III. HARDWARE CUDA

La Tecnología CUDA esta soportada por una gran variedad de tarjetas gráficas con chipset de NVIDIA y comercializadas por distintos fabricantes.

Las distintitas soluciones han sido diseñadas para satisfacer las necesidades particulares de cada nicho comercial. Aunque las características difieren, el objetivo permanece el mismo, acelerar la ejecución de código en Paralelo.

Las tarjetas gráficas de NVIDIA que soportan la arquitectura CUDA pertenecen principalmente a 3 familias:

- Familia GeForce
- Familia Quadro
- Familia Tesla

1. FAMILIA GEFORCE

La familia GeForce está diseñada fundamentalmente para satisfacer las necesidades gráficas de los juegos informáticos y las aplicaciones de fotografía y vídeo.

Las tarjetas que soportan la tecnología CUDA son aquellas que disponen de al menos 256 MB de memoria gráfica. Existen varias series distintas GeForce como 8, 9, 100 y 200 que se dividen según que sea para ordenador de sobremesa o portátil.

En los comienzos de desarrollo CUDA se realizaron aplicaciones de cálculo basadas en GeForce 8800, sin embargo en la actualidad ha quedado relegada. De hecho esta tarjeta ya no se encuentra disponible en el mercado.

2. FAMILIA QUADRO

Las soluciones NVIDIA QUADRO son herramientas de cálculo y visualización para resolver problemas gráficos y computacionales. Al estar compuestas de varias GPUs, proporcionan niveles de rendimiento altos.

Es de destacar el uso de la tarjeta FX 5600 en el desarrollo de aplicaciones de cálculo.

En la actualidad están más enfocadas hacia el tratamiento de gráficos e imágenes en entornos muy exigentes por ejemplo con AUTOCAD. Algunas soluciones disponibles:

- NVIDIA Quadro FX 5800
- NVIDIA Quadro FX 5600
- NVIDIA Quadro FX 4800
- NVIDIA Quadro FX 4600

3. FAMILIA TESLA

La familia Tesla se ha desarrollado con el objeto de proporcionar soporte a necesidades altas de cálculo computacional.

La primera serie fue la 8, seguida de la 10. En la actualidad se ha lanzado la serie 20 que aparte del aumento de la capacidad de rendimiento del Hardware mediante el aumento del número de GPUs, se ha introducido una nueva arquitectura llamada Fermi que cambia la estructura de distribución de las distintas clases de memoria e introduce la memoria Caché para mejorar el rendimiento.

La solución de alta computación (HPC) NVIDIA Tesla es una herramienta para resolver problemas exigentes. Pensada como una solución dedicada al cálculo intensivo en la GPU (GPU Computing), el procesador Tesla puede aportar capacidad de cálculo elevada en una estación de trabajo, un servidor o en cluster de servidores compatibles.

a. SERIE TESLA 8:

Entre las características más importantes de los productos Tesla Serie 8 se incluyen:

- Arquitectura de procesamiento multihilo masivo, con un núcleo de computación de 128 núcleos en el procesador.
- Entorno de programación en C para GPU.
- Una gama de herramientas de desarrollo (compilador de C, depurador, analizador de rendimiento y bibliotecas optimizadas).
- Flexibilidad para adaptarse a los entornos de alta computación (HPC) existentes.

La GPU Tesla C 870 se encuentra en el centro de la serie 8 con soluciones que pretenden aprovechar sus capacidades más relevantes, y que según el fabricante, son:

Alta capacidad de procesamiento en paralelo: Un procesador multihilo de alta capacidad analiza grandes volúmenes de datos para extraer la información y encontrar las respuestas con rapidez. Entre las funciones más importantes de las GPU Tesla se incluyen el gestor de ejecución de subprocesos (Thread Execution Manager), que se encarga de coordinar la ejecución simultánea de miles de subprocesos (threads) en paralelo, y la caché de datos paralelos (Parallel Data Cache), encargada de que esos subprocesos puedan compartir los datos con facilidad.

Comunidad de desarrollo: La comunidad de desarrolladores online de NVIDIA proporciona acceso interactivo a foros, material de formación, y otros recursos y herramientas de utilidad. En realidad está herramienta está disponible para la Tecnología CUDA en general y no exclusivamente para desarrollo sobre Tesla.

