HUAWEI

Wang Huaxin

Cloud and AI Business Solutions

May 21, 2026

Uzbekistan, Tashkent

Building a fully connected, intelligent world.

HUAWEI

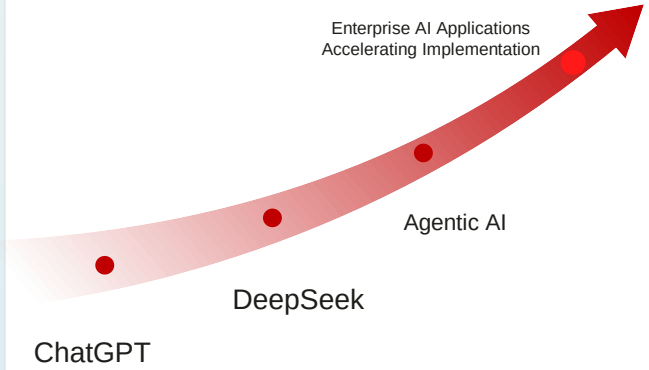
Building a fully connected, intelligent world.

Huawei Cloud MaaS and HCF For AI Industry

Model Maturity Drives AI APP into the Token Industrialization Phase, and Enterprise AI APP Enter the Trillion-Token Era

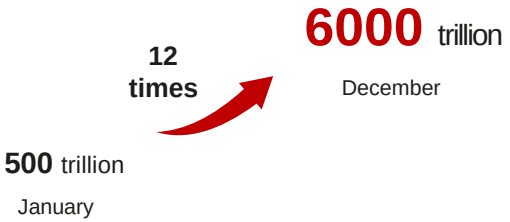
AI applications are moving towards large-scale implementation

- Narrowing gap in model capabilities
- Agentic AI becoming the mainstream model for AI applications
- AI applications are being implemented at scale in enterprises.

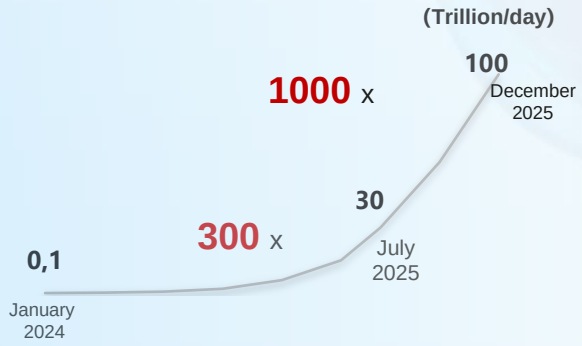


Token consumption is growing explosively

Global Token Scaling Growth



China's Token Index grows exponentially



Data sources: IDC reports, National Data Bureau, and other publicly available information from the internet, compiled and calculated accordingly

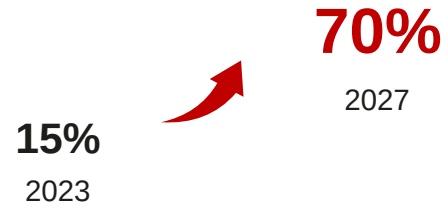
Enterprise production is moving towards large-scale token calls

