



# **NVIDIA GB10 Superchip**

White Paper

# Document History

WP-12718-001\_v02

| Version | Date          | Authors | Description of Change                 |
|---------|---------------|---------|---------------------------------------|
| 01      | March 9, 2026 | JB, SM  | Initial release                       |
| 02      | April 2, 2026 | JB, SM  | Minor update to the “Arm CPU” section |

# Table of Contents

GB10 Maximum Efficiency and Minimum Footprint ..... 1

Two Tiles..... 2

    Arm CPU ..... 6

    Blackwell GPU..... 7

    Encoder and Decoder Engines..... 8

Unified Memory Architecture ..... 11

NVIDIA GB10 Software..... 13

## List of Figures

|                                                                              |    |
|------------------------------------------------------------------------------|----|
| Figure 1. GB10 SoC.....                                                      | 2  |
| Figure 2. GB10 SoC Package with GPU (Lower) and CPU (Top) Tiles Visible..... | 3  |
| Figure 3. Component Ownership Map of S-Die and G Die.....                    | 4  |
| Figure 4. SoC Topology on DGX Spark Application.....                         | 5  |
| Figure 5. 4x + 4x SoC to NIC Interface using ConnectX Socket Direct.....     | 6  |
| Figure 6. GB10 SoC Die Layout.....                                           | 12 |
| Figure 7. Software Stack for GB10 Based Systems with DGX OS.....             | 13 |

## List of Tables

|                                             |   |
|---------------------------------------------|---|
| Table 1. GB10 Superchip Specifications..... | 4 |
| Table 2. Uncore Features except GPU .....   | 5 |
| Table 3. GB10 CPU Specifications.....       | 7 |
| Table 4. GB10 GPU Specifications.....       | 8 |
| Table 5. NVENC Specifications .....         | 9 |
| Table 6. NVDEC Specifications .....         | 9 |

---

# GB10 Maximum Efficiency and Minimum Footprint

The GB10 chip brings the vision of generative AI driving a new industrial revolution into life. Built on NVIDIA's Blackwell GPU architecture - a culmination of three decades advancing accelerated computing - this CPU+GPU design is built from scratch for AI-native development, as the engine behind this transformation, tightly integrating general-purpose processing with massive parallel acceleration.

GB10 is a system-on-a-chip designed for AI first. CPU and GPU packaged together with NVIDIA Chip-to-Chip (C2C) interconnect unlocks full memory cohesion and much faster transfer rates than those of a standard PCI Express expansion slot on a typical personal computer.

This white paper explains the characteristics, benefits and the software available to help developers and system designers build the next generation of accelerated applications at an unprecedented power efficiency.

---

# Two Tiles

NVIDIA, in partnership with MediaTek, went to the drawing board to create a chip that would be natively optimized for accelerated work. The goal was to get the GPU as close as possible to the CPU from a logical perspective, but also physically. Leveraging the knowledge acquired from the 72-core NVIDIA Grace Hopper™ Superchip, NVIDIA developed a very fast and effective chip-to-chip interface that provides about five times the bandwidth of a comparable PCI Express link. However, more importantly than the speed alone, this interface allows full cache coherency between CPU and GPU, reducing the number of copies needed whenever these two components need to exchange data.

