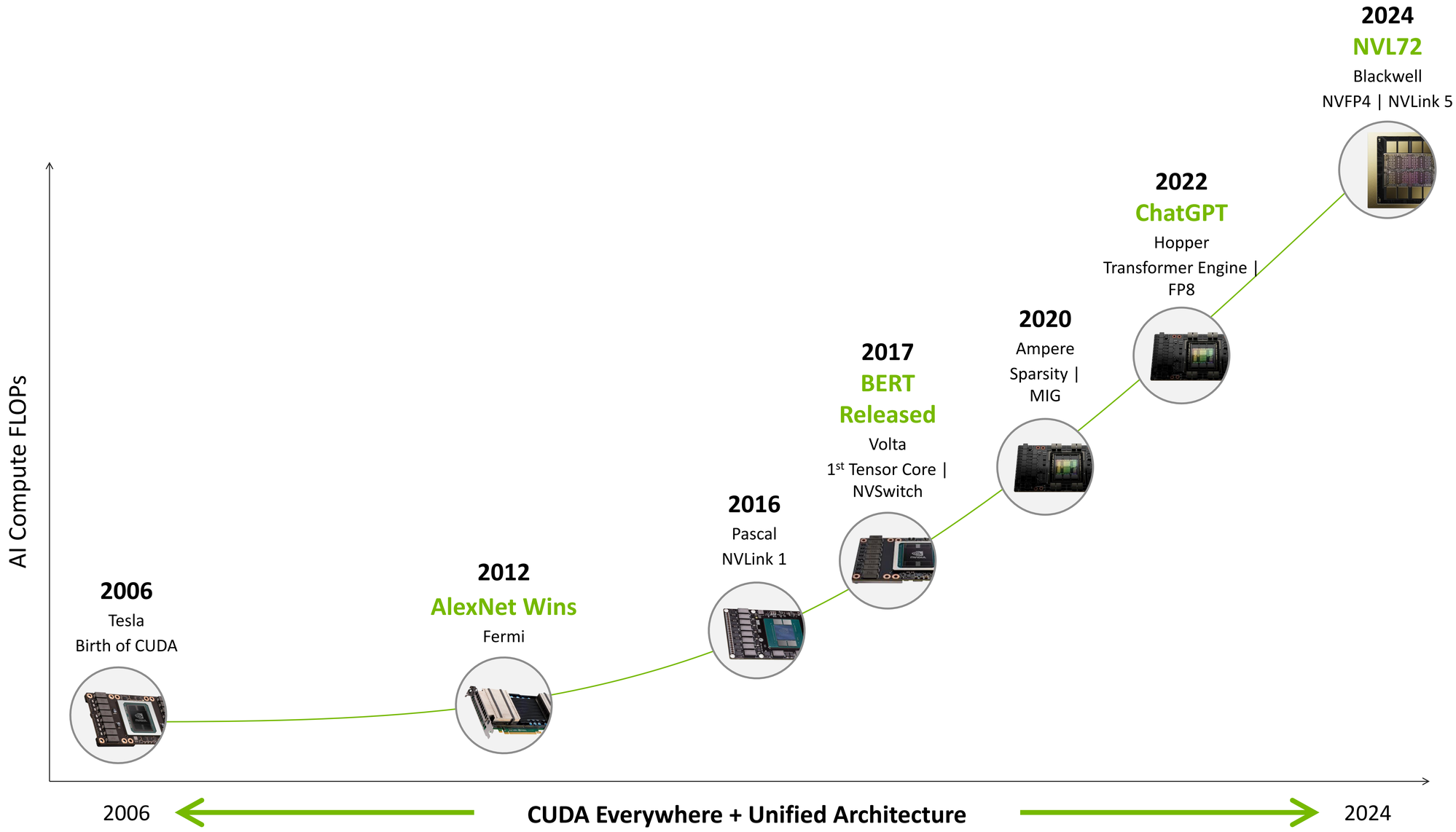
NVIDIA GB10 SoC:
AI Supercomputer On Your Desk

