

COMPUTEX 2026 · GTC TAIPEI KEYNOTE

# One Keynote, Five Fronts

In a single keynote, NVIDIA moved on five layers of the stack at once: developer tools, robotics processors, agentic runtimes, a robotics testing platform, and consumer-grade compute. **The pattern matters more than any one product. For anyone building AI applications today, the platform they will build on is converging into one place.**

<p>01</p> <p><b>Developer Tools</b></p> <p>Cosmos 3 open omnimodel and a full physical-AI toolkit, from data generation to deployment.</p>	<p>02</p> <p><b>Robotics Processors</b></p> <p>Jetson Thor positioned as the edge platform for autonomous machines.</p>	<p>03</p> <p><b>Agentic Runtimes</b></p> <p>Vera Rubin in full production, with a Vera CPU purpose-built for agents.</p>	<p>04</p> <p><b>Robotics Testing</b></p> <p>An open GROOT humanoid reference design with Isaac Sim and Isaac Lab in the loop.</p>	<p>05</p> <p><b>Consumer Compute</b></p> <p>RTX Spark brings petaflop-class local AI to Windows laptops and desksides.</p>
--	---	--	---	--

01. THE MOMENT

## Three converging eras, one ecosystem

Jensen Huang framed the keynote around three eras of AI arriving at the same time: physical AI in embodied robots, personal AI agents running on-device, and agentic AI infrastructure running at factory scale. The argument was not that any one of these is finished. It was that NVIDIA now ships a piece at every scale, and that all of them resolve to the same software ecosystem.

The sections below take each era in turn, with the announcements that back it. Specifications are as presented in the keynote and NVIDIA's newsroom. The physical-AI layer is covered first because it is where the Jetson ecosystem and the production runtime question sit.

02. PHYSICAL AI &amp; ROBOTICS

## The robot gets a model, a body, and a brain

This was the heaviest emphasis of the keynote and the most consequential layer for anyone shipping robots. NVIDIA released a frontier model for physical AI, an open humanoid reference design, and reinforced Jetson Thor as the compute under it.

OPEN MODEL

### NVIDIA Cosmos 3

[open frontier omnimodel for physical AI](#)

Built on a mixture-of-transformers architecture, Cosmos 3 unifies **vision reasoning, world simulation, and action generation** from any perspective, whether first-person teleop data, simulation, or re-projected video. It can serve as the robot's policy itself, and ships with an open toolkit for the full pipeline: data generation, Isaac Sim validation, training, and deployment. Every Jetson-powered robot is a target.

## NVIDIA Isaac GROOT Reference Humanoid Robot

[first open humanoid reference design](#)

An open hardware-plus-software design for research, used by Ai2, ETH Zurich, Stanford Robotics Center, and UC San Diego. It pairs a Unitree H2 Plus body and Sharpa Wave tactile hands (the body) with Jetson Thor compute and the Isaac GROOT stack (the brain). Software is on GitHub and Hugging Face now; hardware ships through Unitree in late 2026.

CHASSIS	Unitree H2 Plus · ~6 ft · 68 kg · 31 DoF
HANDS	Sharpa Wave tactile five-finger · +22 DoF each
TOTAL ARTICULATION	75 degrees of freedom
COMPUTE	Jetson AGX Thor T5000
SOFTWARE	Isaac GROOT models, Isaac Sim, Isaac Lab, Isaac ROS

## Jetson AGX Thor, reinforced

[the edge platform for physical AI](#)

Called out as the premier edge AI platform for physical AI and autonomous machines, and the compute powering the GROOT reference robot directly. It is the consistent on-robot target across the whole physical-AI stack announced today.

GPU	Blackwell architecture
CPU	14-core Arm
MEMORY	128 GB unified
POWER	40 to 130 W, configurable
PEAK AI	up to 2,070 FP4 TFLOPS

## A new class of PC, built for on-device agents

NVIDIA's first Arm-based consumer superchip, co-designed with MediaTek and paired with a deep Microsoft partnership for a native Windows agent runtime. Huang called it the biggest reinvention of the PC in 40 years.

## NVIDIA RTX Spark

[Arm superchip for Windows laptops and desksides](#)

A custom 20-core Grace CPU fused with a Blackwell RTX GPU, delivering local AI performance with full native CUDA, RTX, and DLSS support. The pitch is **always-on local personal agents**, large on-device LLMs and VLMs up to 120B-plus parameters, alongside RTX 5070-class gaming and creative work. Slim laptops as thin as 14 mm and deskside DGX Station for Windows systems arrive this fall from Dell, HP, Lenovo, ASUS, Microsoft Surface, and MSI, with Acer and GIGABYTE to follow.

CPU	20-core custom Grace (Arm)
GPU	Blackwell RTX · 6,144 CUDA cores
AI PERFORMANCE	1 petaflop
MEMORY	128 GB unified
LOCAL MODEL SIZE	up to 120B+ parameters

## The factories that run the agents

At the other end of the scale, the data-center platform moved into full production, with a CPU and an open model both tuned for agentic workloads rather than raw training throughput.

NOW SHIPPING

### NVIDIA Vera Rubin

full-stack AI factory, in full production

The Vera Rubin NVL72 platform, with new networking and storage, optimized for agentic AI, reinforcement learning, and large-scale inference. Huang confirmed it is now in full production, with rack assembly times cut to minutes.

AGENTIC CPU

### NVIDIA Vera CPU

88-core Arm CPU for orchestration, code, and RL

Built specifically for agentic workloads rather than general compute. NVIDIA cites **1.8x faster task completion than x86 rivals** with high memory bandwidth. Ships in systems from Dell, HPE, Lenovo, and Supermicro, and in the major clouds, this fall.

OPEN MODEL

### Nemotron 3 Ultra

550B open MoE for enterprise agents

A 550-billion-parameter open mixture-of-experts model tuned for enterprise agents, with NVIDIA citing up to **5x faster and around 30% cheaper inference** than leading open models.

## Why the magnitude is the story

**Cosmos 3 and Jetson Thor for physical robots. RTX Spark for personal agents on Windows. Vera Rubin and Vera CPU for the giant agentic factories. One software ecosystem underneath all of it.**

Read product by product, these are five separate announcements. Read together, they are one move. NVIDIA is laying a continuous path from the smallest robot to the largest factory, and tying every layer back to the same software surface: Isaac, Omniverse, CUDA, and the open models that now sit on top of them. That is the part the headlines undercount.

For teams building AI applications, the practical takeaway is simple. The silicon and the models are converging into a single ecosystem faster than most roadmaps assume. The open question each team still owns is the layer in between: how a fleet of these devices actually ships, updates, and stays in production once the demo is over. That layer does not come in the box, and it is where the real engineering cost lives.

### SOURCES & NOTES

NVIDIA Newsroom and investor relations (GROOT reference humanoid, Cosmos 3, Vera Rubin, Vera CPU), NVIDIA GTC Taipei / Computex 2026 keynote, and contemporaneous live coverage. Specifications are as presented and may be refined as products ship. Compiled by Peridio for internal and partner use.