

PUBLIC EXHIBIT 50

US Patent No. 11,841,803

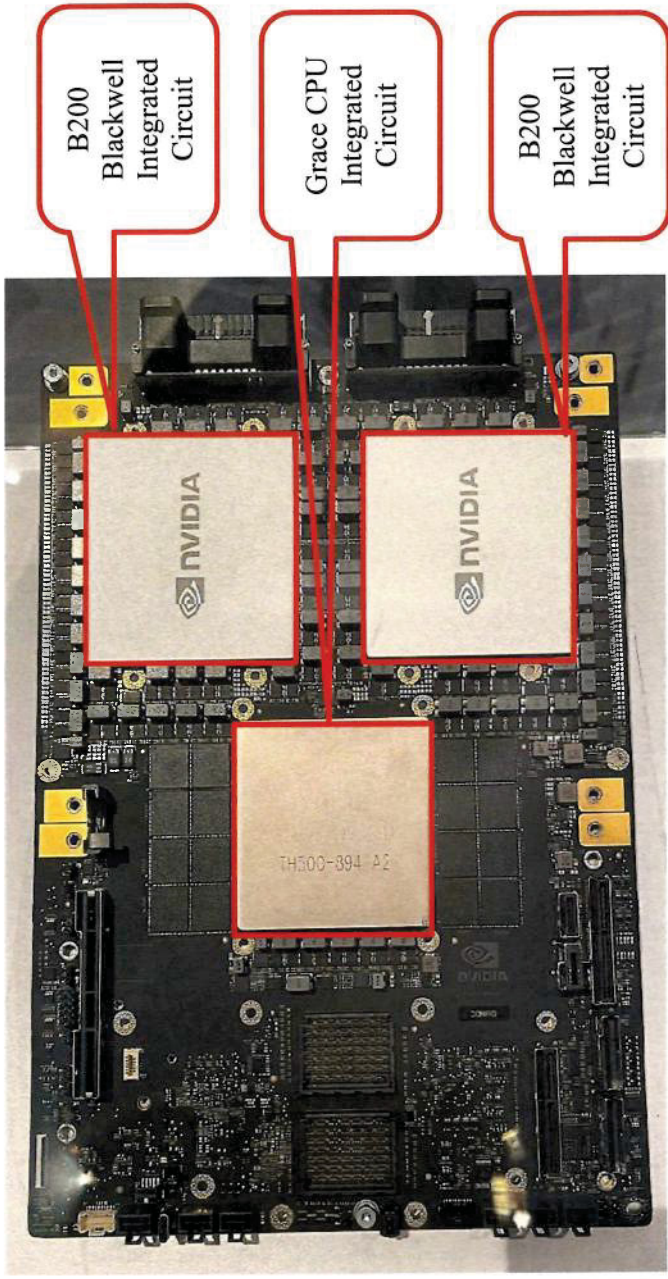
NVIDIA Representative Article B
(NVIDIA Blackwell GB200 Superchip Circuit Board Assembly)

Claim 1(pre): "A system, comprising:"

1(pre). A system, comprising:

The NVIDIA Grace Blackwell GB200 Superchip (the NVIDIA Representative Article B) (hereinafter, the "NVIDIA Product") is a circuit board assembly that includes a system.

For example, the NVIDIA Product is a circuit board assembly, which includes "1 Grace CPU" integrated circuit and "2 Blackwell GPUs", which are integrated circuits.



See GB200 Photograph, CES - NVIDIA GB200 1.jpg.

