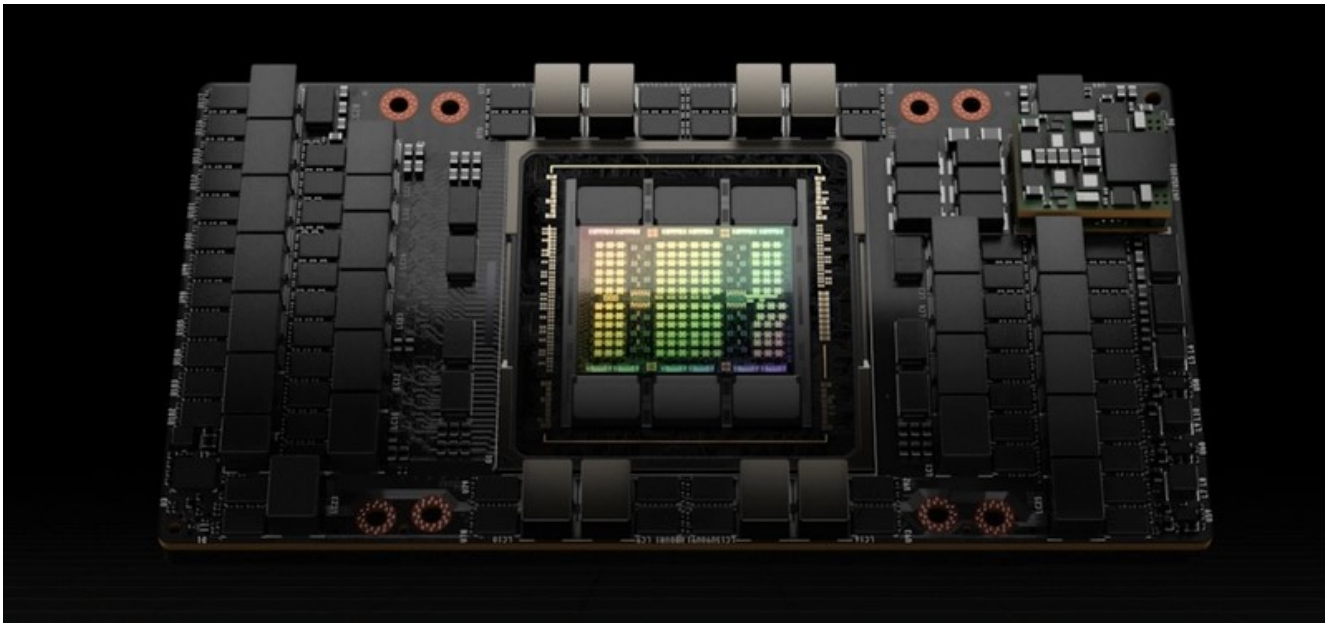
NVIDIA anuncia la arquitectura Hopper con el GH100 de 814 mm^2 a 4 nm, y la H100 de 16 896 CUDA y 80 GB HBM3

NVIDIA ha anunciado finalmente su nueva arquitectura de cómputo, **Hopper**, y el primer chip que la usa, el **GH100**, aunque se usará en otros chips que también ha comentado en la presentación. Ese chip GH100 se usará principalmente en la **aceleradora H100** en formato SXM5 y PCIe 5 las cuales siguen la senda marcada por la A100 y que tan buenos resultados ha dado en cómputo a las empresas que han tenido el dinero para comprarla.

El modelo como tarjeta SXM5 es el más potente y consume 700 vatios. El modelo como tarjeta PCIe consume 350 W y en realidad solo pierde un 20 % de rendimiento sobre el papel, aunque también ve reducida su ancho de banda de memoria de los 3 TB/s a los 2 TB/s. Esa memoria son 80 GB de HBM3 a 4.8 Gb/s en el modelo superior con un bus de datos de 5120 bits.