Figure 1. GB10 SoC

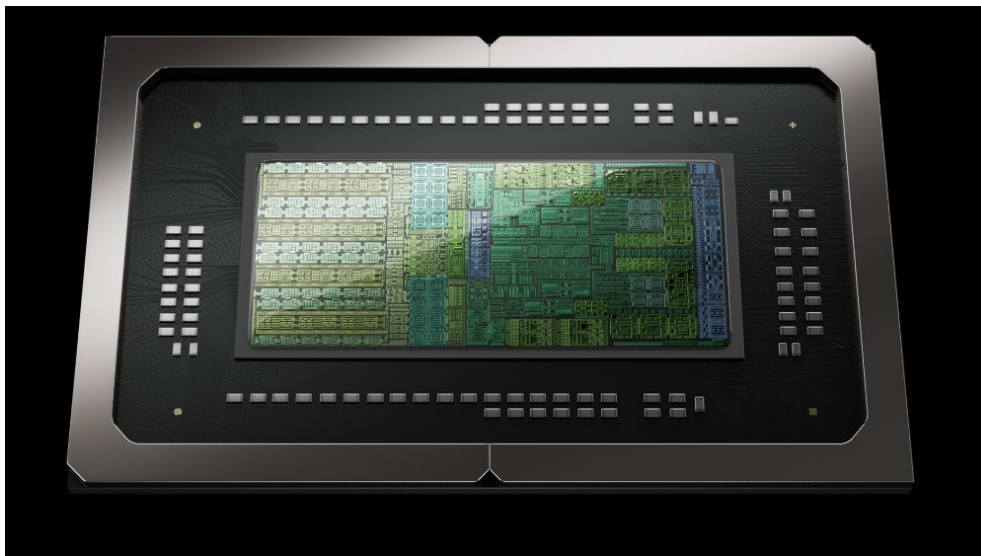
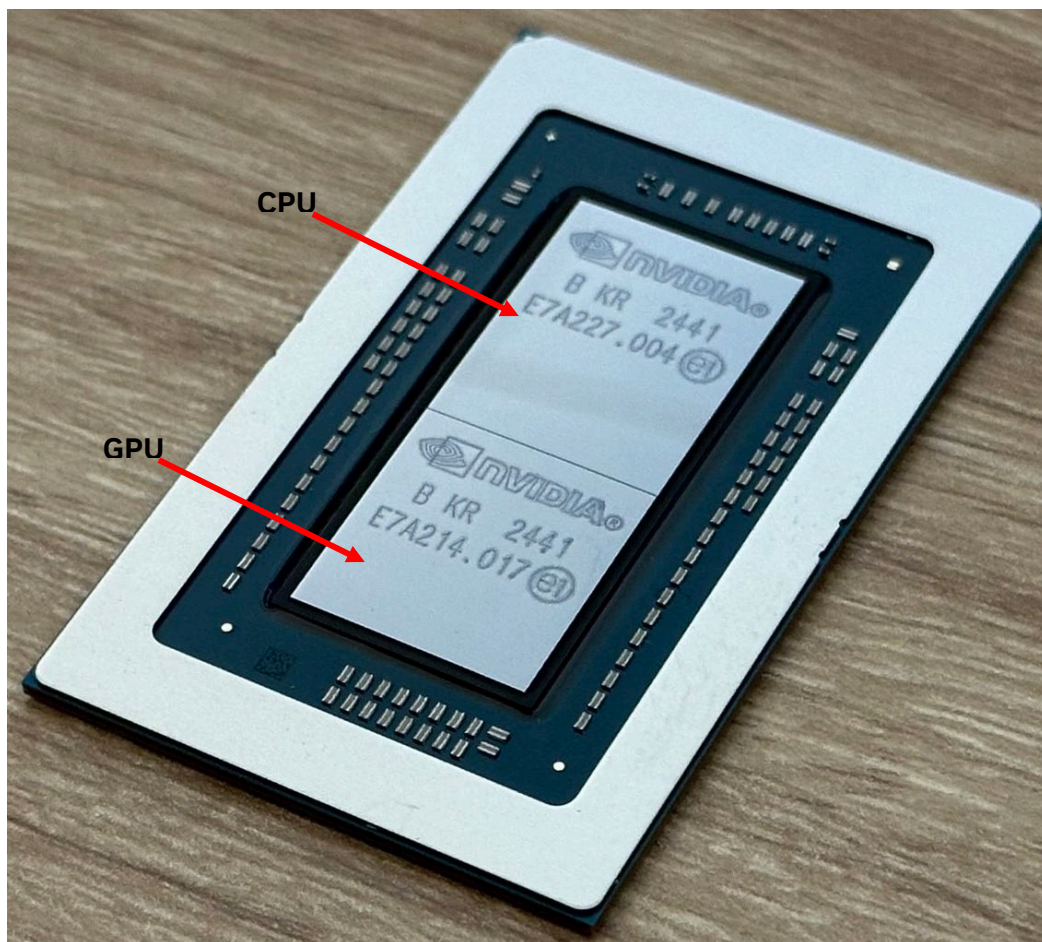


Figure 2. GB10 SoC Package with GPU (Lower) and CPU (Top) Tiles Visible



Using advanced 2.5D packaging, the S-die and G-die built on TSMC's 3 nm process are joined together using the NVIDIA Chip-to-Chip (C2C) interface, with Arm AMBA CHI-G coherency protocol support.