Building and Inspiring AI Across Generations

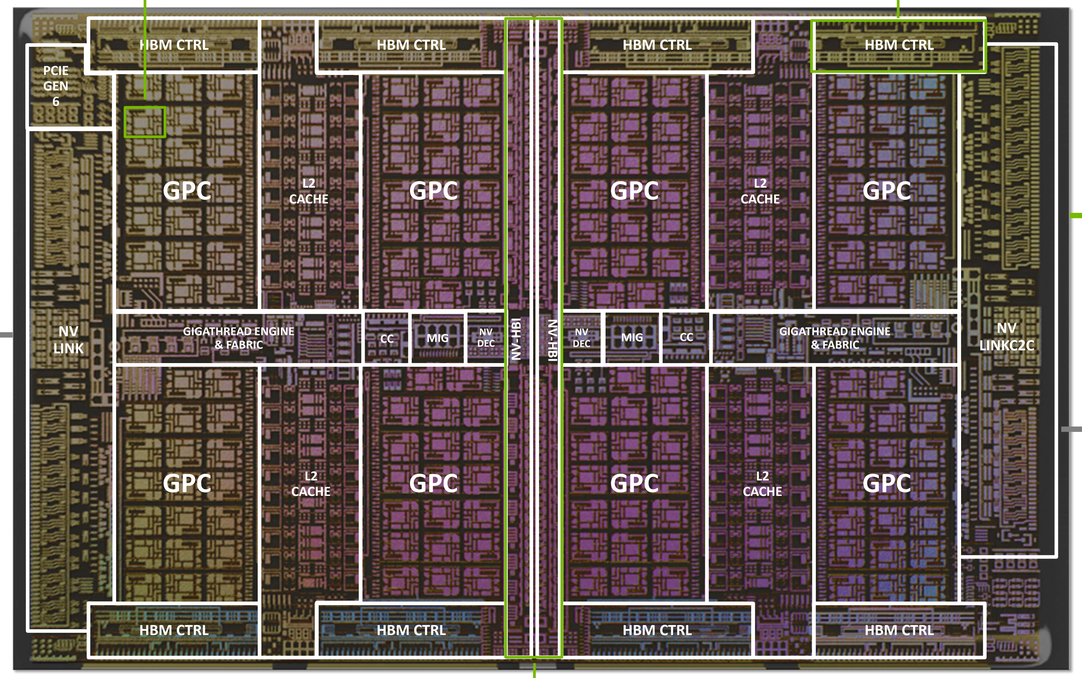


Scaling the Blackwell Architecture

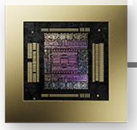
Compute Density
4x more TFLOPs/mm^2
vs Hopper

NVFP4 Tensor cores
15 PetaFLOPS
Dense NVFP4

288GB HBM3E Memory
(8 Stacks, Up to 8 TB/s)



