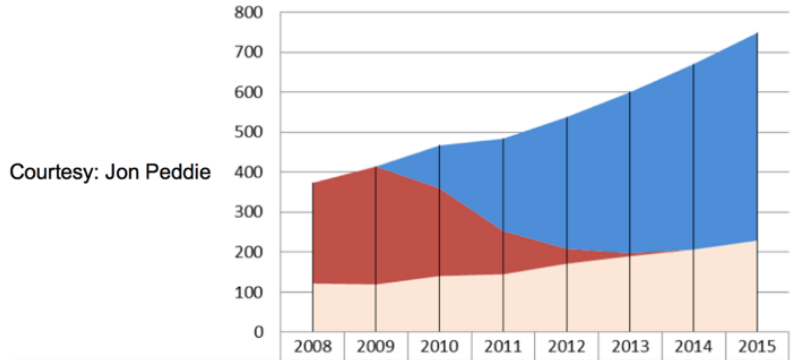
- Número total de chips gráficos vendidos en un trimestre:
  - 124 millones el último trimestre de 2011.
  - 138.5 millones el penúltimo trimestre de 2011.
  - 114 millones el último trimestre de 2010.
- El mercado de los chips gráficos sigue creciendo, aunque se notan los efectos de la crisis.
- Comparado con las ventas de PCs, que fueron de 93.5 millones en el último trimestre de 2011, tenemos 1.5 chips gráficos por cada PC, y este porcentaje del 150% viene creciendo de forma sostenida desde el 115% en 2001.

Por marcas

	Nvidia	AMD/ATI	Otros
Cuota de mercado en el último trimestre de 2010	60,5%	39,0%	0,5%
Cuota de mercado en el penúltimo trimestre de 2011	59,7%	39,9%	0,4%
Cuota de mercado en el último trimestre de 2011	63,4%	36,3%	0,3%
Variación respecto al trimestre anterior	+ 3,7%	-3,6%	-0,1%
Variación en 2011 respecto al mismo período del año anterior	+2,9%	-2,7%	-0,2%

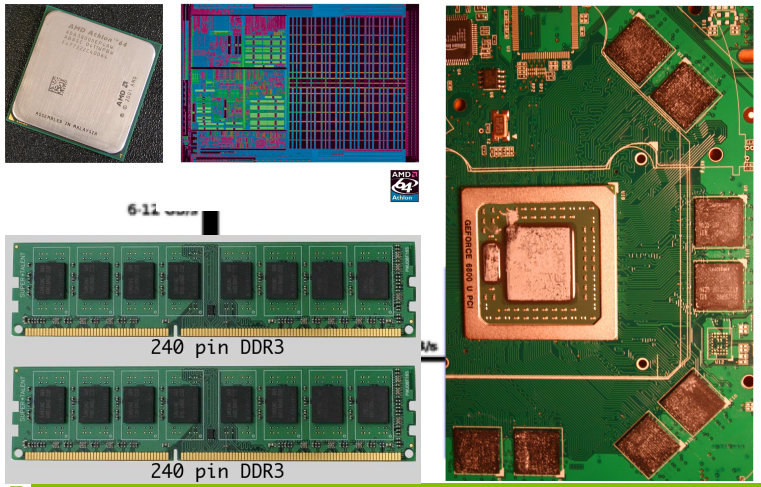
- Nvidia duplica en ventas a AMD/ATI.
- El resto de firmas tiene una presencia residual.

Predicciones de ventas hasta 2015: Desaparece IGP en favor de EPG y más tarde HPU



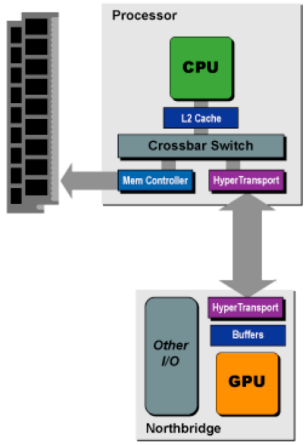
	2008	2009	2010	2011	2012	2013	2014	2015
HPU & EPG, CAGR 2010-2015: 37.11%	-	-	107.07	230.65	328.63	402.37	463.29	518.86
IGP, CAGR 2010-2015: -88%	251.30	294.73	219.30	108.14	37.57	7.36	0.19	0.01
Discrete GPU, CAGR 2010-2015: 10.28%	121.88	119.52	140.51	144.91	171.46	190.22	206.76	229.22