Figure 3. Component Ownership Map of S-Die and G Die

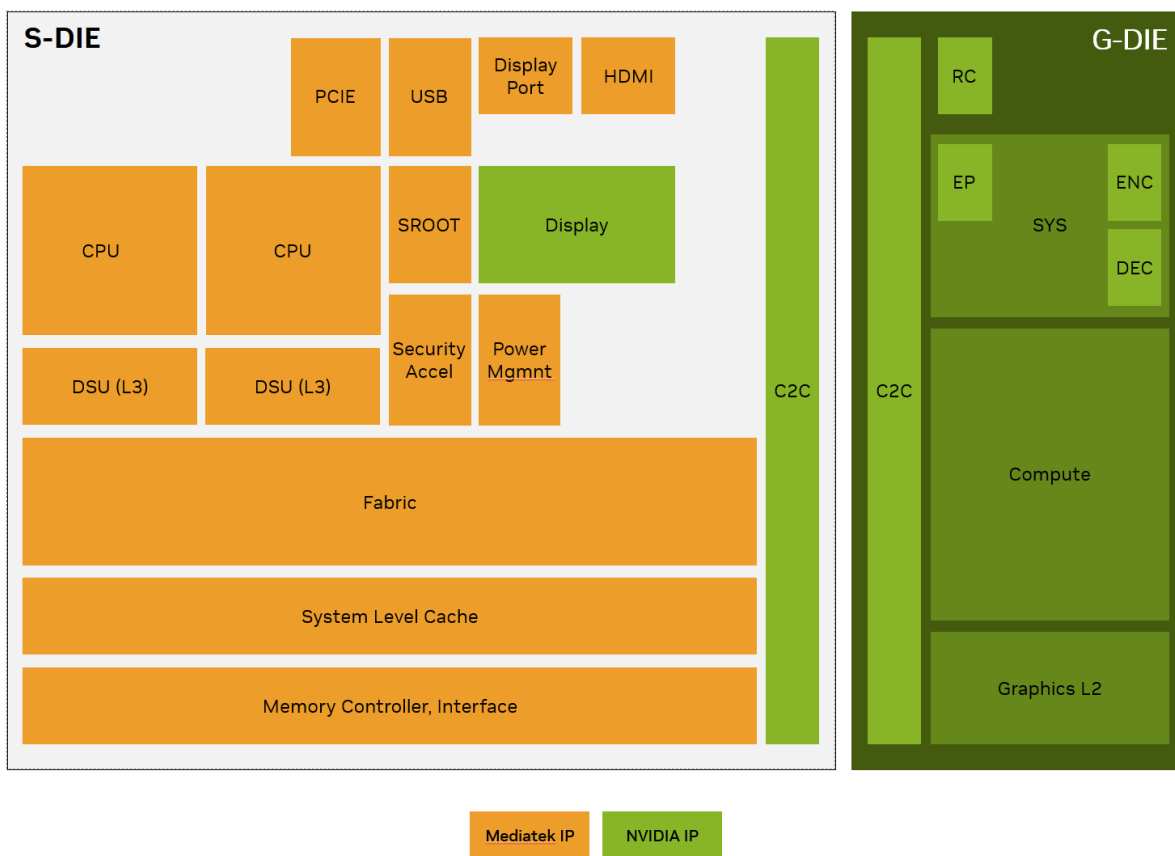


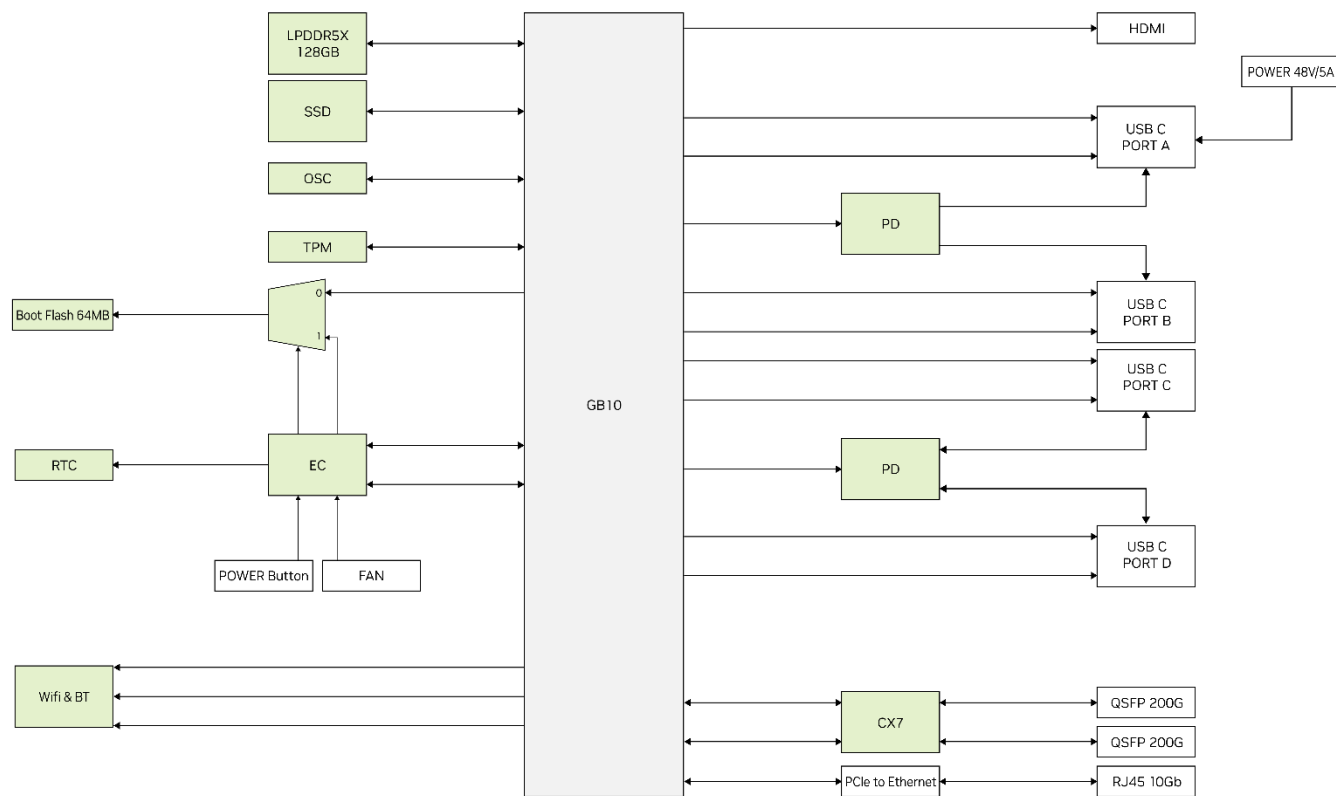
Table 1. GB10 Superchip Specifications

| Specification                                         | Description               |
|-------------------------------------------------------|---------------------------|
| Package type                                          | Flip-Chip Ball Grid Array |
| Pin count                                             | 2,970                     |
| Package dimensions                                    | 31 × 50 mm                |
| Power*                                                | 140 Watts                 |
| Memory type                                           | LPDDR5x (external to SoC) |
| Memory capacity                                       | 128 GB                    |
| Memory interface speed*                               | 8,533 Mbps                |
| <b>Note:</b>                                          |                           |
| *SoC only, typical. Can vary depending on application |                           |

**Table 2.      Uncore Features except GPU**

| Specification       | Description                        |
|---------------------|------------------------------------|
| USB version         | 3.2                                |
| USB ports           | 4                                  |
| Audio               | HiFi 1s and HiFi 5s Tensilica DSPs |
| PCI Express         | 16 Shared lanes of PCIe Gen4/5     |
| AI TOPS performance | 1 PFLOP NVFP4 Sparse               |

The following figure is the SoC logical diagram showing the SoC on the NVIDIA® DGX Spark™ application.