Soluciones compatibles:

Como solución basada en estándares, Tesla se integra con facilidad en los entornos HPC existentes. Entre los productos disponibles se incluyen una GPU (C 870) para que los usuarios mejoren su actual estación de trabajo, un sistema de cálculo en la GPU para escritorio (D 870) que permite añadir capacidad suplementaria a las estaciones de trabajo y un servidor de GPU para rack de tamaño de 1U (S 870) que puede integrarse en los CPD corporativos.

Por un lado tenemos la propia **GPU Tesla 870** que es instalada en un ordenador directamente, pero como no posee dispositivos de entrada o salida y se comunica con el ordenador por la PCIx16. Por tanto se hace necesaria añadir una tarjeta gráfica adicional con salida vídeo.

La **C870** introduce una arquitectura de procesamiento multihilo masivo para aplicaciones de cálculo de alto rendimiento HPC (High Precision Computing) destinadas a usos científicos, técnicos e industriales.

El modelo computacional del Tesla C870 transformaría, por tanto, los sistemas estándar en supercomputadoras personales con picos de capacidad de más de 500 gigaflops. Está Dotado de un núcleo de cálculo de 128 procesadores y C870 puede utilizarse en combinación con sistemas de CPU multinúcleo para crear un superordenador personal.

En la Tabla 1 se resumen las características técnicas más relevantes del producto en base al catálogo del fabricante.

Producto	Tesla C870
Formato	ATX, 4.38” x 12.28”
N° de GPUs Tesla	1
Memoria dedicada	1,5 GB de GDDR3
Máximo de ops. en coma flotante/segundo	Más de 500 gigaflops
Precisión de las operaciones en coma flotante	Precisión simple según norma IEEE 754
Interfaz de memoria	384 bits
Ancho de banda de memoria	76,8 GB/s
Consumo máximo	170W
Interfaz del sistema	PCI Express x16
Conectores de alimentación auxiliares	Sí (2)
N° de ranuras	2
Solución de disipación térmica	Disipador/ventilador activo

Tabla 1: Características técnicas de Tesla C870.

Además, existe la solución Deskside o sistema de cálculo llamada **D 870** que es conectado mediante una tarjeta PCI x 16 a una estación de trabajo duplicando las capacidades de la GPU. Este dispositivo está formado por dos GPUs C 870 e incorpora dispositivos independientes de conexión y control de temperatura. Esta solución también permite combinar dos unidades en un rack montado obteniendo un sistema con cuádruple GPU. Esta supercomputadora se puede conectar a estaciones de trabajo mediante PCI Express para crear una solución HPC basada en dos GPU de 128 procesadores.

En la Tabla 2 se resumen las características técnicas más relevantes del producto en base al catálogo del fabricante.

Producto	Tesla D870
Formato	Sistema de escritorio
N° de GPUs Tesla	2
Memoria total dedicada	3 GB (1,5 GB de GDDR3 por GPU)
Máximo de ops. en coma flotante/segundo	Más de 500 gigaflops por GPU
Precisión de las operaciones en coma flotante	Precisión simple según norma IEEE 754
Adaptador del sistema	PCI Express x16 o x8, SFF, pasiva (10 W)
Potencia	520 W máx., 100-240 VCA, detección automática
Emisión acústica	40 dB
Peso	~8,5 Kg aprox.

Tabla 2: Características técnicas de Tesla D870.

Finalmente encontramos El sistema de cálculo bautizado como **S 870,** con cuatro GPUs y que se conecta al sistema Host con un cable a una tarjeta PCI Express x16. Esta solución permite el desarrollo de sistemas cluster.

Se trata de un servidor de pequeño formato (1U) que se puede integrar en los clusters de servidores de las empresas. A esto se añade su capacidad de expansión, que se puede incrementar el rendimiento para resolver los problemas de cálculo más complejos.

En la Tabla 3 se resumen las características técnicas más relevantes del producto en base al catálogo del fabricante.

Producto	Tesla S870
Formato	Carcasa de 1U para racks estándar de 19”
N° de GPUs	4
Memoria dedicada	6 GB (1,5 GB de GDDR3 por GPU)
Máximo de ops. en coma flotante/segundo	2,072 Teraflops
Precisión de las operaciones en coma flotante	Precisión simple según la norma IEEE 754
Potencia	550W. (800W máx), 100-240 VCA, detección automática
Adaptador del sistema	PCI Express x16 o x8, SFF, pasiva (10 W)

Tabla 3: Características técnicas de Tesla S870.