Number of enterprises with cumulative token consumption exceeding trillions

700+

World's Leading Automakers | Life Service Platforms | Top global education apps

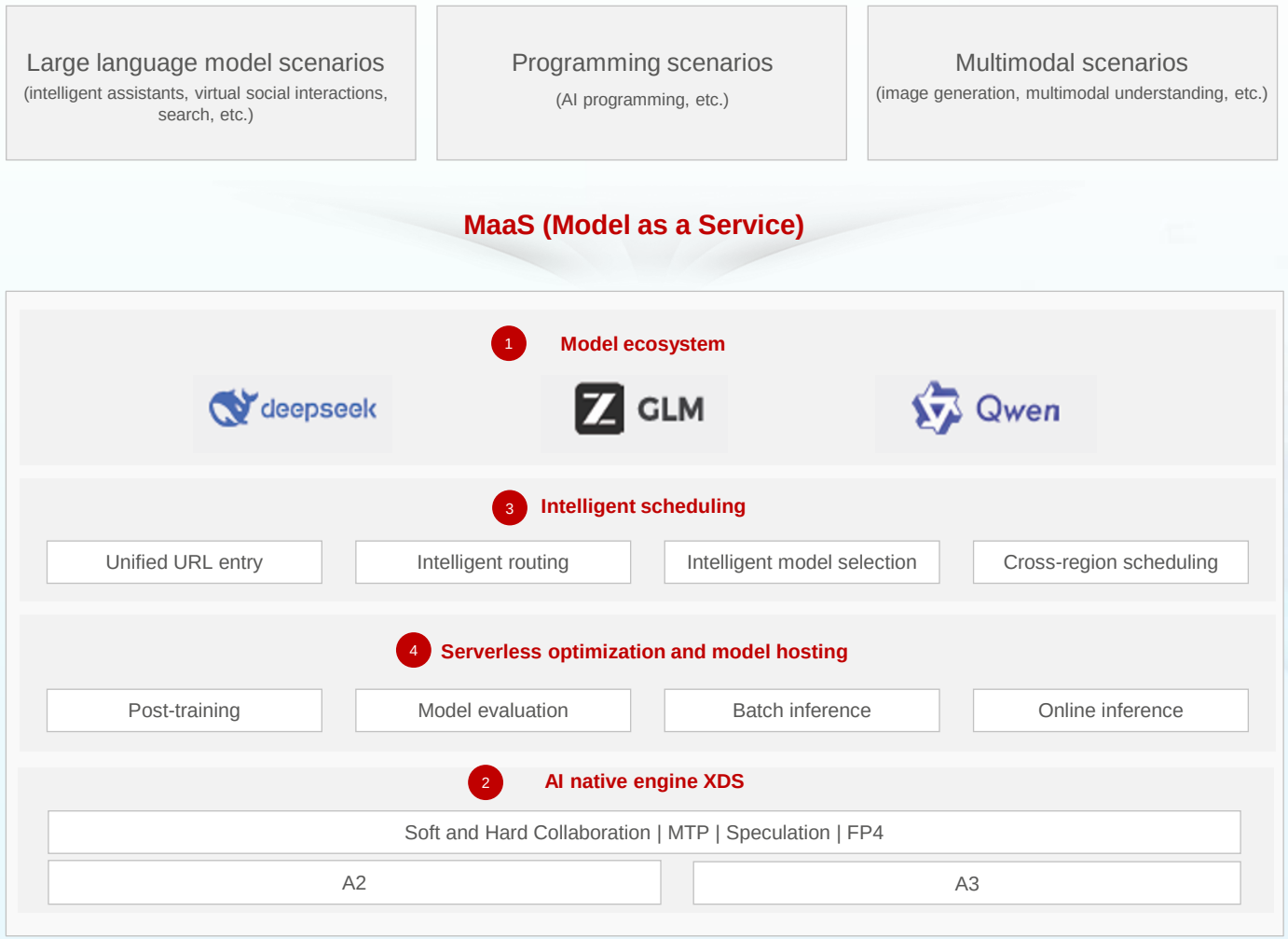
The token approach is driving a continuous rise in enterprise trends



Source: Gartner Report

HUAWEI

Huawei Cloud MaaS Builds Secure and Reliable Token Services with a Rich Model Ecosystem, Low Latency, and High Reliability.



1. Rich model ecosystem

Mainstream LLMs/Code models will be launched at the same time as China.
(All models support direct API calls.)