**Figure 4.      SoC Topology on DGX Spark Application**

# Arm CPU

NVIDIA and MediaTek leveraged industry-proven Cortex core designs, which are known for excellent single-thread performance and power efficiency. The CPU is designed for a balanced power and performance ratio, with 10 high-performance Cortex-X925 cores (p-cores) and 10 high-efficiency Cortex-A725 cores (e-cores). The p-cores can deliver impressive 4 GHz of peak frequency on demand for intensive single-thread operations, while the e-cores can keep the system running at very low power when idling.

The CPU is fully compliant with Arm v9.2 specification, which is used in many enterprise-class servers, and offers advanced features such as secure boot.

On the I/O side, this SoC offers 12 links of PCI Express Gen5, four links of PCI Express Gen4, and allows high-speed networking interface.

The SoC can provide up to 4-lane wide PCIe links, but with the NVIDIA® ConnectX®-7 (CX-7) Socket Direct technology. Two sets of 4-lane links can be combined for an 8-lane wide link between the SoC and the NIC. This enables networking interfaces as fast as 200 Gbps to be integrated into compact designs without PCIe switching. On DGX Spark, this interface can provide 200 Gbps of bandwidth even if only one port is physically linked.

Multi-socket mode creates a duplicate entry of the physical interface, so each port can be accessed through each half of the x8 link. Consequently, each half-link is expressed as an individual 100G interface within the operating system, resulting in 4x 100G links. These can be aggregated in the software layer for ease of use.