NVSwitch



SHARP
In Network
Compute

NVLink 5
1,800 GB/s

High Bandwidth Interface
10 TB/s Die-to-Die

Grace CPU

NVLink-C2C
Up to 900GB/s Coherent CPU-GPU Interface

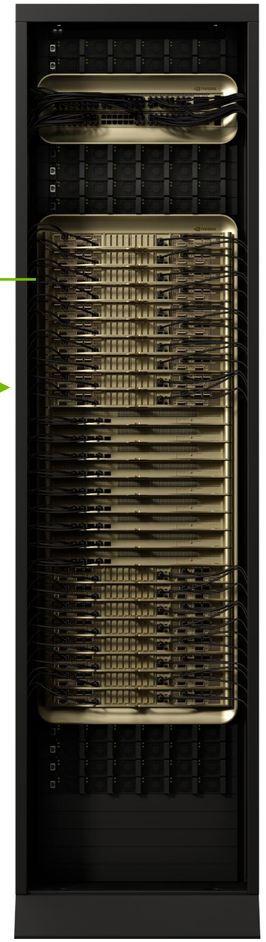
Datacenter Integration

Same architecture,
optimized for developers

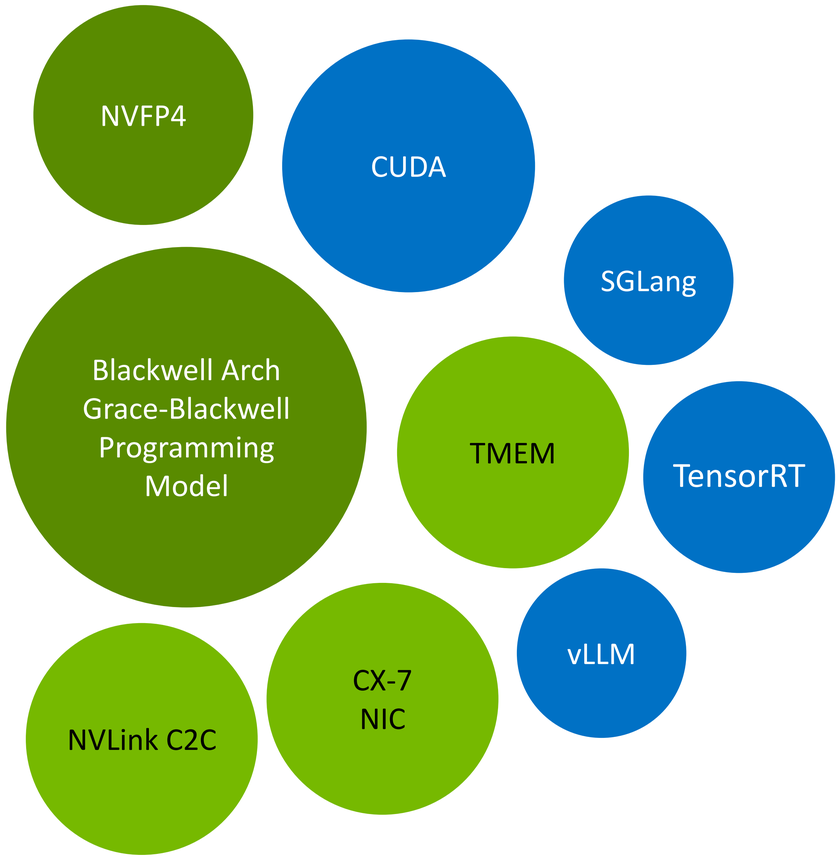
DGX Spark

NVIDIA Spectrum-X
& ConnectX-8
800 Gbps Networking

GB300 NVL72 Rack
72 Blackwell Ultra GPUs
in a coherent domain



Building a “Mini” AI Supercomputer with Blackwell



Inheriting from the Datacenter

Small Form Factor
Multi-die Packaging
(Advanced 2.5D)

Low Power
C2C

Unified Memory
Architecture (UMA)

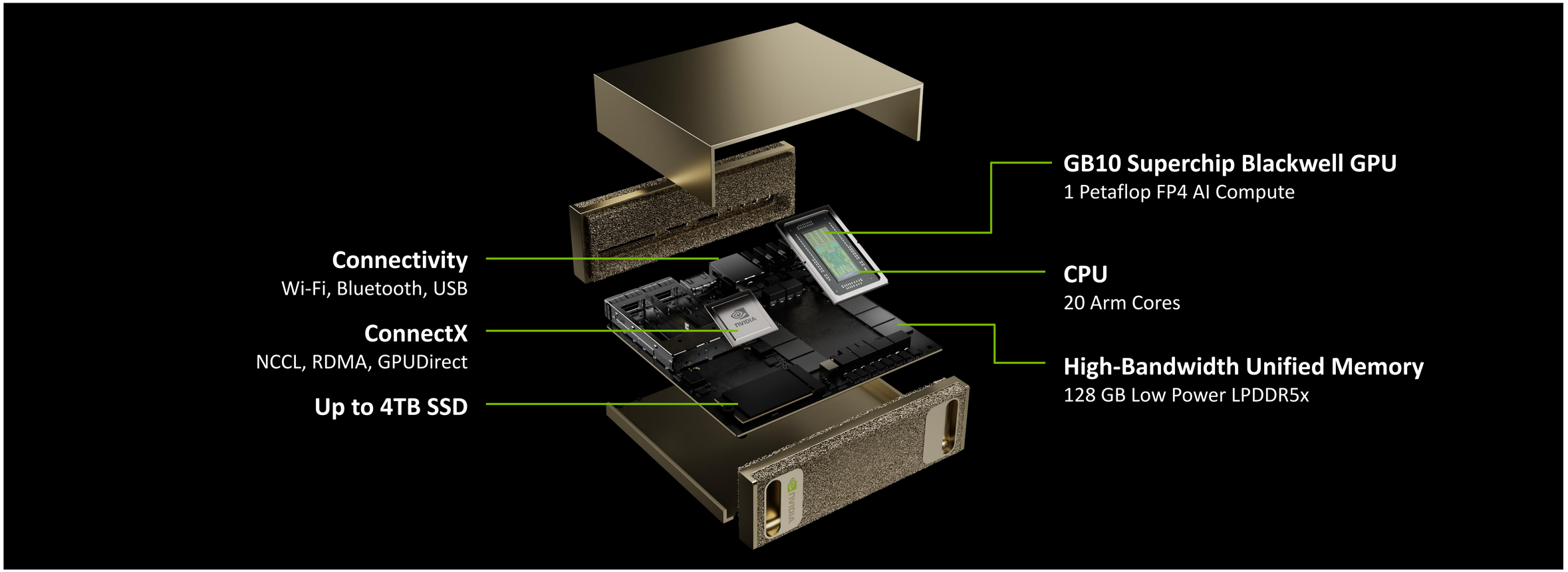
GB10 Adaptations

DGX Spark

DGX Spark Workstation Key Features and Benefits

- **GB10 Grace Blackwell Superchip:** Accelerates AI, Data Science, compute, rendering, and visualization workloads
- **128GB coherent unified system memory:** Work with large AI models of up to 200 billion parameters, fine-tune models of up to 70 billion parameters
- **ConnectX-7 networking:** Connect two DGX Spark systems together to work with models of up to 405 billion parameters
- **DGX Base OS and NVIDIA AI software stack:** Seamlessly move workloads from DGX Spark to DGX Cloud or any accelerated data center or cloud infrastructure
- **Flexible deployment configurations:** Configure as an AI workstation or a network-connected personal AI cloud
- **Great desktop experience:** Multi-head display support and flexible connectivity
- **Compact, power efficient design:** Easily fits on any desk, powered by standard wall outlet