GPU (Graphics Processing Unit): El caso típico que originó todo

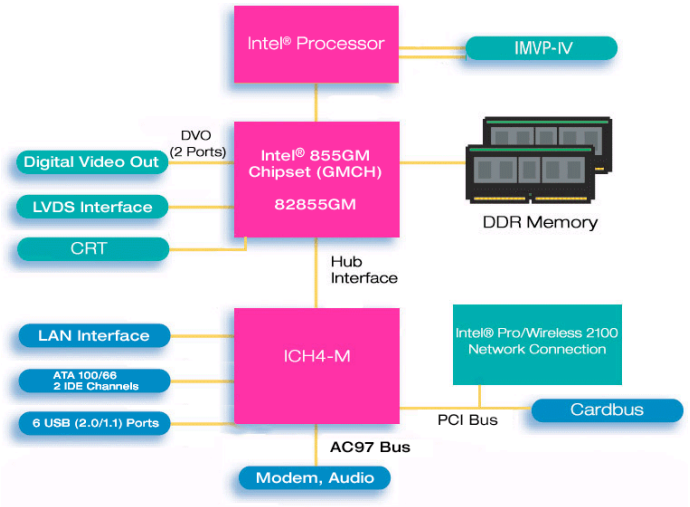


IGP (Integrated Graphics Processor) en AMD

- AMD:

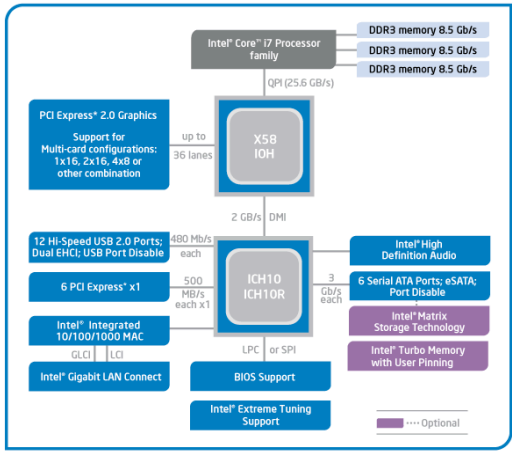


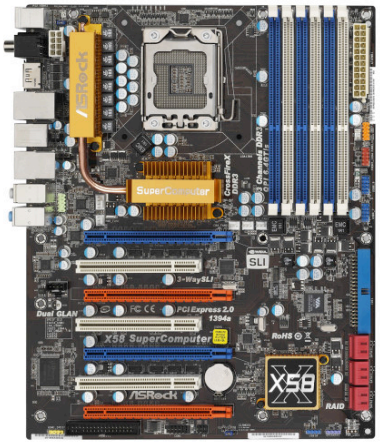
IGP (Integrated Graphics Processor) en Intel



La evolución hacia Sandy Bridge e Ivy Bridge [Intel, 2010-12]

La primera generación de la familia i3-i5-i7 (Nehalem) son IGPs

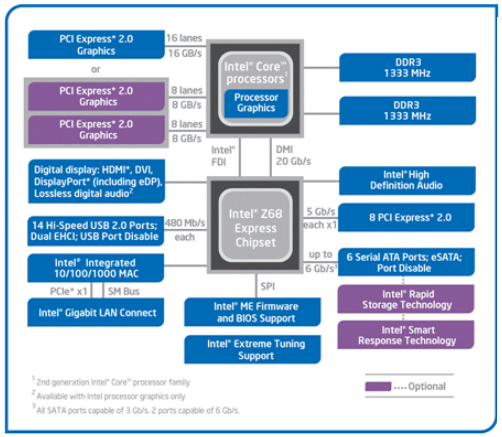




Intel® X58 Express Chipset Block Diagram

La segunda generación de i5 (Arrandale, Clarkdale) ya es EPG (Embedded Processor)

- Debut en el mercado por parte de Intel: Enero de 2010.