**Figure 5. 4x + 4x SoC to NIC Interface using ConnectX Socket Direct**

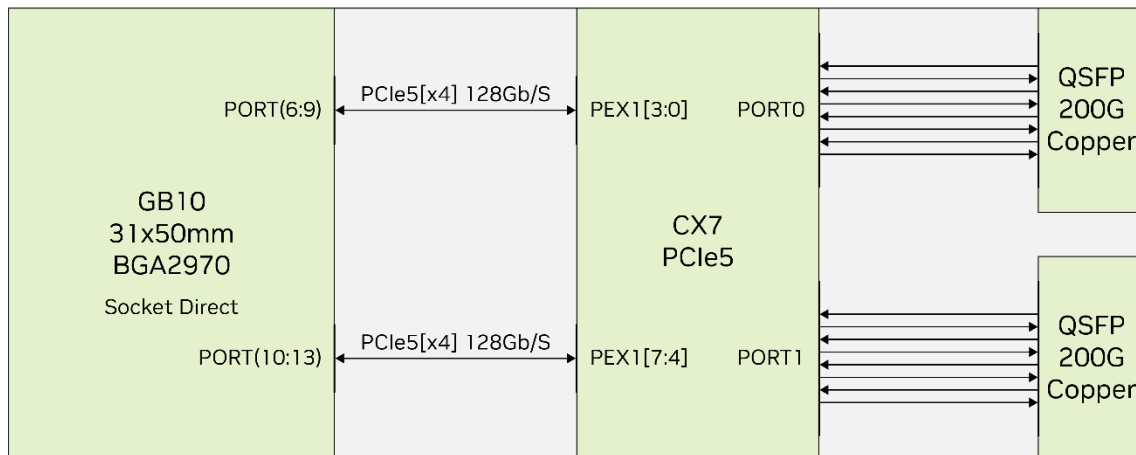


**Table 3. GB10 CPU Specifications**

| Specification                   | Description                 |
|---------------------------------|-----------------------------|
| Total core count                | 20                          |
| Performance core (p-core) count | 10                          |
| p-core maximum clock            | 4 GHz                       |
| Efficiency core (e-core) count  | 10                          |
| e-core maximum clock            | 2.8 GHz                     |
| L1 cache (p-core and e-core)    | 64 KiB per core             |
| L2 cache (p-core)               | 2 MiB per cluster           |
| L2 cache (e-core)               | 512 KiB per cluster MiB     |
| L3 cache (Cluster 0)            | 16 MiB                      |
| L3 cache (Cluster 1)            | 8 MiB                       |
| Instruction set                 | ARM V9.2                    |
| Virtualization support          | No                          |
| SVE and SVE2 support            | Yes                         |
| Secure Boot support             | Yes                         |
| TPM support                     | Yes, discrete or fTPM, v2.0 |

## Blackwell GPU

This class-defining chip would not be complete without a modern, powerful GPU. NVIDIA equipped this SoC with the state-of-the-art Blackwell architecture, the same used on flagship enterprise systems today. Developers can create applications using the same driver and libraries found on production systems, simplifying the workflow. This Blackwell GPU supports FP4 precision, allowing very high parameter count models to be executed locally. Now, when most other GPUs depend on their own dedicated memory to run accelerated applications, the GB10 Superchip shares its entire memory space, allowing the GPU to allocate more than 100 GiB, which is more memory than any consumer Blackwell GPU today. This vast memory opens a range of possibilities previously limited to data center equipment.

Along with its AI prowess, GB10 is still a fully featured Blackwell GPU capable of driving display output up to 8k in resolution, hardware encoding and decoding, and 3D rendering, enabling a wider range of applications beyond the traditional AI space.

**Table 4. GB10 GPU Specifications**

| Specification             | Description                                                       |
|---------------------------|-------------------------------------------------------------------|
| Architecture              | Blackwell                                                         |
| Streaming multiprocessors | 48                                                                |
| GPU boost clock           | 2.4 GHz                                                           |
| Global (L2) cache         | 24 MiB                                                            |
| Hardware encoder          | 1x NVENC                                                          |
| Encoder support           | h.265, h.264, AV1, YUV422, YUV44, VP9 10bit                       |
| Hardware decoder          | 1x NVDEC                                                          |
| Decoder support           | h.265, h.264, AV1, VP8, VP9, YUV422, YUV44, VC1(AP), MPEG2, MPEG4 |
| NVIDIA® CUDA® cores       | 6,144 (128 per SM)                                                |
| Tensor cores              | 48 (1 per SM)                                                     |
| Tensor core generation    | Fifth                                                             |
| Supported precisions      | CUDA: FP16, FP32<br>Tensor: FP4, FP8, FP16, BF16                  |
| Ray-tracing cores         | 48 (Two per TPC)                                                  |
| Ray-tracing generation    | Fourth                                                            |
| GPC                       | 4                                                                 |
| VESA® DisplayPort™ output | v.2.01.4, over USB only, up to four displays                      |
| HDMI™                     | v.2.1a, single Type-A port                                        |
| Maximum resolution        | Up-to FLR/12G + TDMS, 8K @ 120 fps SDR/HDR                        |
| Variable refresh rate     | Supported                                                         |
| DLSS 4                    | Supported                                                         |

## Encoder and Decoder Engines

Every GB10 is equipped with state-of-the-art Blackwell generation Standard Video Encoder (NVENC) and Multi-Standard Video Decoder (NVDEC). Common tasks, such as video playback and desktop capture recording can be greatly improved by offloading codec tasks to the discrete hardware engines. This subsystem supports timestamp for:

- > Audio and video synchronization
- > CBR and VBR rate controls
- > Programmable intra-refresh for error resiliency
- > Macro-block based and bit-based packetization
- > A motion estimation-only mode

Table 5. NVENC Specifications

| Stream Standard | Profiles                                      | Resolution and FPS (Max Number of Streams)                       |
|-----------------|-----------------------------------------------|------------------------------------------------------------------|
| H.264           | Baseline, Main, and High                      | 4k60 (1)<br>4k30 (3)<br>1080p60 (6)<br>1080p30 (13)              |
|                 | High 444, High 444 Predictive, MVC (per view) | 4k30 (1)<br>1080p60 (3)<br>1080p30 (6)                           |
| H.265 (HEVC)    | Main, Main 10                                 | 8k30 (1)<br>4k60 (3)<br>4k30 (7)<br>1080p60 (11)<br>1080p30 (23) |

Standards supported for NVDEC are H.265 (HEVC), H.264, VP9, VP8, AV1, MPEG-4, MPEG-2 and VC-1.

