

Advancing the Agentic World, Building a Solid Silicon Foundation

Descrizione

default watermark

COMUNICATO STAMPA â?? CONTENUTO PROMOZIONALE

Key takeaways:

SHANGHAI, CHINA â??
Media OutReach Newswire
â?? 17 September 2026 â?? HUAWEI CONNECT 2026 kicked off today in Shanghai. The first keynote of the date was by David Wang, the Deputy Chairman of the Board and Rotating Chairman at Huawei. In his speech (Advancing the Agentic World, Building a Solid Silicon Foundation), Wang highlighted the work the company is doing alongside industry stakeholders to build powerful AI infrastructure, lay a solid computing foundation, and address the challenges and opportunities in the intelligent world to come.

AI is sweeping the world faster than any previous technological revolution. Today, foundation model parameters are rapidly approaching 10 trillion, and are projected to exceed 100 trillion by 2030. AI agents can now work on the same task continuously, for hours on end. By 2030, they will be able to handle tasks that span months.

In China alone, the average number of inference tokens consumed every day has surged to around 500 trillion, and is expected to reach quintillions (1018) by 2030.

On-device AI is also advancing rapidly. On-device models for smartphones have expanded from three billion parameters in 2024 to 30 billion today, and will push toward hundreds of billions in the near future.

These trends will set a much higher bar for the scale, performance, and reliability of underlying technical systems. Only by building powerful AI infrastructure can the industry lay a solid foundation for the future intelligent world.

An intelligent world is approaching â?? and faster than ever. To lead the charge into this new world, Huawei is laser-focused on building out AI infrastructure â?? the silicon foundation for the future to come.

default watermark

In particular, Huaweiâ??s AI strategy is centered on computing power, with a focus on monetizing hardware. The company is also sharpening its competitive edge through systems and architectural innovation. Centering these efforts on SuperPoDs and SuperClusters, the company aims to build a solid computing foundation and offer a new option for the world.

Huawei is a major contributor to open computing ecosystems, and will continue to support native training for mainstream foundation models on its systems, as well as supporting a vast range of models and applications.

For customers, Huawei provides flexible on-premises and cloud compute solutions for its customers to accelerate intelligent transformation across industries.

With diverse forms of compute, including solutions for micro-, low-tier, mid-range, and massive computing power â?? Huawei is driving the expansion of on-device and in-vehicle AI, making intelligence truly ubiquitous.

Additionally, Huawei is dedicated to building next-generation communications networks to bring readily available compute and intelligence to every person, home, and organization.

SuperPoDs gain broad consensus, with adoption growing in industries, academia, and research institutes

To date, over 1,000 Atlas 900 A3 SuperPoDs have been deployed, and Atlas 950 SuperPoD is seeing large-scale commercial use. While adoption continues to grow, SuperPoDs have gained broad acceptance across industry, academia, and research institutions as a key direction for AI infrastructure. Currently, a SuperPoD is explicitly defined as a computing system in which multiple computing nodes are tightly coupled through high-speed interconnect protocols, featuring unified memory addressing across physical nodes â?? functioning like a single logical computer.

SuperPoDs are the go-to choice for AI infrastructure buildout. Right now, 100k-NPU computing clusters have become the baseline for training SOTA models. However, traditional server architectures result in intra-cluster communications that account for over 40% of total training time, severely constraining Model FLOPs Utilization (MFU). Simulation results from Huaweiâ??s Markov Lab show that a 100k-NPU cluster built with 4k-NPU SuperPoDs can deliver a 2.75x increase in MFU compared to a 100k-NPU cluster composed of 8-NPU servers.

default watermark

11-chip UnifiedBus-powered portfolio for SuperPoDs and SuperClusters; the Atlas 960E SuperPoD â??the industryâ??s first to use NPO

The Ascend series of chips is the most critical component in Huaweiâ??s 11-chip UnifiedBus-powered portfolio for SuperPoDs and SuperClusters. Development on Ascend 960 has exceeded the companyâ??s expectations, with performance doubling as planned. Ascend 960DT will be available in Q1 2027, three quarters ahead of the companyâ??s original roadmap. And the Ascend 960PR will be ready in Q3 2027, one quarter ahead of schedule.