Como características comunes de la serie TESLA 8 se destacan:

Plataformas compatibles

- Sistema NVIDIA Tesla certificado
- Microsoft® Windows® XP (32-bit)
- Linux® (64 y 32 bits)
    - Red Hat Enterprise Linux 3, 4 y 5
    - SUSE 10.1, 10.2 y 10.3

Arquitectura Tesla de NVIDIA

- Arquitectura de cálculo paralelo masivo con 128 procesadores multihilo por GPU
- Procesador escalar con operaciones de enteros y en coma flotante
- Thread Execution Manager: gestor de ejecución que permite ejecutar miles de subprocesos (hilos) simultáneos por GPU
- Parallel Data Cache: caché de datos paralelos que permite a los procesadores compartir la información de la caché y, por tanto, acelerar el rendimiento
- Acceso ultrarrápido a la memoria con picos de ancho de banda de 76,8 GB/s por GPU
- Precisión simple de las operaciones en coma flotante según la norma IEEE 754

Soluciones ampliables

- Ampliable de una a millares de GPU
- Disponible como procesador para GPU Computing, supercomputadora de escritorio y servidor de GPU Computing de 1U para rack

Herramientas de desarrollo de software comunes a CUDA

- Compilador de C, herramienta de análisis y modo de emulación para depuración
- Bibliotecas numéricas estándar para FFT (Fast Fourier Transform) y BLAS (Basic Linear Algebra Subroutines)

b. SERIE TESLA 10

La serie Tesla 10 está basada en el procesador Tesla C1060, dotado de múltiples núcleos para cálculo masivo en paralelo, que se combina con el entorno de programación en C CUDA a fin de simplificar la programación en sistemas multinúcleo. Su arquitectura se representa en la figura 1.

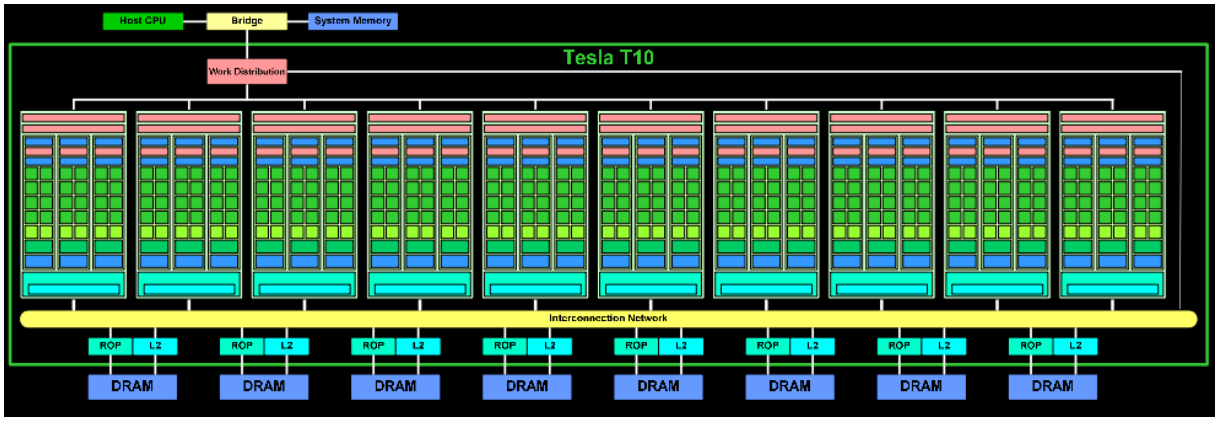

Fig. 1: Arquitectura Tesla 10

La Tesla™ C1060 es una tarjeta PCI Express 2.0. Según NVIDIA es capaz de realizar 933 GFLOPs/s y se ofrece con 4 GB de memoria GDDR3 con un ancho de banda de 102 GB/s. Sus características fundamentales se resumen en la tabla 4:

Producto	Tesla C1060
Form Factor	10,5" x 4,376", doble ranura
N° de GPUs Tesla	1
Memoria dedicada **Frecuencia de la memoria**	4 GB de GDDR3 800 MHz
N° de núcleos de procesamiento para streaming **Frecuencia de los núcleos de procesamiento**	240 1.3 Ghz.
Precisión de las operaciones en coma flotante	Precisión doble y simple según norma IEEE 754
Interfaz de memoria	512-bit GDDR3
Ancho de banda de memoria	102 GB/s
Consumo máximo	187.8 W
Interfaz del sistema	PCI Express x16 Generación 2
Conectores de alimentación auxiliares	Sí (2 de 6 patillas o 1 de 8 patillas)
Máximo de ops. en coma flotante/segundo	933 GFLOPs/s
Solución de disipación térmica	Disipador/ventilador activo

Tabla 4: Características técnicas de Tesla C1060.