Table 6. NVDEC Specifications

| Stream Standard | Profiles | Resolution and FPS (Max Number of Streams)          |
|-----------------|----------|-----------------------------------------------------|
| H.264           | UHP      | 4k60 (1)<br>4k30 (3)<br>1080p60 (7)<br>1080p30 (14) |
|                 | HP       | 4k60 (1)<br>4k30 (2)<br>1080p60 (4)<br>1080p30 (8)  |
|                 | HQ       | 4k30 (1)<br>1080p60 (2)<br>1080p30 (4)              |
| H.265 (HEVC)    | UHP      | 4k60 (1)<br>4k30 (3)<br>1080p60 (7)<br>1080p30 (15) |
|                 | HP       | 4k30 (1)<br>1080p60 (3)<br>1080p30 (7)              |

| Stream Standard | Profiles | Resolution and FPS (Max Number of Streams)          |
|-----------------|----------|-----------------------------------------------------|
|                 | HQ       | 1080p60 (1)<br>1080p30 (30)                         |
| AV1             | UHP      | 4k60 (1)<br>4k30 (3)<br>1080p60 (7)<br>1080p30 (15) |
|                 | HP       | 4k30 (1)<br>1080p60 (3)<br>1080p30 (7)              |

---

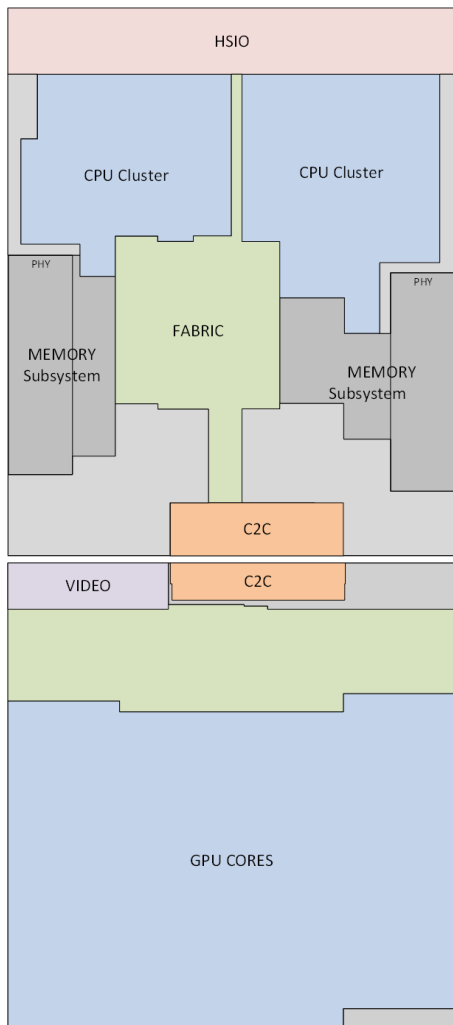
# Unified Memory Architecture

The NVIDIA Chip-to-Chip (C2C) interconnect is the underlying fabric that allows CPU and GPU to share memory in a very effective way. Traditionally, the CPU and GPU have their own memory spaces and any data that needs to flow between them must be copied before they can be accessed - this operation that can take multiple CPU cycles and slow down applications. With the increased pervasiveness of accelerated software, the narrow bridge between memory areas quickly becomes a performance bottleneck.