NVIDIA DGX Spark: Powered by the GB10 Superchip



AI Superchip

FP4 Tensor Core

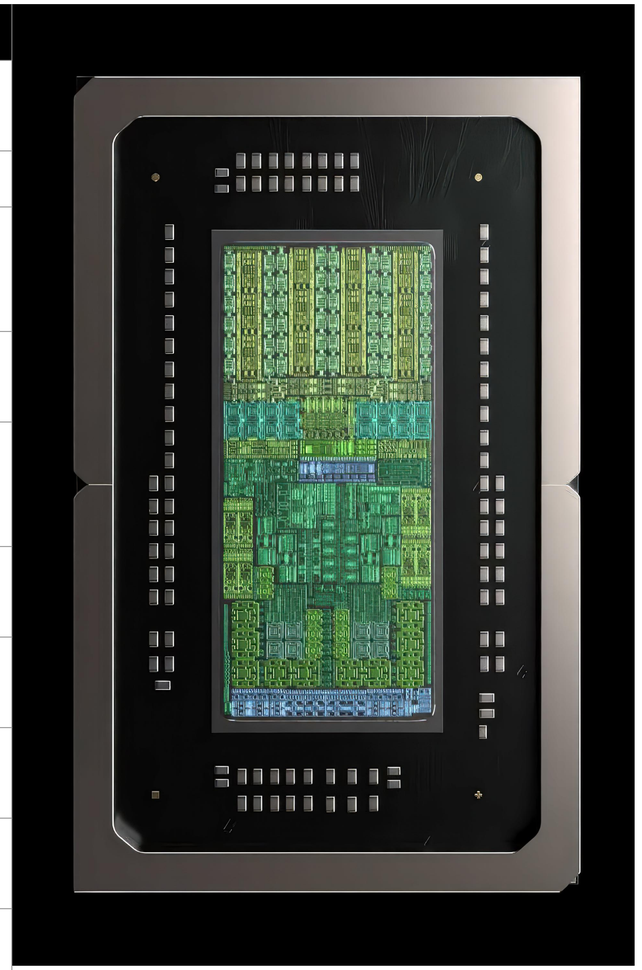
128GB Unified Memory

ConnectX

NVIDIA GB10

Specifications

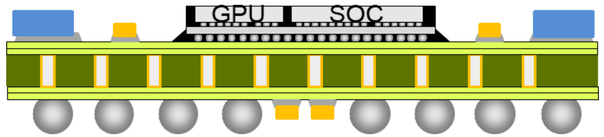
SoC Composition	S-dielet (CPU, memory subsystem, etc.) + G-dielet (GPU core) Advanced 2.5D packaging
Process Technology	Both S-dielet and G-dielet in TSMC 3nm
GPU	iGPU based on NVIDIA Blackwell Architecture 5th Generation Tensor Core DLSS-4 and RayTracing
GPU Performance	CUDA: 31 TFLOPS (fp32) AI: 1000 TOPS NVFP4
CPU	20-cores, 2 clusters of 10 cores each, each core has a private L2 cache 16MB L3 cache per cluster ARM Arch v9.2
System Fabric	High Performance Coherent Fabric Support for CHI-E Coherency Protocol
Memory	256b LPDDR5x Coherent Unified System Memory (UMA) Up-to 9400 Mbs, ~301GB/s raw bandwidth
System Level Cache	16MB Serves as L4 cache for CPU, enables power efficient data-sharing between engines
Chip-2-Chip Interface	High bandwidth, low power C2C interface NVIDIA NVLINK Architecture
HSIO Connectivity	PCIE, USB, Ethernet over PCIE