Al igual que en el caso de la C 870, la C 1060 se encuentra en el centro de las soluciones de la serie 10.

Una de estas soluciones es el sistema computacional S 1070 que es un sistema montado rack 1U con 4 tarjetas Tesla C1060. El sistema puede conectarse a uno o dos Host con cables PCI Express. Una tarjeta interfaz para host (HIC) es usada para conectar cada cable a un host. Las tarjetas son compatibles tanto con sistemas PCI Express 1x como con PCI Express 2x.

c. SERIE TESLA 20 Y ARQUITECTURA FERMI

La serie NVIDIA® Tesla™ 20 está diseñada para sistemas de cálculo de alto rendimiento. Se basa en la arquitectura CUDA de última generación, "Fermi", e incorpora muchas funciones diseñadas para los sistemas de cálculo de entornos empresariales y técnicos. Esto incluye memoria ECC para proporcionar máxima precisión y capacidad de crecimiento, soporte de C++ y 8 veces más rendimiento en las operaciones de coma flotante de doble precisión que los productos de la serie Tesla 10.

NVIDIA® Tesla™ C2050 y Tesla C2070 son tarjetas tipo PCI Express 2.0 basadas en la GPU Tesla T20. Han sido diseñadas para sistemas de computación PCI Express. Ambas son capaces de realizar hasta 600 GFLOPs/sec con doble precisión y tiene un ancho de banda de 170 GB/s. Tesla C2050 se comercializa con 3 GB de memoria GDDR5, mientras que Tesla C2070 lo está con 6 GB de memoria GDDR5.

Ambas tarjetas pueden ser configuradas para activar o desactivar el código de control de error (ECC) que permite corregir errores de un bit y detectar errores de dos bits. Su activación conlleva un gasto de memoria que la disminuiría a 2.625 GB en la primera y 5.25 GB en la segunda.

Estas tarjetas están basadas en la arquitectura Fermi que pretende convertir el coprocesamiento distribuido entre la GPU y la CPU en algo generalizado. Esta arquitectura está diseñada para C++ y disponible con un entorno de desarrollo Visual Studio, y pretende facilitar la programación paralela y acelerar el rendimiento de aplicaciones y procesos, lo que incluye numerosos algoritmos de búsqueda y operaciones de trazado de rayos, física, análisis de elementos finitos, cálculo científico de alta precisión, álgebra lineal dispersa y ordenación.

Fermi incorpora algunas innovaciones importantes:

- 512 núcleos CUDA.
- Tecnología Parallel DataCache de NVIDIA.
- Motor GigaThread 3.0 de NVIDIA.
- Soporte completo de ECC.

4. ANEXOS:

ANEXO I:Comparativa de precios de Tarjetas compatibles CUDA

Modelo	Precio		Modelo	Precio		Modelo	Precio
Tesla C870	692.91 €		GeForce GT 220	61.05€		Quadro 4600	1.418,74 €
Tesla C870	1264.50 €		GeForce GTS 250	120,00 €		Quadro 4800	2.193,13 €
Tesla™ C1060 Compute Board TCSC1060-PB	938,70 €		GeForce GT 240	87.70 €		Quadro 5600	2.579,74 €
Tesla™ C2050 TCSC2050-PB	1.805,58 €		GeForce GTX 260	179,00 €		Quadro 5800	$2,950.00
Tesla™C2070 TCSC2070-PB	2.886,66 €		**GeForce GTX 285**	364.32 €		Quadro® PLEX 2200 D2 PCI-E x16 8GB	7.865,00 €
Tesla™ S1070-500 PCIE-X16-2M TCSS10702MX165B	6.494,04 €		GeForce 9800 GTX	178.66 €		Quadro® PLEX 2200 D2 PCI-E x8 8GB	7.865,00 €
Tesla™ S1070-500 Standard Configuration	6.494,04 €		**GeForce 9600GT**	69,27 €			
			GeForce 9800GT	89,01 €			
			GeForce 9500 GT	59.48 €			
			GeForce 295 GTX	$559,99			
			GeForce 280 GTX	$349,99			

Los precios corresponden al periodo Marzo-Abril 2010.

ANEXO II: Comparativa de Características Técnicas de las tarjetas compatibles con CUDA de la serie GeForce 200.