Intel® Z68 Express Chipset Platform Block Diagram

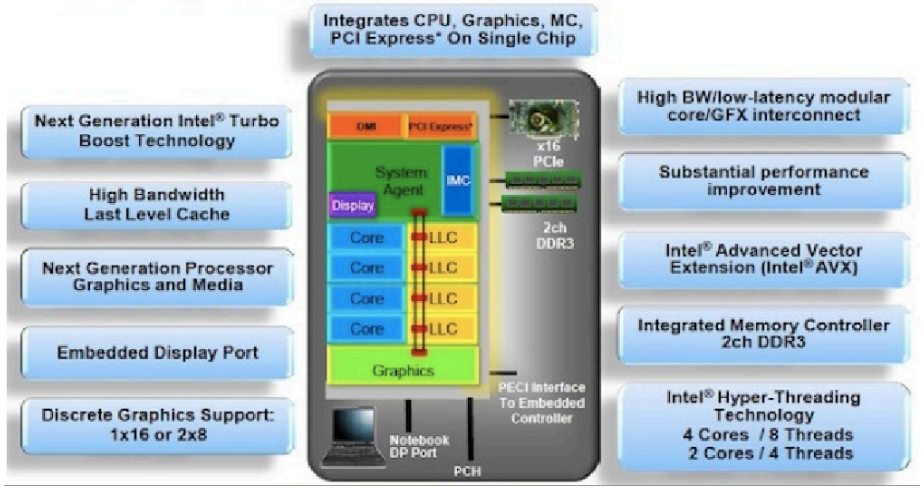
La 2ª generación del i7 también tiene algún modelo EPG, pero ya es menos frecuente

- Listado de modelos de Intel i7 según www.intel.com a mediados de 2012.

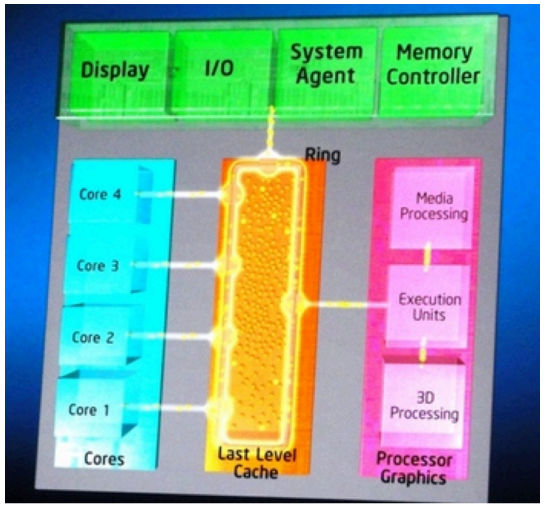
Compare Select: All | None	Product Name	Status	Embedded Options Available	Max TDP	Recommended Customer Price	Graphics	Intel® Turbo Boost Technology
Select	Intel® Core™ i7-3930K Processor (12M Cache, up to 3.80 GHz)	Launched	No	130 W	$583 - $594		2.0
Select	Intel® Core™ i7-3820 Processor (10M Cache, up to 3.80 GHz)	Launched	No	130 W	$294 - $305		2.0
Select	Intel® Core™ i7-2700K Processor (8M Cache, up to 3.90 GHz)	Launched	No	95 W	$332 - $342	Intel® HD Graphics 3000	2.0
Select	Intel® Core™ i7-2600S Processor (8M Cache, up to 3.80 GHz)	Launched	No	65 W	$294 - $305	Intel® HD Graphics 2000	2.0
Select	Intel® Core™ i7-2600K Processor (8M Cache, up to 3.80 GHz)	Launched	No	95 W	$317 - $326	Intel® HD Graphics 3000	2.0
Select	Intel® Core™ i7-2600 Processor (8M Cache, up to 3.80 GHz)	Launched	Yes	95 W	$294 - $305	Intel® HD Graphics 2000	2.0

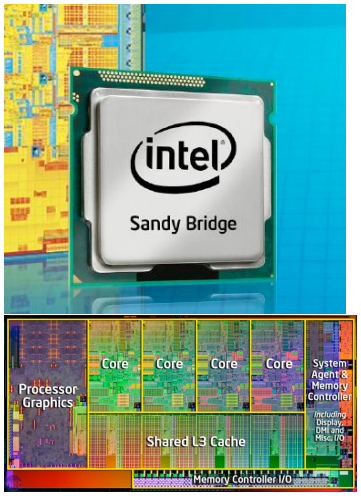
Como era de esperar, el EPG es el mas barato de la gama.