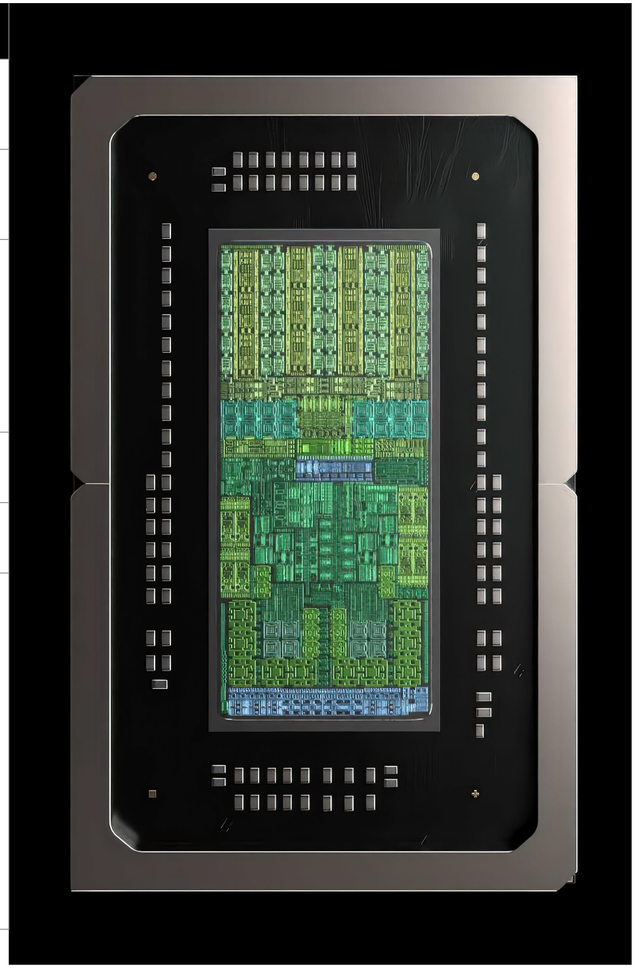


NVIDIA GB10

Specifications

Display	Up-to 4 concurrent displays 3x DisplayPort(s) + 1x HDMI
Display Output	3x DP Alt-Mode 4k @ 120Hz 1x HDMI 2.1a, FLR/12G + TDMS, 8K @ 120Hz SDR/HDR
Security	Dual Secure Root support SROOT Processor: NVIDIA Root for Secure Boot and NVIDIA secrets management OSROOT Processor: Authentication and measurement of UEFI & OS & other System SW components Support for both fTPM and discrete TPM
SoC TDP	140W
OS	Ubuntu Linux, DGX Base OS





GB10 GPU

Blackwell iGPU

- Same architecture as GB100

5th generation NVIDIA Tensor Core, 4th generation NVIDIA RT Core

- Work with large AI models of up to 200 billion parameters, fine-tune models of up to 70 billion parameters

GPU has access to entire system bandwidth over the C2C interface

- ~600GB/s of aggregate bandwidth

Large graphics L2 (GL2)

- 24MB, provides GPU compute units with large local bandwidth
- Enables CPU – GPU coherency

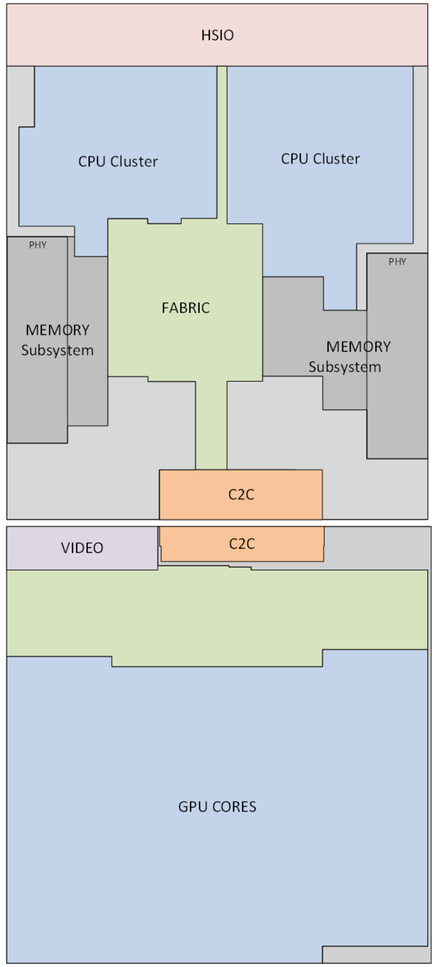
Hardware-enabled bidirectional coherency between CPU and GPU

- Simpler / easier application and software programming model
- Removes the performance overhead of Cache Management Operations (CMOs)
- Graphics L2 is physically tagged. Graphics L2 caches data in SPA (System Physical Address) space
- Support for ATS (Address Translation Services) through Graphics MMU + System MMU