Claim 1(pre): “A system, comprising:”

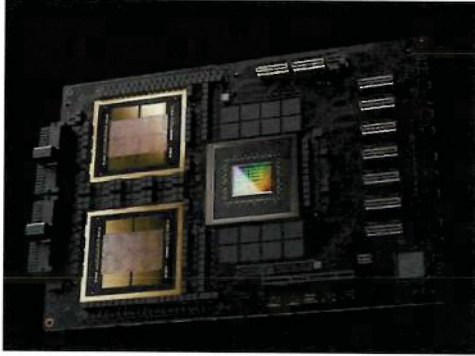
NVIDIA GB200 Superchip and GB200 NVL72

The NVIDIA GB200 Grace Blackwell Superchip connects two high-performance NVIDIA Blackwell Tensor Core GPUs and an NVIDIA Grace CPU using the NVIDIA® NVLink®-C2C interconnect that delivers 900 gigabytes per second (GB/s) of bidirectional bandwidth to the two GPUs.

Table 1. NVIDIA Blackwell GB200 Specifications

GPU Spec	GB200 Superchip
Configuration	1 Grace CPU : 2 Blackwell GPUs

See NVIDIA, NVIDIA Blackwell Architecture Technical Brief, <https://resources.nvidia.com/en-us-blackwell-architecture>, p. 11.



See NVIDIA, NVIDIA Blackwell Platform Arrives to Power a New Era of Computing, <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>.

“The [Blackwell] architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU.”

Claim 1(pre): "A system, comprising:"

This architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU. Each of the two dies are the largest die possible within the limits of reticle size, as big as can possibly be built today. The two dies are connected and unified with a single 10 terabyte-per-second (TB/s) chip-to-chip NVIDIA High-Bandwidth Interface (NV-HBI), providing one fully coherent, unified GPU.

See NVIDIA, NVIDIA Blackwell Platform Arrives to Power a New Era of Computing, <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>; see also NVIDIA, NVIDIA Blackwell Architecture Technical Brief, <https://resources.nvidia.com/en-us-blackwell-architecture>, p. 7.

The first thing to note is that the Blackwell GPU is going to be big. Literally. The B200 modules that it will go into will feature two GPU dies on a single package. That's right, NVIDIA has finally gone chiplet with their flagship accelerator. While they are not disclosing the size of the individual dies, we're told that they are "reticle-sized" dies, which should put them somewhere over 800mm² each. The GH100 die itself was already approaching TSMC's 4nm reticle limits, so there's very little room for NVIDIA to grow here – at least without staying within a single die.

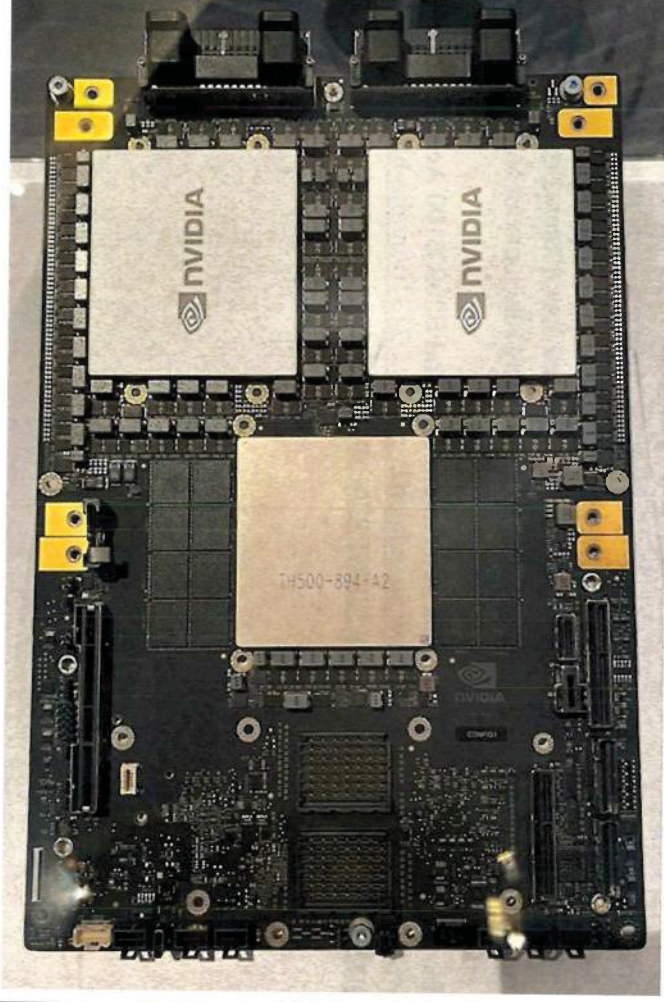
See AnandTech, NVIDIA Blackwell Architecture and B200/B100 Accelerators Announced: Going Bigger with Smaller Data, <https://www.anandtech.com/show/21310/nvidia-blackwell-architecture-and-b200b100-accelerators-announced-going-bigger-with-smaller-data>.