HPU (Heterogeneous Processing Unit). Ya es Sandy/Ivy Bridge (2012)



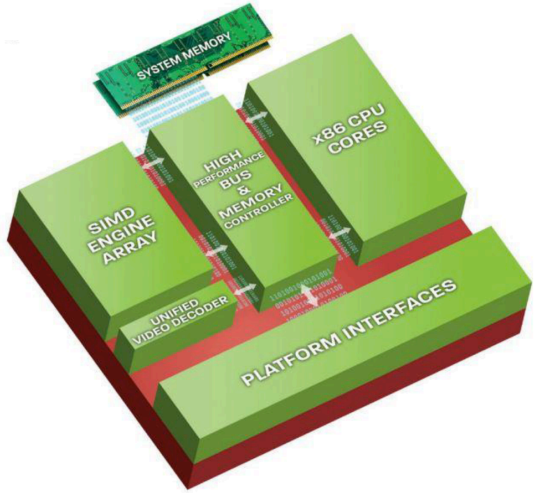
Sandy Bridge: Algunos detalles



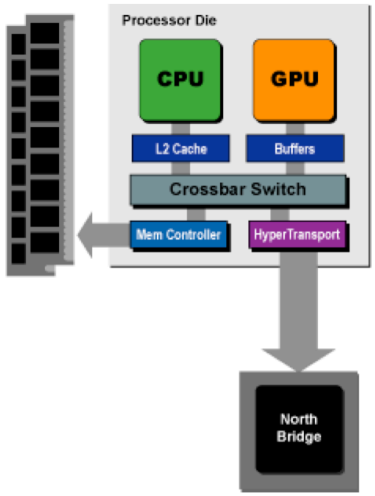


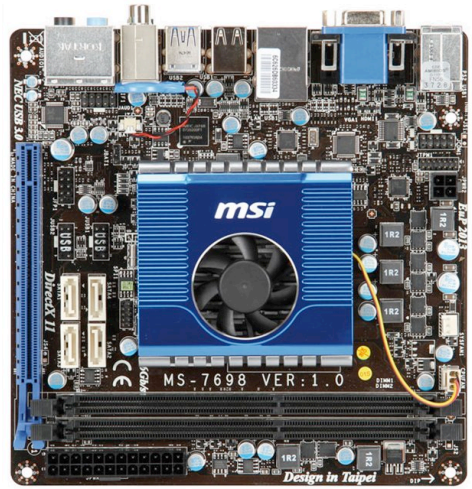
Las arquitecturas Fusion [AMD, 2010-12]

Fusion: La idea

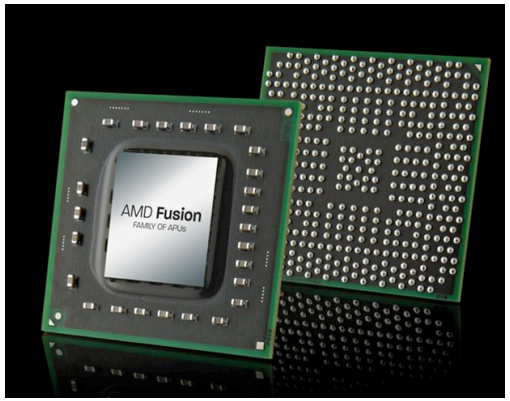


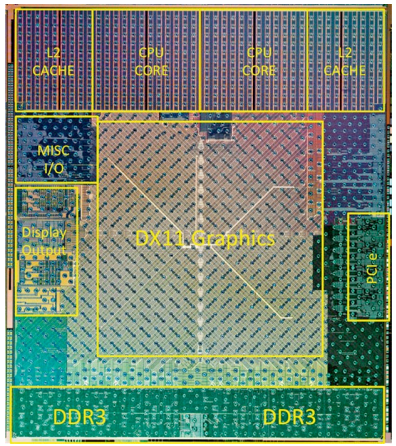
Los modelos HPU de AMD (Fusion) salieron antes al mercado. Ejemplo: E-350