Another advantage of C2C is the unified physical memory pool. With this technology, a common memory bank, such as the 128 GB LPDDR5x memory of DGX Spark, can be shared between CPU and GPU needs. This flexibility allows large models that otherwise would require high-end GPUs with large, dedicated memory capacity to run, not requiring more memory to be added to the CPU to run the base operating system and other parallel tasks.

The combination of direct cross-access between CPU and GPU along with increased bandwidth compared to a traditional PCI Express link boosts the performance of this SoC to unprecedented levels at its power envelope, enabling new possibilities of high-performance accelerated applications using a very small form factor and low power requirements.

**Figure 6. GB10 SoC Die Layout**



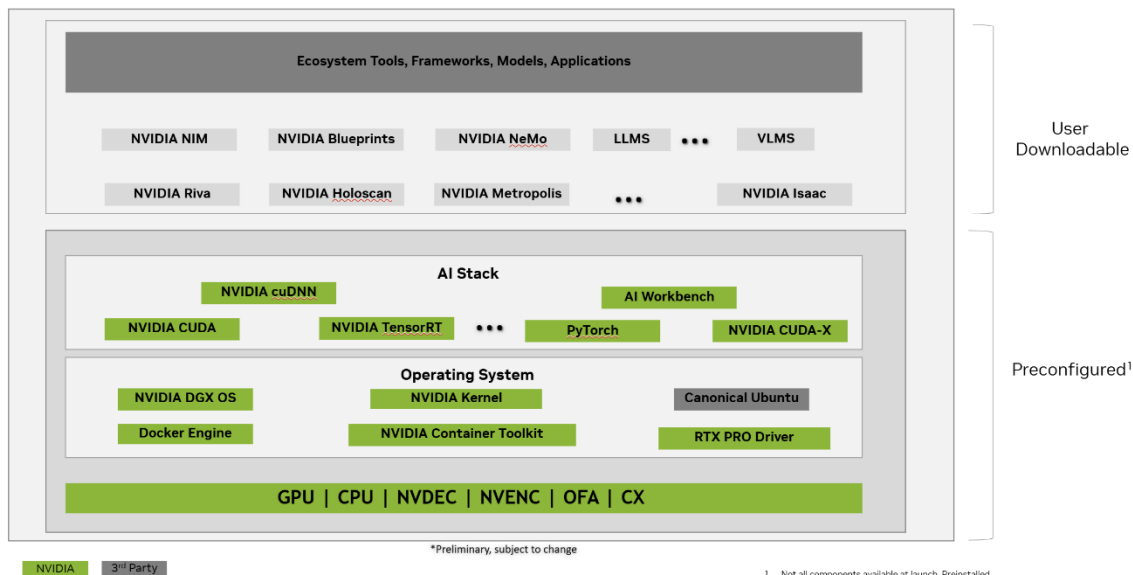
# NVIDIA GB10 Software

Engineers, researchers, and related professionals can effectively transition development work to production assets. With GB10, advanced AI workloads, such as fine tuning LLMs, can be done anywhere, whether at one's cube, dorm, or home office. NVIDIA offers a series of blueprints based on GB10 hardware, which are intended to demonstrate libraries and tools used in real-world AI applications. NVIDIA AI Enterprise supports the platform as well, providing developers with the stability, security and performance of enterprise-class software.

The GB10 represents NVIDIA's commitment to developers, bringing the latest data center technologies right to the hands of practitioners. Sharing software and hardware with larger data center systems, the hardware gap between development platforms and production becomes smaller. GB10 is a class-defining chip design built for the new era of AI-native software development.

The GB10 powers the NVIDIA DGX Spark and comparable systems from OEM partners. To get started with GB10, visit [www.nvidia.com/dgx-spark](http://www.nvidia.com/dgx-spark).

Figure 7. Software Stack for GB10 Based Systems with DGX OS



## Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

## Trademarks

NVIDIA, the NVIDIA logo, ConnectX, CUDA, DGX Spark, and NVIDIA Grace Hopper are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

## VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

## HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

## Arm

Arm, AMBA, and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS, and Arm Sweden AB.

## Copyright

© 2026 NVIDIA Corporation. All rights reserved.