â??Weâ??re evolving our Ascend chip series on a one-generation-a-year cycle,â?• said Wang in his keynote. â??In 2028 and 2029, we will roll out the Ascend 970 and 980 chips, respectively. Thanks to the Tau (Ï?) Scaling Law, not only will their compute specifications continue to double, but you can also expect to see huge improvements across the board in terms of memory bandwidth, memory capacity, interconnect bandwidth, and more.â?•

In addition to Ascend chips, Huawei has also developed a complete portfolio of chips for AI infrastructure, based on UnifiedBus, delivering key capabilities that cover computing, interconnect, storage, and management.

â??SuperPoDs are designed to coordinate multiple NPUs through interconnect,â?• continued Wang. â??We have developed a next-generation optical interconnect product based on near-packaged optics (NPO): the High-density Optical-interconnect-Node Engine (Hi-ONE).â?• Built on Huaweiâ??s proprietary technologies, Hi-ONE has a multi-physics design for balancing optical, mechanical, electrical, electromagnetic, and thermal performance, realizing a transmission capacity of 7.2 Tbit/s per single engine.

â??This is the industryâ??s first NPO product ready for mass production, delivering the largest transmission capacity. It is also the industryâ??s first NPO product with a built-in light source.â?•

This product combines high bandwidth and high reliability with low latency and low power consumption. This, coupled with UnifiedBus, will make it far easier to scale up SuperPoD interconnect systems.

default watermark

Recently, Huawei submitted an implementation agreement (IA) on NPO to the Optical Internetworking Forum (OIF), a standards organization. The response from numerous industry partners has been widely positive. Huawei will continue its efforts to further refine the NPO industry ecosystem.

Using Ascend 960 chips and Hi-ONE, Huawei has developed the industryâ??s first NPO-based SuperPoDs: the Atlas 960E SuperPoDs. A single Atlas 960E SuperPoD can scale up to 4,096 NPUs, delivering 8 EFLOPS of FP8 compute performance, with up to 1 petabyte of HBM capacity. With 5,500 Hi-ONE units, this SuperPoD doesnâ??t need the 48,000 800G optical modules that would traditionally be required to connect all the NPUs. This cuts power consumption by over 550 kilowatts, while doubling the systemâ??s fault-free operating time, achieving 99.8% system availability.

Combining the upgraded TaiShan 950 SuperPoD and context memory storage to power an ultrascale cluster with 1 million NPUs

As SOTA models scale to 10 trillion parameters, training and inference can no longer rely on a single AI server or AI SuperPoD â?? they require a more complex computing system. This system includes AI SuperPoDs, general-purpose SuperPoDs, and an interconnect system that features peer-to-peer interconnect and zero protocol conversion. For inference, including a petabyte-scale KV cache cluster is also a must.

To meet these demands, Huawei has fully upgraded its TaiShan 950 SuperPoD. Powered by UnifiedBus all-optical networking, this new SuperPoD supports up to 4,096 nodes with a unified memory pool of up to 256 TB. This setup significantly improves agent performance. For sandbox-intensive workloads, startup speeds for 100,000 sandboxes are 30 times faster than traditional servers, and sandbox density can be improved by an additional 25%. For vector search across 10 billion x 1,000-dimensional vectors, this SuperPoD delivers twice the search efficiency of traditional servers.

Huawei has also launched OceanStor M900 â?? a UnifiedBus-powered context memory storage cluster that delivers multi-tier KV caching for agent-heavy and longer-context workloads. Designed for agentic inference, this cluster supports one-hop direct access and provides a petabyte-scale KV cache for the L3.5 layer. OceanStor M900 also uses hybrid media and an optimized retention algorithm, extending SSD read/write lifespan by 16-fold. This ensures a higher KV cache hit rate alongside long-term stability and reliability from the ground up.

default watermark

Combining its strengths in computing and communications, Huawei has built a brand-new agentic SuperCluster to accelerate training and inference for 10-trillion-parameter models. This SuperCluster uses UnifiedBus to consolidate multiple interconnect protocols into a single unified protocol, significantly reducing protocol conversion overhead. This delivers peer-to-peer interconnect between subsystems like Ascend SuperPoDs, Kunpeng SuperPoDs, and KV cache clusters. The SuperCluster also comes with a multi-tier, high-bandwidth, and large-capacity storage system that enables direct single-hop access for all KV cache tiers.

With a two-tier, four-plane Clos architecture, the SuperCluster can interconnect up to 512,000 NPUs. When combined with a multi-rail topology, this cluster can support up to one million NPUs.