Virtualization support through PCIe SR-IOV

- 1 Physical function + 255 virtual functions

Integrated NVIDIA video encode and decode engines

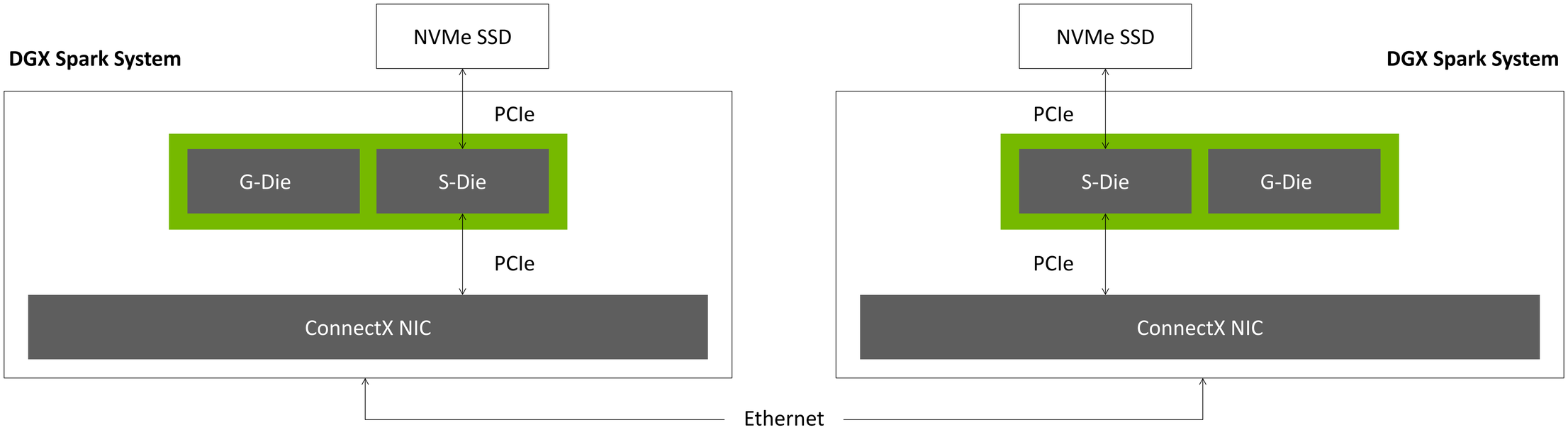


GB10 Scalability

Connect multiple GB10 chips through NVIDIA ConnectX Technology

- Scale compute throughput, bandwidth and DRAM capacity to support larger and more complex AI models
- PCIe Gen5 x8 connectivity from GB10 SoC to CX NIC

Maximize efficiency of multi-GPU parallel computing through NVIDIA NCCL framework



GB10: A Successful Collaboration Between NVIDIA and Mediatek

Merging of two architectures

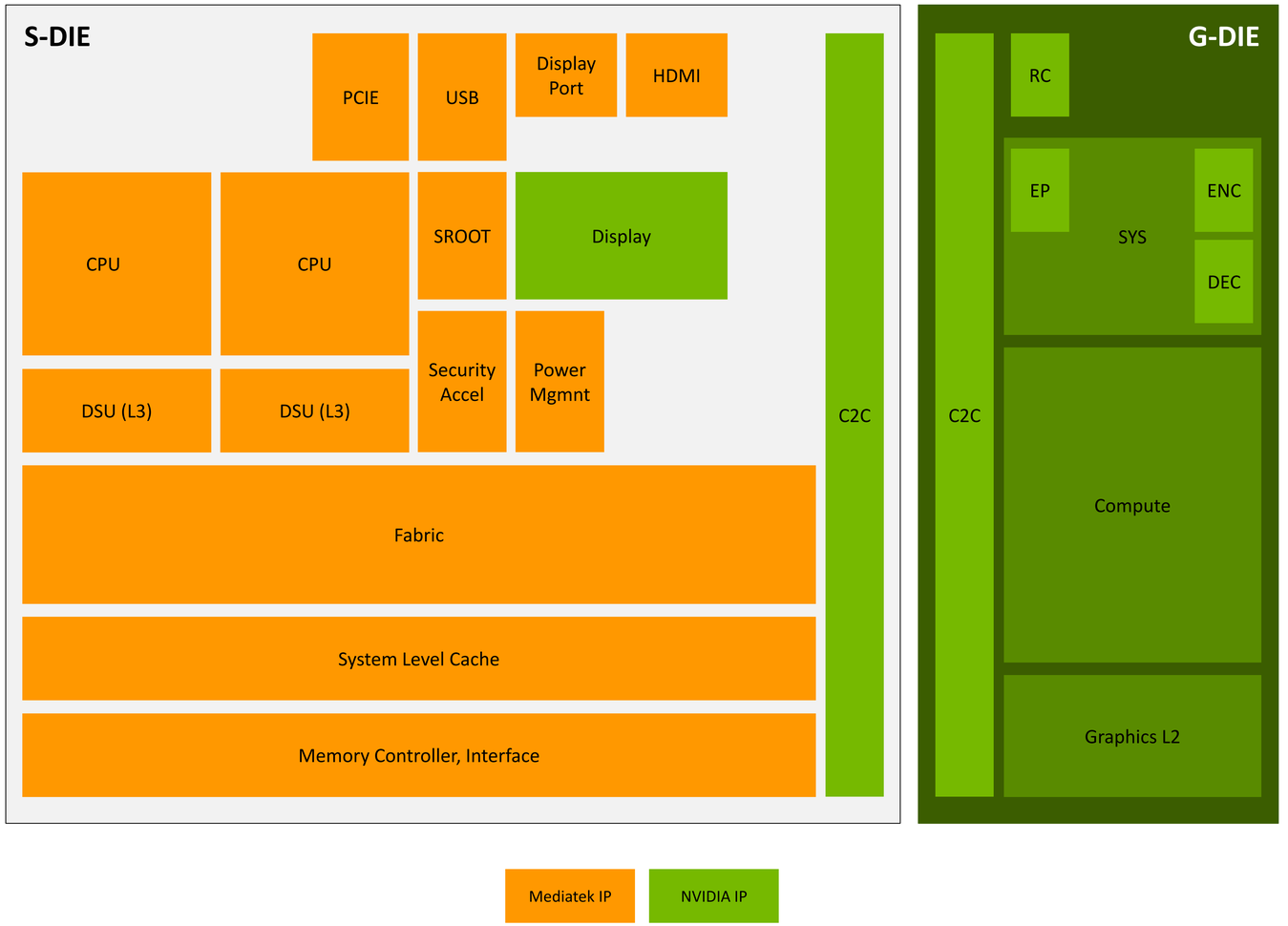
- GB10 combines MTK's power efficient CPU & Memory Subsystem implementation with NVIDIA's high-performance GPU
- Possible through implementation of industry defined standard interfaces and protocols
- Extensive performance modeling of GPU memory traffic into MTK's memory subsystem