2. Self-developed AI acceleration engine, delivering a low-latency experience

An acceleration engine with ultimate throughput, low latency
35 ms @TPOT, ranking among the top-performing models (DeepSeek-V3)

③ Dynamic scheduling, ensuring stable and reliable services

Dynamic load balancing across multiple PUs and DUs
Rapid scaling, capable of increasing 10 million TPM within 2 minutes

④ Post-training & inference toolchain, easy to use

SFT/LoRA/RL post-training toolchain
Hosted/Custom Model Inference Deployment

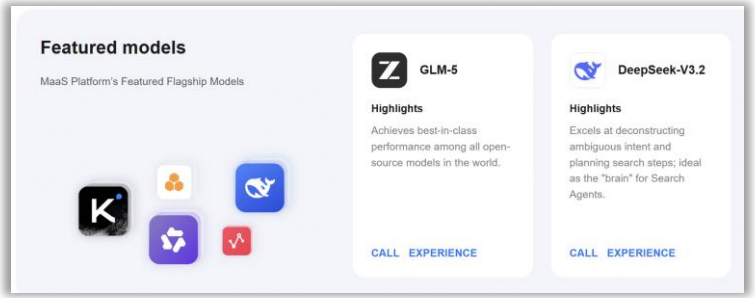
HUAWEI

High-Quality Open-Source Foundation Models in the Large Language and Programming Fields are Quickly Available

Large Language Models	GLM-5
	DeepSeek-V3.2
	DeepSeek-V3.1
	Qwen3-32B
	DeepSeek R1
	DeepSeek R1
Programming	GLM-5
	DeepSeek-V3.2

Monthly iteration
Continuous experience optimization

Mainstream open source foundation models, which will be launched at the same time as those in China



*GLM and DeepSeek models launched on Day 0 in China.

Building a Neutral MaaS Service Platform

Not utilizing Monetizing customer data

Embracing the mainstream High-quality models

HUAWEI

Use Models Appropriately Based on the Expertise

Application Scenario	Recommended Model	Model Features
General Scenarios	GLM 5	This model focuses on four key capabilities: long context, strong reasoning, top-tier coding, and autonomous agents, achieving a SOTA level of overall performance among open-source models.
	DeepSeek V3.2	Cost-effective, excels in balanced reasoning, and excels at interpreting ambiguous user intents.
	DeepSeek R1	Proficient in competition-level mathematics and complex logical reasoning, with a rigorous and transparent thought process.
Code Generation	GLM 5	Suitable for complex software engineering and end-to-end software development, large project refactoring, automated testing, and DevOps.
	DeepSeek V3.2	Proficient in balancing code engineering and efficiency, capable of generating reliable code quickly with the first-tier accuracy and low cost in the open-source code domain.
Chat Q&A	Qwen3-32B	Cost-effective, fast, and highly human-like, suitable for chat Q&A and emotional companionship scenarios.
	DeepSeek R1	Competition-level mathematics and complex logical reasoning, designed for tackling intricate problems, professional conversations, and deep problem-solving.
	DeepSeek V3	Stable responses, ultra-long context memory, seamless chat Q&A & emotional companionship conversations, with strict content control.
Content Translation	DeepSeek V3.1	Excellent performance in professional scenarios such as technical documents and academic papers, with accurate terminology, strict logic, and stable format.

Self-Developed Acceleration Engine, Software-Hardware Collaboration Optimized, Delivers a Low-Latency Service Experience

