The fertile ground for compute, the AI pioneers for industries
做厚算力黑土地 成就行业AI先锋

COMUNICATO STAMPA SPONSORIZZATO â?? Huawei Cloud: Fostering the Fertile Ground for Compute, Empowering AI Pioneers for Industries

Descrizione

(Immediapress) â?? SHANGHAI, CHINA â?? Media OutReach Newswire â?? 22 September 2025 â?? On the second day of HUAWEI CONNECT 2025, Zhang Pingâ??an, Huaweiâ??s Executive Director of the Board and CEO of Huawei Cloud, delivered a keynote speech titled â??All Intelligence: Empowering AI Pioneers for Industriesâ?•. He shared Huawei Cloudâ??s innovation and practices in AI compute services, foundation models, embodied AI, AI agents, and much more.

Constant innovation in AI Compute Service: Unleashing powerful compute in the intelligent era

This year, Huawei Cloud announced its AI Compute Service powered by CloudMatrix384. The specifications of the Huawei CloudMatrix supernode will be upgraded from 384 cards to 8,192 cards. The supernodes can support a hyperscale cluster running on 500,000 to 1 million cards, thus providing robust AI compute, an invaluable resource in the intelligent era. Huawei Cloud also announced innovative memory storage with its Elastic Memory Service (EMS), which achieved an industry first by expanding video RAM with memory. This drastically reduces the latency of multi-round conversations on foundation models, greatly improving user experience.

Huawei Cloud has deployed fully liquid-cooled AI data centers in Chinaâ??s Guizhou, Inner Mongolia, and Anhui. These AI data centers support 80 kW heat dissipation per cabinet, reduce power usage effectiveness (PUE) to 1.1, and offer AI-enabled O&M. This means

enterprises do not need to reconstruct traditional data centers or build new ones. Instead, they require only a pair of optical fibers in order to connect to the data center and access efficient AI compute, as well as full-stack dedicated AI cloud services, on Huawei Cloud.

Zhang Pingâ??an pointed out that Huawei Cloudâ??s AI Token Service abstracts away the underlying technical complexity and directly provides users with the final AI computing results. This allows users to utilize the inference computing power in the most efficient way possible. The CloudMatrix384 supernode realizes the full pooling of compute, memory, and storage resources, decouples compute tasks, storage tasks, and AI expert systems, and converts serial tasks into distributed parallel tasks, greatly improving the inference performance of the system. In scenarios involving inference tasks with different latency requirements, such as online, nearline, and offline inference, CloudMatrix384 delivers an average inference performance per card that is 3 to 4 times that of H20.

At the conference, Zhang Pingâ??an announced the official launch of the AI Token Service powered by CloudMatrix384. The service delivers superior performance, service, and quality to customers.

Tackling challenges head-on: Helping enterprises build their own models

Huawei Cloud has been honing its Pangu Models by diving into industry-specific scenarios, and has worked with its customers to tackle their most pressing challenges head-on, reimagining what is possible in these industries. Huawei uses openPangu to provide best practices for AI training and inference, making it easier for developers to efficiently use AI computing power. Zhang Pingâ??an noted that, at the same time, Huawei is developing the closed-source Pangu Model. Huawei will continually increase investment in Pangu Models, and accelerate intelligent transformation across industries.

Pangu Models have been applied in more than 500 scenarios across over 30 industries, such as finance, manufacturing, healthcare, coal mining, steel, railways, autonomous driving, and meteorology.

Moving beyond terminals: Enabling infinite intelligence evolution on the cloud

This year, Huawei Cloud launched the CloudRobo Embodied AI Platform, which deploys complex algorithms and intelligent logic on the cloud to realize more lightweight robots. By taking advantage of the massive computing power and advanced AI models on the cloud, the platform makes robot execution more intelligent. Cloud intelligence overcomes the limitations that have been holding robots back, making them applicable to more scenarios.

To build a unified, open, and secure communication channel between robots and the cloud, Huawei Cloud has launched the Robot to Cloud (R2C) Protocol. Zhang Pingâ??an announced that the first 20 partners of the R2C Protocol were officially onboard.

Kunpeng Cloud Services: Empowering industry innovation with software-hardware synergy and an open ecosystem

One of Huawei Cloudâ??s key strategies is to develop Kunpeng-powered ARM cloud services that deliver performance, security, and reliability. In the past year, the number of Kunpeng compute cores on Huawei Cloud has increased from 9 million to 15 million, an increase of 67%.

GaussDB: Building efficient, reliable data foundations based on supernodes and full pooling

Based on general-purpose computing supernodes, Huawei Cloudâ??s GaussDB databases realize the layered pooling of compute, memory, and storage resources, and allow multi-read and multi-write on any node at the same time, breaking free from the restrictions of the traditional architecture where only the primary node supports data read/write. A GaussDB cluster deployed based on computing supernodes can process 5.4 million transactions per minute, marking a 2.9-fold performance increase over non-supernode clusters.

Contatti:
ImmediapressMedia Contact:Corporate.comms@huawei.com

COMUNICATO STAMPA SPONSORIZZATO: Immediapress Ã¨ un servizio di diffusione di comunicati stampa in testo originale redatto direttamente dallâ??ente che lo emette. Lâ??Adnkronos e Immediapress non sono responsabili per i contenuti dei comunicati trasmessi

â??

immediapress

Categoria

1. Comunicati

Tag

1. ImmediaPress

Data di creazione
Settembre 22, 2025
Autore
andreaperocchi_pdnrf3x8