Claim 1(a): “a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and”

1(a). a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and

The NVIDIA Product includes a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus.

For example, the NVIDIA Product “connects two NVIDIA B200 Tensor Core GPUs to the NVIDIA Grace CPU over a 900GB/s ultra-low-power NVLink chip-to-chip interconnect.”



See GB200 Photograph, CES - NVIDIA GB200 1.jpg.

A Massive Superchip

The NVIDIA GB200 Grace Blackwell Superchip connects two NVIDIA B200 Tensor Core GPUs to the NVIDIA Grace CPU over a 900GB/s ultra-low-power NVLink chip-to-chip interconnect.

See NVIDIA, NVIDIA Blackwell Platform Arrives to Power a New Era of Computing, <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>.

Claim 1(a): “a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and”



The NVIDIA Product includes “NVLink v5” and “PCIe 6.0” interconnects.

Table 1. NVIDIA Blackwell GB200 Specifications

GPU Spec		GB200 Superchip
Configuration		1 Grace CPU : 2 Blackwell GPUs
NVLink Support		NVLink v5
NVLink Bandwidth Bidirectional		2x 1.8 TB/s
PCIe Gen 6 Bandwidth Bidirectional		2x 256 GB/s Gen6

See NVIDIA, NVIDIA Blackwell Architecture Technical Brief, <https://resources.nvidia.com/en-us-blackwell-architecture>, pp. 11-12.

Claim 1(a): “a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and”

GB200	
Type	Grace Blackwell Superchip
Interconnects	2x NVLink 5 (1800GB/sec) 2x PCIe 6.0 (256GB/sec)
GPU	2x “Blackwell GPU”
GPU Transistor Count	416B (2x2x104B)
TDP	2700W
Manufacturing Process	TSMC 4NP
Interface	Superchip
Architecture	Grace + Blackwell

See AnandTech, NVIDIA Blackwell Architecture and B200/B100 Accelerators Announced: Going Bigger with Smaller Data, <https://www.anandtech.com/show/21310/nvidia-blackwell-architecture-and-b200b100-accelerators-announced-going-bigger-with-smaller-data>.

NVIDIA’s NVLink-C2C “extends the industry-leading NVLink technology to a chip-to-chip interconnect. This enables the creation of a new class of integrated products with NVIDIA partners, built via chiplets, allowing NVIDIA GPUs, DPUs, and CPUs to be coherently interconnected with custom silicon.”

NVIDIA® NVLink®-C2C extends the industry-leading NVLink technology to a chip-to-chip interconnect. This enables the creation of a new class of integrated products with NVIDIA partners, built via chiplets, allowing NVIDIA GPUs, DPUs, and CPUs to be coherently interconnected with custom silicon. See <https://www.nvidia.com/en-us/data-center/nvlink-c2c/>.

Claim 1(a): “a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and”

NVIDIA GB200 NVL72

The NVIDIA GB200 Superchip combines two Blackwell GPUs and an NVIDIA Grace™ CPU connected over NVLink-C2C and can scale up to the GB200 NVL72, a massive 72-GPU system connected by NVIDIA® NVLink®.

Learn More >

See <https://www.nvidia.com/en-us/data-center/nvlink-c2c/>.

Further, for example, “[t]he [Blackwell] architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU.”

This architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU. Each of the two dies are the largest die possible within the limits of reticle size, as big as can possibly be built today. The two dies are connected and unified with a single 10 terabyte-per-second (TB/s) chip-to-chip NVIDIA High-Bandwidth Interface (NV-HBI), providing one fully coherent, unified GPU.

See NVIDIA, NVIDIA Blackwell Platform Arrives to Power a New Era of Computing, <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>; see also NVIDIA, NVIDIA Blackwell Architecture Technical Brief, <https://resources.nvidia.com/en-us-blackwell-architecture>, p. 7.

The “two silicon dies” in each Blackwell B200 GPU “are linked in the same package by a 10 terabyte per second connection.”

The B200 is composed of about 1600 square millimeters of processor on two silicon dies that are linked in the same package by a 10 terabyte per second connection, so they perform as if they were a single 208-billion-transistor chip.

See IEEE Spectrum, Nvidia Unveils Blackwell, Its Next GPU, <https://spectrum.ieee.org/nvidia-blackwell#:~:text=The%20B200%20is%20composed%20of,208%2Dbillion%2Dtransistor%20chip.>

Claim 1(a): “a central processing unit (CPU) communicably coupled to a first graphics processing unit (GPU) chiplet of a GPU chiplet array, wherein the GPU chiplet array includes: the first GPU chiplet communicably coupled to the CPU via a bus; and”

This architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU. Each of the two dies are the largest die possible within the limits of reticle size, as big as can possibly be built today. The two dies are connected and unified with a single 10 terabyte-per-second (TB/s) chip-to-chip NVIDIA High-Bandwidth Interface (NV-HBI), providing one fully coherent, unified GPU.

See NVIDIA, NVIDIA Blackwell Platform Arrives to Power a New Era of Computing, <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>; see also NVIDIA, NVIDIA Blackwell Architecture Technical Brief, <https://resources.nvidia.com/en-us-blackwell-architecture>, p. 7.

A New Class of AI Superchip

Blackwell-architecture GPUs pack 208 billion transistors and are manufactured using a custom-built TSMC 4NP process. All Blackwell products feature two reticle-limited dies connected by a 10 terabytes per second (TB/s) chip-to-chip interconnect in a unified single GPU.

See NVIDIA, NVIDIA Blackwell Architecture, <https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/>.

Claim 1(b): “a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink, wherein the passive crosslink is dedicated for inter-chiplet communications.”

1(b). a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink, wherein the passive crosslink is dedicated for inter-chiplet communications.

The NVIDIA Product includes a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink, wherein the passive crosslink is dedicated for inter-chiplet communications.

For example, “[t]he [Blackwell] architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU.”

This architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU. Each of the two dies are the largest die possible within the limits of reticle size, as big as can possibly be built today. The two dies are connected and unified with a single 10 terabyte-per-second (TB/s) chip-to-chip NVIDIA High-Bandwidth Interface (NV-HBI), providing one fully coherent, unified GPU.

See NVIDIA, NVIDIA Blackwell Platform Arrives to Power a New Era of Computing, <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>; see also NVIDIA, NVIDIA Blackwell Architecture Technical Brief, <https://resources.nvidia.com/en-us-blackwell-architecture>, p. 7.

The “two silicon dies” in each Blackwell B200 GPU “are linked in the same package by a 10 terabyte per second connection.”

The B200 is composed of about 1600 square millimeters of processor on two silicon dies that are linked in the same package by a 10 terabyte per second connection, so they perform as if they were a single 208-billion-transistor chip.

See IEEE Spectrum, Nvidia Unveils Blackwell, Its Next GPU, <https://spectrum.ieee.org/nvidia-blackwell#:~:text=The%20B200%20is%20composed%20of%20a%20208%2Dbillion%20transistor%20chip.>

This architecture is able to incorporate a significant amount of computing power by merging two dies into a single, unified GPU. Each of the two dies are the largest die possible within the limits of reticle size, as big as can possibly be built today. The two dies are connected and unified with a single 10 terabyte-per-second (TB/s) chip-to-chip NVIDIA High-Bandwidth Interface (NV-HBI), providing one fully coherent, unified GPU.