	GeForce GTS 250	GeForce GT 240	GeForce GTX 260	GeForce GTX 280	GeForce GTX 295	GeForce GTX 285	GeForce GTX 220
Especificaciones de motor de GPU:							
Núcleos de procesamiento	128	96	192	240	480 (240x2)	240	48
Reloj de gráficos (MHz)	738	550	1242 MHz	602 MHz	576	648	625
Reloj de procesador (MHz)	1836	1340	576 MHz	1296 MHz	1242	1476	1360
Tasa de relleno de texturas (10^9/s)	47,2		36,9	48,2	92,2	51,8	
Especificaciones de memoria:							
Reloj de la memoria (MHz)	1100	1700 MHz GDDR5, 1000MHz GDDR3, 900MHz DDR3	999 MHz	1107 MHz	999	1242	790MHz
Config. de memoria estándar	512 MB/1GB	512MB/1GB	896MB	1G	1792MB (896MBx2) GDDR3	1024 MB GDDR3	1GB
Interfaz de memoria	256-bit	128-bit	448-bit	512-bit	896-bit (448-bitx2)	512-bit	128-bit DDR3
Ancho de banda de memoria (GB/s.)	70,4	57,6	111,9	141,7	223,8	159	25,3
Especificaciones térmicas y de alimentación:							
Temperatura máxima de la GPU (en C)	105	105			105 C	105 C	
Potencia máxima en la tarjeta gráfica (W)	150	69			289 W	204 W	
Requisitos mínimos de alimentación del sistema (W)	450	300	500		680 W	550 W	

ANEXO III: Características Técnicas de las tarjetas de la serie Tesla y Quadro FX

GPU	C 870	C 1060	C 2050	C 2070
Número de Procesadores	128	240	448	448
Reloj del núcleo	1.35 GHz	1.296 GHZ	1.25 GHz to 1.40 GHz	1.25 GHz to 1.40 GHz
Tamaño	42.5 mm × 42.5 mm 1449-pin flip-chip ball grid array (FCBGA)	45.0 mm × 45.0 mm 2236-pin flip-chip ball grid array	42.5 mm × 42.5 mm 1981-pin ball grid array (BGA)	42.5 mm × 42.5 mm 1981-pin ball grid array (BGA)
Memoria	1536 MB	4 GB	3 GB	6 GB
	24 pieces 16M × 32 GDDR3 136-pin BGA SDRAM	32 pieces 32M × 32 GDDR3 136-pin BGA, SDRAM	24 pieces 32M × 32 GDDR5 136-pin BGA, SDRAM	24 pieces 64M × 32 GDDR5 136-pin BGA, SDRAM
Reloj de Memoria	800 MHz	800 MHz	1.8 GHz to 2.0 GHz	1.8 GHz to 2.0 GHz
BIOS	128 K × 8 Serial ROM,	1 Mbit Serial ROM	2Mbit Serial ROM	2Mbit Serial ROM
Memoria I/O	384-bit GDDR3		384-bit GDDR5	384-bit GDDR5
Conectores externos	None	None	Single port, dual-link DVI-I	Single port, dual-link DVI-I
Conectores a la placa	PCI Express Gen 1 x16	PCIe x16	PCI Express Gen2 ×16 system interface	PCI Express Gen2 ×16 system interface
Conectores internos	Two 6-pin PCI Express	8-pin PCI Express power connector	8-pin PCI Express power connector	8-pin PCI Express power connector
	4-pin fan connector	6-pin PCI Express power connector	6-pin PCI Express power connector	6-pin PCI Express power connector
	Two SLI 26-pin edge connectors are present	4-pin fan connector	4-pin fan connector	4-pin fan connector
Consumo total	170.9 W	187.8 W	< = 225 W	< = 225 W
Pico Operaciones en float	500 GFLOPs/s	933 GFLOPs/s		
Pico Operaciones en double	N/A	78 GFLOPs/s	600 GFLOPs/s	600 GFLOPs/s
Ancho de banda de memoria	76,8 GB/s	102 GB/sec	170 GB/sec	170 GB/sec

	FX 5600	FX 4600	FX 4800	FX 5800
Núcleos CUDA	128	112	192	240
Memoria	1.5 GB GDDR3	768MB GDDR3	1.5 GB GDDR3	4GB
Interfaz de Memoria	384 bit	384 bit	384 bit	512 bit
Ancho de banda de Memoria	76.8 GBps	67.2 GBps	76.8 GBps	102 GBps
Consumo	171 W	134 W	150 W	189 W