NVIDIA IP Integration

- C2C and other NVIDIA IP delivered for integration by MTK through a well-defined IP model

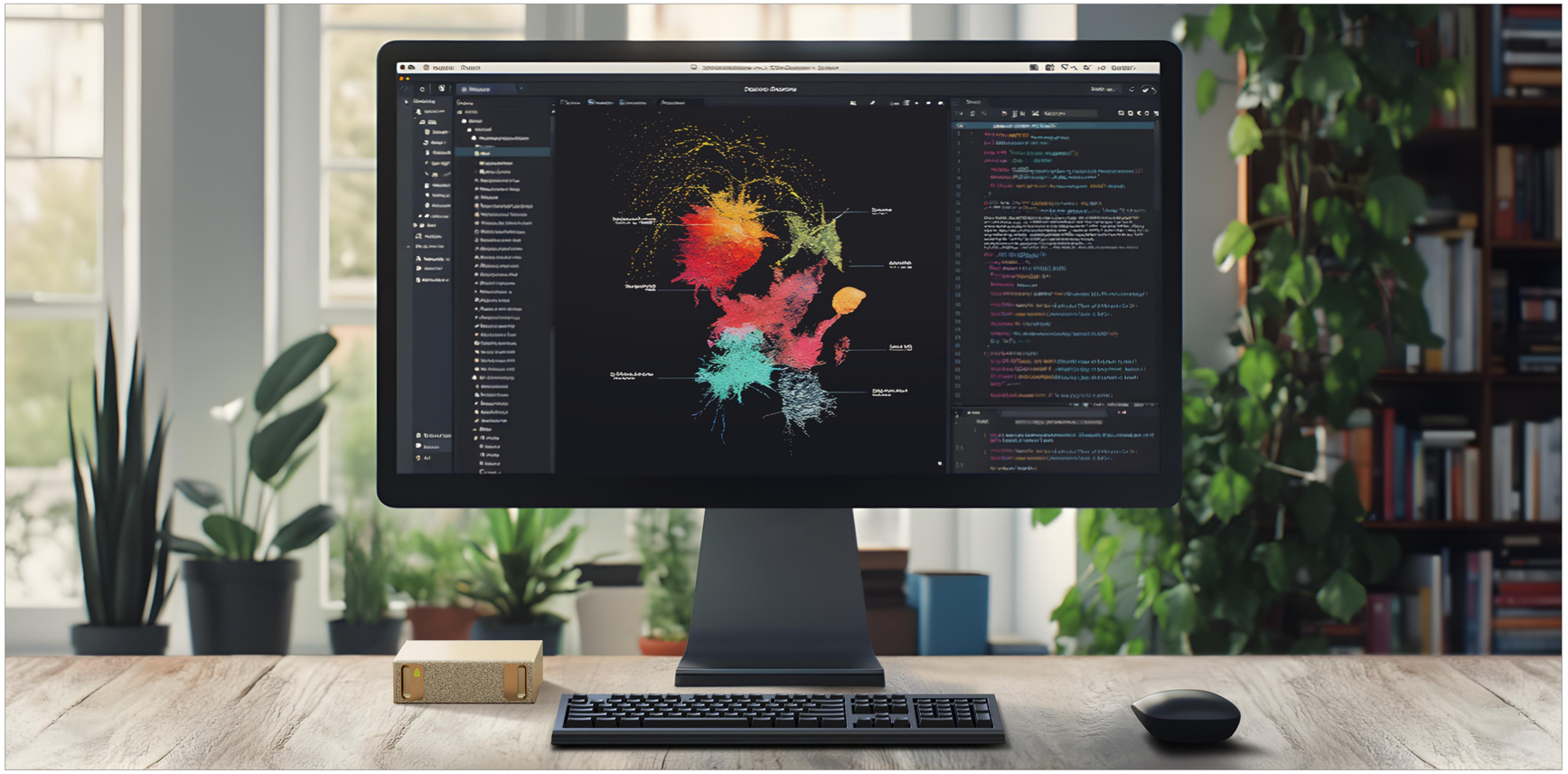
Design and Verification flow alignment

- Hierarchical functional verification approach with heavy reliance on BFMs as well as co-simulation for complex end-to-end features
- Heavy reliance on emulation to validate MTK's + NVIDIA's HW, FW and SW working together



NVIDIA DGX Spark

NVIDIA DGX™ Spark is part of a new class of computers designed from the ground up to build and run AI

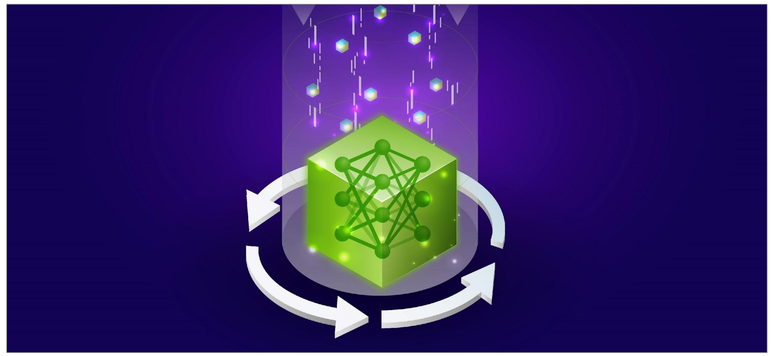


Brings the power of Blackwell GPU to millions of developers

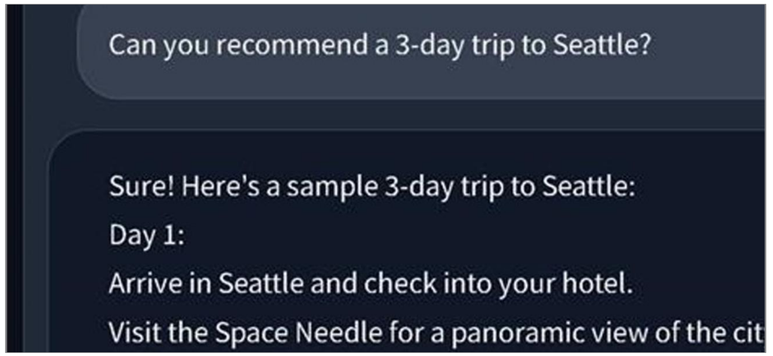
Key Workloads

Accelerating AI workloads for developers, researchers, data scientists, and students

Prototyping

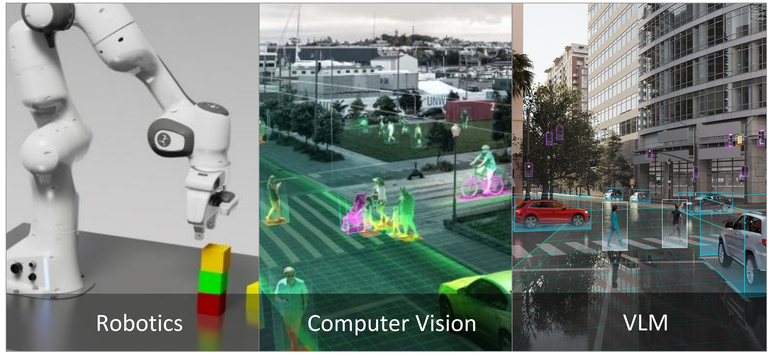


Fine Tuning



Inference & Gen AI

Data Science



Edge Applications