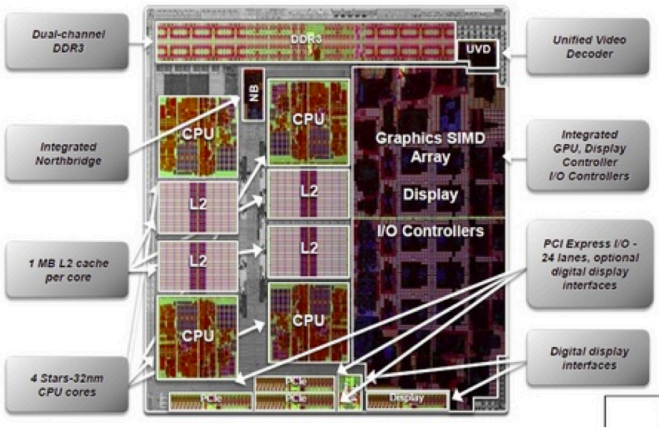
Más detalles de Fusion sobre el modelo comercial E-350





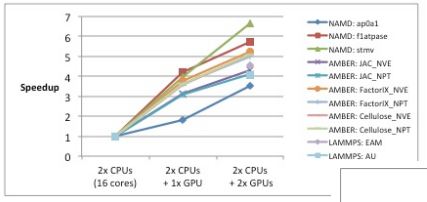
La segunda generación de HPUs en AMD son las plataformas Llano (APUs según AMD)

Llano

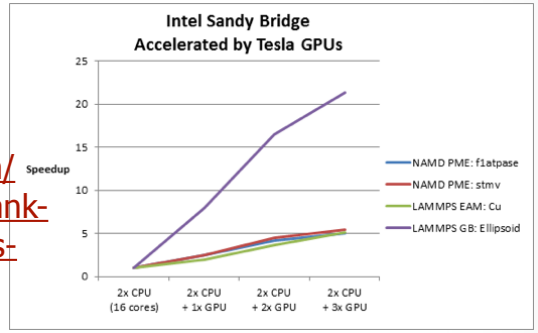


Tiempos de ejecución y resultados experimentales

Pruebas realizadas en Nvidia: Dell PowerEdge R720 con CPUs Sandy Bridge y GPUs Teslas



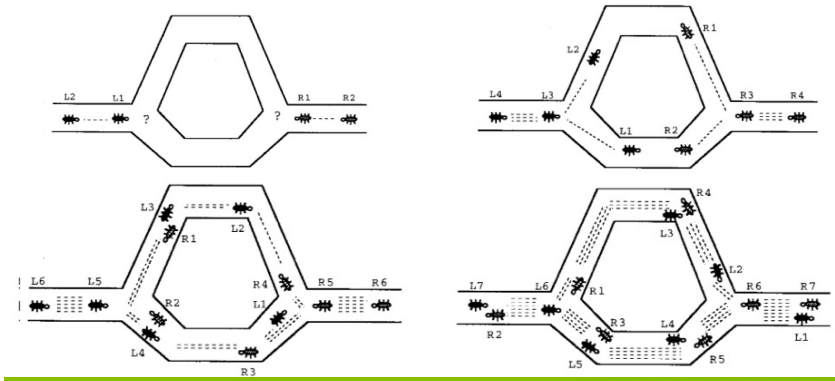
Dos conocidas aplicaciones de dinámica molecular: NAMD and LAMMPS.



Autor: Sumit Gupta.
Más información en:
http://blogs.nvidia.com/2012/03/tesla-gpus-crank-intel-sandy-bridge-cpus-up-to-11

Un algoritmo bioinspirado: ACO

- Código ACO (Ant Colony Optimization). Inspirado en la manera que tienen las hormigas de encontrar siempre la distancia más corta entre la comida y el hormiguero.



¿Para qué sirve el algoritmo ACO?

- Utilizado para resolver numerosos problemas de optimización que acontecen en la vida real.
- Uno de los más populares es el TSP (Travelling Salesman Problem), en la que un comercial tiene que visitar "m" ciudades interconectadas por una red de carreteras, y se trata de encontrar la ruta que minimice la distancia recorrida.
- Hay variantes más sofisticadas, por ejemplo, otorgando pesos para la "calidad de la carretera" en cada una de las aristas de conexión entre ciudades, pero aquí veremos el caso más sencillo.