One of Huaweiâ??s core strategies: Going open source and open system to build out computing ecosystems

The Kunpeng ecosystem is driving digital and intelligent innovation across a wide range of industries. To date, the Kunpeng ecosystem has attracted over 4.16 million developers and more than 7,200 ecosystem partners from around the globe. The community currently supports over 560 open-source projects worldwide. openEuler has seen more than 20 million installations, securing the largest share in Chinaâ??s server OS market.

The Ascend ecosystem has reached a new inflection point. The Compute Architecture for Neural Networks (CANN) is the foundation of the Ascend ecosystem. Today, CANN has moved to sustained,

community-driven open-source development, which has brought the platform from usable to user-friendly.

External CANN developers now comprise 61% of all CANN developers, outnumbering internal developers for the first time. With over 5,200 monthly active developers, the CANN community has become the most vibrant open-source community in China. Whatâ??s more, over 40 models have been natively pre-trained on Ascend and CANN, making it the only proven domestic stack capable of model pre-training.

Ascend now supports over 90 leading third-party open-source projects, including PyTorch, Triton, vLLM, and veRL. With strong support from the Linux Foundation, Ascend is the first official Chinese compute platform on PyTorchâ??s website. This gives developers around the world ready access to new innovations in the Ascend ecosystem.

default watermark

Diverse forms of compute for ubiquitous on-device and in-vehicle AI

AI is expanding faster into all kinds of devices. To deliver an unparalleled AI experience across all scenarios, Huawei will continue to strengthen capabilities in four key areas:

First, Huawei will combine Kirin and Ascend chips to drive self-reliance and autonomy in on-device compute.

Second, Huawei will bring together Pangu models and third-party models to make on-device intelligence better and easier to use.

Third, HarmonyOS, as an Agent OS for ubiquitous intelligence, will be completely redefined from the ground up â?? spanning system architecture, how it operates, and interaction logic â?? to enable human-agent collaboration.

Fourth, Huawei will keep cultivating a diverse AI ecosystem, which is the foundation for its system agent Celia to thrive.

Huawei plans to build four on-device computing platforms: for AI phones, AI PCs, vehicles, and homes. Through cross-device and device-cloud compute synergy, Huawei will be able to provide distributed swarm intelligence, delivering integrated and continuous intelligent services across personal mobile, office, vehicle, and home spaces, ultimately bringing intelligence to every person and every space.

Building next-generation communications that prioritize readily available compute, because without networks, all compute is siloed

Next-generation communications networks are crucial for fully unleashing the value of AI compute. We are driving the upgrade to networks that, in addition to connecting people, will prioritize delivering readily available compute. These networks will be underpinned by 5G-A/6G, 10-gigabit optical networks, and multi-tier, low-latency bearer networks, delivering intelligent connectivity across data centers, the edge, and devices.

default watermark

Concluding his keynote, Wang expressed that AI â??may well be the final technological revolution in human history,â?• noting that its impact is deeper and broader, and coming faster than anyone could have ever imagined. â??No single company,â?• he said, â??can build an intelligent world alone.â?•

He stressed Huaweiâ??s ongoing commitments moving forward:

â??And we will continue to work together to drive shared success, growing together with our customers and partners around the world,â?• Wang concluded. â??Letâ??s work together to build a fully connected, intelligent world.â?•

Themed Advancing the Agentic World, HUAWEI CONNECT 2026 will delve into AI across three dimensions: strategy, technology, and ecosystems. You can expect an in-depth look at our latest strategic initiatives, and weâ??ll also be unveiling our all-new digital and intelligent infrastructure products, scenario-specific solutions for industries, and development tools. The event will run from September 17 to 19 at the Shanghai World Expo Exhibition & Convention Center and Shanghai Expo Center. For more information, please visit HUAWEI CONNECT 2026 online at www.huawei.com/en/events/huaweiconnect

Contatti:
Immediapress
comunicati@immediapress.it

Media Contact:HuaweiCorporate.comms@huawei.com

COMUNICATO STAMPA â?? CONTENUTO PROMOZIONALE

ResponsabilitÃ editoriale di Immediapress

â??

immediapress

default watermark

Categoria

1. Comunicati

Tag

1. ImmediaPress

Data di creazione
Settembre 18, 2026
Autore
redazione