La potencia de cómputo del H100 SXM es de **60 TFLOPS en FP32** y el modelo H100 PCIe tiene 48 TFLOPS. El plato fuerte viene al calcular en otros tipos de formatos: 500/1000 TFLOPS (cálculo normal y disperso) en TF32, 1000/2000 TFLOPS en FP16 de tensores, 2000/4000 TOPS en INT8 o 1000/2000 TFLOPS en BFLOAT16.

Todos son formatos muy usados en inteligencia artificial, y por tanto el rendimiento es excelente. Eso sí, los datos de potencia son *estimaciones* de NVIDIA que deja claro que pueden cambiar en el producto final.

El chip desbloqueado **GH100, fabricado a 4 nm por TSMC y con un tamaño de 814 mm²,** es de hasta 144 SM para un total de 18 432 núcleos CUDA, y dispone de hasta 576 núcleos tensoriales con 60 MB de caché de nivel 2 y seis chips de HBM3 con doce controladores de 512 bits lo que permite hasta 96 GB de VRAM. **La implementación usada en la H100 SXM5** es de 16 896 CUDA a 1775 MHz con 528 tensoriales con 80 GB de HBM3 usando diez controladores de 512 bits y tiene 50 MB de caché N2. **La implementación del H100 PCIe 5** es de 14 592 CUDA con 456 núcleos tensoriales y misma configuración de caché y VRAM.

En cuanto a esos núcleos tensoriales, ahora tienen motores de transformación que se encargarán de procesar modelos de

aprendizaje automática. Al ser una unidad especializada es mucho más potente y eficiente que el núcleo tensorial general. Esto resuelve el hecho de que los transformadores son los modelos más usados en IA en los últimos tiempos ya que son de inestimable ayuda en el procesamiento del lenguaje natural y la visión computerizada.

Otro truco que se ha sacado NVIDIA de la manga son las **instrucciones DPX orientadas a la programación dinámica**. Este tipo de programa está orientado a dividir la solución de un problema en subproblemas que puedan resolverse de manera óptima, subdividiéndose a su vez en subproblemas, y que el conjunto de soluciones permita llegar a la solución al problema principal. En esencia es trocear los problemas y aprovechar el fuerte paralelismo de las GPU para resolver problemas muy complejos a través de la recursividad.

Las reglas de división de los problemas están claras en la programación dinámica, y estas instrucciones DPX vienen a quitarle trabajo a las CPU y FPGA en que se suelen hacer este tipo de cálculos haciéndo de Hopper una arquitectura más apetecible para las empresas. Los casos de uso que pone NVIDIA son en campos tan diversos como análisis del genoma hasta la optimización de enrutamiento en redes, pasando por el análisis de grafos o procesamiento de datos de distintos tipos.

Hay muchos otros cambios a nivel interno para proporcionar mayor seguridad a la transformación de datos y reforzar la ***computación confidencial***. Esto se aplica al entorno de la virtualización donde múltiples usuarios van a compartir los recursos del mismo chip GH100, y por eso se aporta mayor protección de datos y aislamiento para evitar que unos usuarios accedan a los datos de otros, o incluso protección frente a manipulación física de las tarjetas que puedan afectar a los datos o dañar la propia GPU.

No faltan tampoco mejoras a la parte multimedia. Se aumentan a 340 los flujos de datos procesando HEVC, a 170 los de H.264 y

a 260 los de VP9. Las características decodificación son: H.264 a 8 bits y subcrominancia 4:2:0, HEVC hasta a 12 bits y 4:4:4, y VP9 hasta a 12 bits y 4:2:0.

Las conexiones de las tarjetas H100 puede ser una SXM con NVLink 4, que aumenta la velocidad a 900 GB/s frente a los 600 GB/s de la NVLink 3. La conexión PCIe será tipo 5.0, por lo que en estos casos el rendimiento de la tarjeta puede verse perjudicado, sobre todo si tiene que trabajar junto a más tarjetas. Los sistemas podrán tener hasta ocho SXM, las PCIe serán más útiles en servidores menos punteros –y más baratos o tradicionales–.

Las primeras aceleradoras H100 se empezarán a enviar a los que las reserven a partir del tercer trimestre del año.

Fuente: geektopia.