El proceso de simulación computacional

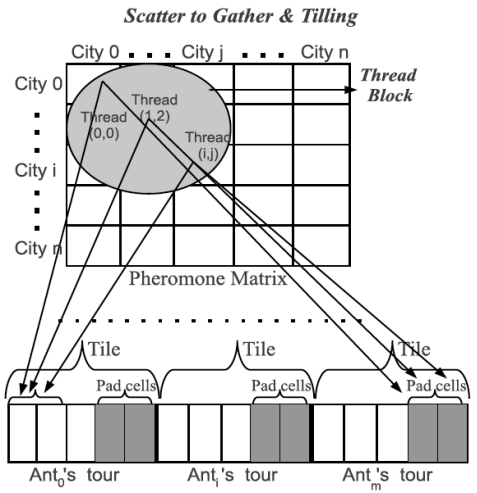
- "m" hormigas construyen sus rutas en paralelo.
- Inicialmente, las hormigas se sitúan aleatoriamente.
- Luego aplican la regla proporcional aleatoria:
- $$p_k(r,s) = \begin{cases} \dfrac{[\tau(r,s)] \cdot [\eta(r,s)]^\beta}{\sum_{u \in J_k(r)} [\tau(r,u)] \cdot [\eta(r,u)]^\beta}, & \text{if } s \in J_k(r) \\ 0, & \text{otherwise} \end{cases}$$
- donde:
  - p_k (r, s) es la probabilidad de que la hormiga "k" de la ruta "r" elija moverse a la ciudad "s".
  - sigma cuantifica el nivel de feromonas de esa arista en la ruta.
  - n es la inversa de la distancia.
  - beta determina la importancia relativa de la feromona frente a la distancia.
  - J_k(r) es el conjunto de ciudades que quedan por visitar por la hormiga "k" posicionada en la ciudad "r".

El banco de pruebas utilizado para nuestra evaluación

- Nuestro conjunto de datos de entrada procede de la librería TSP. La longitud del mejor tour corresponde a la mínima solución encontrada por el algoritmo.

Conjunto de datos de entrada	Nombre del grafo	Número de ciudades en el grafo	Longitud del mejor tour encontrado
Tamaño pequeño	d198	198	15780
	a280	280	2579
	lin318	318	42029
	pcb442	442	55778
Tamaño medio	rat783	783	8806
	pr1002	1002	259045
	pcb1173	1173	56892
	d1291	1291	50801
Tamaño grande	pr2392	2392	378032

Detalle de la implementación en CUDA



Resultados experimentales obtenidos

- Las GPUs se comportan mejor, y CUDA gana a OpenCL.
- La CPU escala mejor (2571x), vs. la mejor GPU (93x con CUDA en Tesla S2050). [escala = T(large)/T(small)]

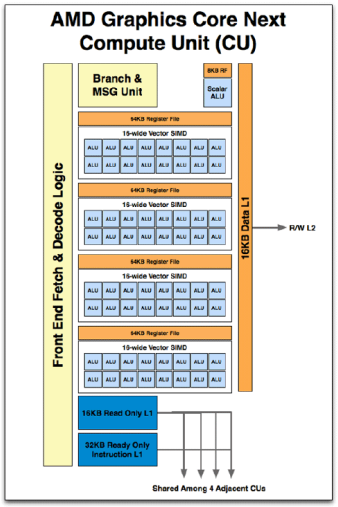
Modelo del procesador	Coste (trimestre)	Clase (procesador)	Lenguaje	Tiempo de ejecución (ms.)		
				Pequeño	Medio	Grande
Intel Westmere	$500 (Q4'09)	CPU (Xeon E5620)	C	43,01	3538,89	110573,04
Nvidia Fermi	$1500 (Q4'09)	GPU (Tesla C2050)	CUDA	4,59	209,96	5131,11
Nvidia Fermi	$1500 (Q4'09)	GPU (Tesla C2050)	OpenCL	5,53	247,94	5991,91
ATI Cypress	$1500 (Q1'10)	GPU (FirePro V8800)	OpenCL	23,16	395,88	8989,04
AMD Llano	$150 (Q1'10)	HPU (E-350)	OpenCL	3200,09	174320	No se dispone de suficiente memoria para esta ejecución
ATI Redwood	$150 (Q1'10)	HPU (HD 6310)	OpenCL	200,00	10620,4	
AMD LLano	$150 (Q2'11)	HPU (A6-3420)	OpenCL	1228,50	67690,9	
ATI Redwood	$150 (Q2'11)	HPU (HD 6520)	OpenCL	148,48	7787,60	
ATI Redwood	$100 (Q1'11)	GPU (HD 6650M)	OpenCL	80,09	3529,49	