See NVIDIA, NVIDIA Blackwell Platform Arrives to Power a New Era of Computing, <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>; see also NVIDIA, NVIDIA Blackwell Architecture Technical Brief, <https://resources.nvidia.com/en-us-blackwell-architecture>, p. 7.

Claim 1(b): “a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink, wherein the passive crosslink is dedicated for inter-chiplet communications.”

A New Class of AI Superchip

Blackwell-architecture GPUs pack 208 billion transistors and are manufactured using a custom-built TSMC 4NP process. All Blackwell products feature two reticle-limited dies connected by a 10 terabytes per second (TB/s) chip-to-chip interconnect in a unified single GPU.

See NVIDIA, NVIDIA Blackwell Architecture, <https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/>.

Further, for example and upon information and belief, the NVIDIA Blackwell B200 GPU “use[s] TSMC’s CoWoS-L packaging, which connect chiplets using a redistribution layer (RDL) interposer with local silicon bridges.”

B200 GPUs are the first processors to use TSMC’s CoWoS-L packaging, which connects chiplets using a redistribution layer (RDL) interposer with local silicon bridges.

See 3DInCites, IFTLE 607: Why Nvidia’s Blackwell is Having Issues with TSMC CoWoS-L Technology, <https://www.3dincites.com/2024/10/iftle-607-why-nvidias-blackwell-is-having-issues-with-tsmc-cowos-l-technology>.

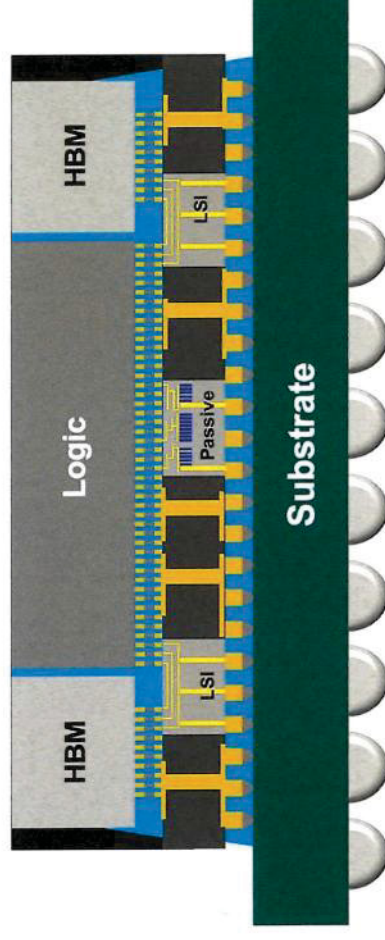
CoWoS-L “us[es] an interposer with a Local Silicon Interconnect (LSI) chip for die-to-die interconnect and RDL layers for power and signal delivery.”

CoWoS®-L is one of the chip-last packages on the CoWoS® (Chip on Wafer on Substrate) platform. It combines the merits of CoWoS®-S and InFO (Integrated Fan-Out) technologies to provide the most flexible integration using an interposer with a Local Silicon Interconnect (LSI) chip for die-to-die interconnect and RDL layers for power and signal delivery.

Claim 1(b): “a second GPU chiplet communicably coupled to the first GPU chiplet via a passive crosslink, wherein the passive crosslink is dedicated for inter-chiplet communications.”

Key CoWoS[®]-L features include:

1. LSI chips for a high routing density die-to-die interconnect through multiple layers of sub-micron copper lines. LSI chips can feature a variety of connection architectures, e.g., System on Chip (SoC)-to-SoC, SoC-to-chiplet, SoC-to-High Bandwidth Memory, within each product, and can be used repeatedly in multiple products. The corresponding metal types, layer counts, and pitches align with the offering from CoWoS[®]-S.



See TSMC, CoWoS, <https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm>.