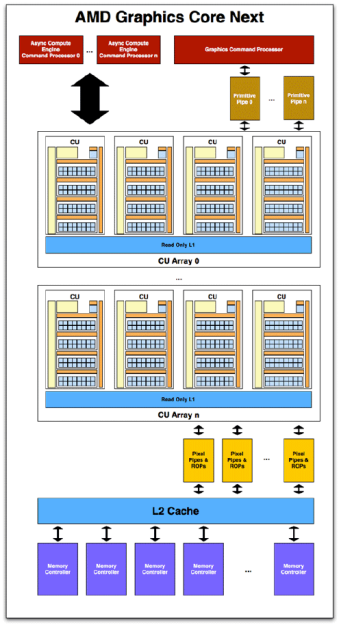
4. La segunda generación de arquitecturas heterogéneas

Los diseños de los principales fabricantes

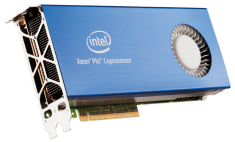
- Graphics Core Next de AMD (2012/13).
- Xeon Phi de Intel (2012/13).
- Denver de Nvidia (aún en desarrollo).

Graphics Core Next (AMD): Es un VLIW SIMD.





Xeon Phi (Intel): Arquitectura



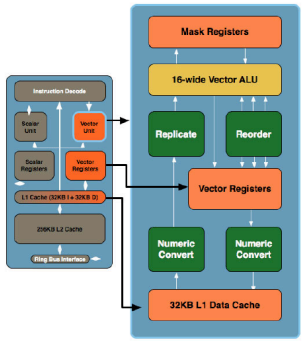
- La versión inicial se ha fabricado a distancia de integración de 22 nm. con transistores 3D tri-gate.
- Primer prototipo en alcanzar la frontera del TFLOPS en DP, con un modelo de 64 cores a 2 GHz (2 x 64 x 8 GFLOPS).
- La principal diferencia arquitectural con respecto a Atom o Xeon es que cada core tiene una unidad vectorial de 512 bits en su unidad de punto flotante, capaz de manejar 8 operaciones SIMD en doble precisión.
- Su topología es la de un gran bus en forma de anillo con 512 bits en cada dirección, las cachés ubicadas en el centro del área de integración y los cores en la periferia.
- Controlador de memoria integrado con soporte para ECC.

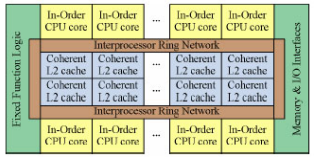
Xeon Phi (Intel): Programación

- Basada en herramientas de programación ya consolidadas para el conjunto de instrucciones x86. Por lo tanto, pueden utilizarse los mismos paradigmas y utilidades de desarrollo que para los modelos de procesadores Xeon tradicionales:
  - OpenMP.
  - MPI.
  - TBB (Intel's Threading Building Blocks).
  - MKL (Math Kernel Library).

La arquitectura ha sufrido muchos cambios

- Precedentes:
  - Larrabee (2008).
  - Knights Corner y Knights Ferry (2010).
  - MIC (2011).
- Cómo era Larrabee en sus orígenes:
  - Procesamiento escalar x86.
  - Procesamiento vectorial de anchura 16.
  - Jerarquía de caché L1 y L2 muy similar a la de las arquitecturas "Core" y "Core 2" coetáneas.
  - Anillo de comunicaciones bidireccional de 512 bits circundando la caché L2 ubicada en el centro neurálgico del chip.





Conclusiones

- Dentro de un PC disponemos de mayor potencial computacional en la vertiente gráfica que en la vertiente de propósito general, y se impone un trasvase de recursos de la primera hacia la segunda a medida que baja su precio.
- Las alternativas Sandy-Ivy Bridge y Fusion-Llano son competitivas para bajo coste y bajo consumo, aunque están limitadas por el ancho de banda de la memoria. Optimizaciones "tipo coalescing" (vectorización) resultan decisivas.
- El final del viaje es la integración conjunta de todos los recursos hardware en un único chip (SoC = System on Chip), y la necesidad de programadores que sepan maximizar su explotación desde la vertiente software.