3x ms

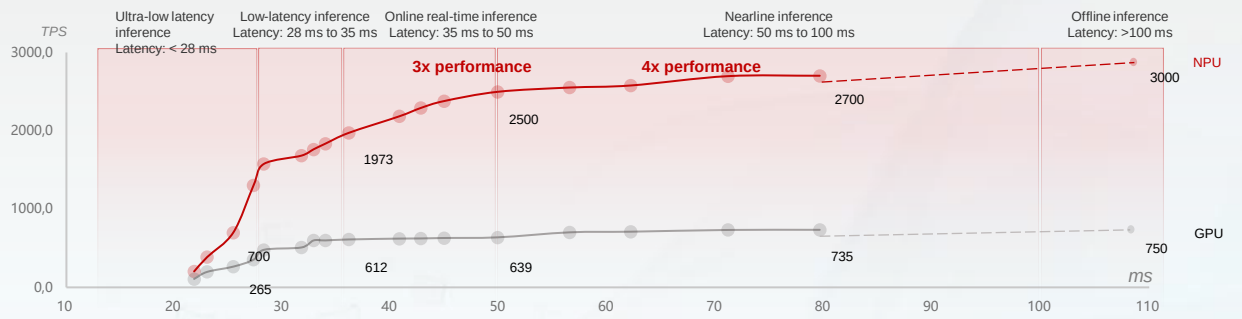
Mainstream vendors in China

vs

35 ms

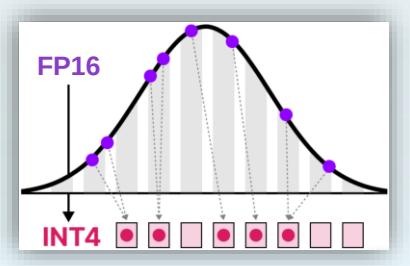
Huawei Cloud

Lower latency per token (DeepSeek-V3.2)



Full-Stack Optimization for a Low-Latency Experience

Model quantization 1.4x



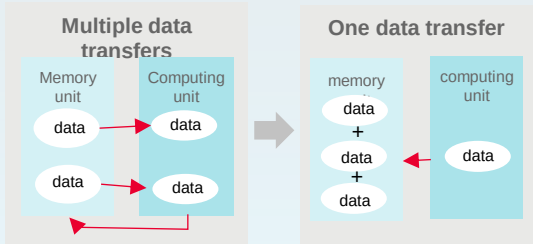
Quantized compression

W4A4/W4A8 | Sparse quantization | Communication quantization

KV cache quantization

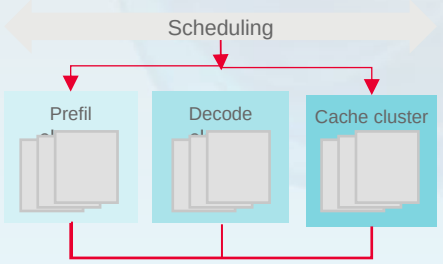
Hierarchical quantization | Context compression | Tiered cache

AI affinity Optimization of operators 2.5x



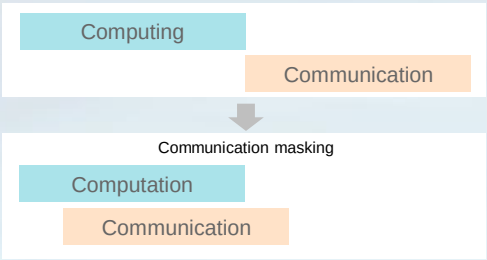
- AMLA refactoring floating-point operations
- Chip-optimized operators reduce redundant data movement.
- Operator fusion reduces data movement overhead.

PDC separation 1.5x



- Prefill is fixed in the Prefill cluster.
- Decode is fixed in the Decode cluster.
- An independent KV Cache cluster is established.

Distributed parallel optimization 1.3x



- Distributed scheduler eliminates single-point bottlenecks.
- FlashComm multi-stream concurrency
- Microbatch dual-stream masking

HUAWEI

Case: Content Moderation Safeguards Hobby

Hobby

HOBBY is a co-creation platform designed for Gen Z through dynamic content collaboration and smart recommendations. It delivers engaging, insightful, and tailored social experiences that resonate deeply with today's youth.

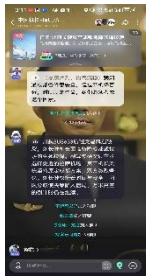
Challenges | Scenarios | Benefits

Customer Services

Dedicated virtual roles

AI board games

AI voice chat rooms

AI life simulation games

Customer Requirements

Use LLMs to simulate dialogues and story lines for diversified virtual life experiences.

- **Concurrency**: Daily active users exceed 500,000 (and five times more in the case of special activities). Each user requires a dedicated virtual role, which places high demands on model performance.
- **Latency**: In real-time interactions, a high latency impacts user experience.
- **Stability**: Core production services must stay stable.
- **Fluctuation:**
  - Traffic during peak hours can be 2 to 5 times heavier than during off-peak hours.
  - Traffic during promotions can be 5 to 10 times heavier than during daily operations.
- **Content moderation**: Sensitive content may be generated in user dialogs, which, in the worse case, may lead to serious complaints and app delisting from the app store.

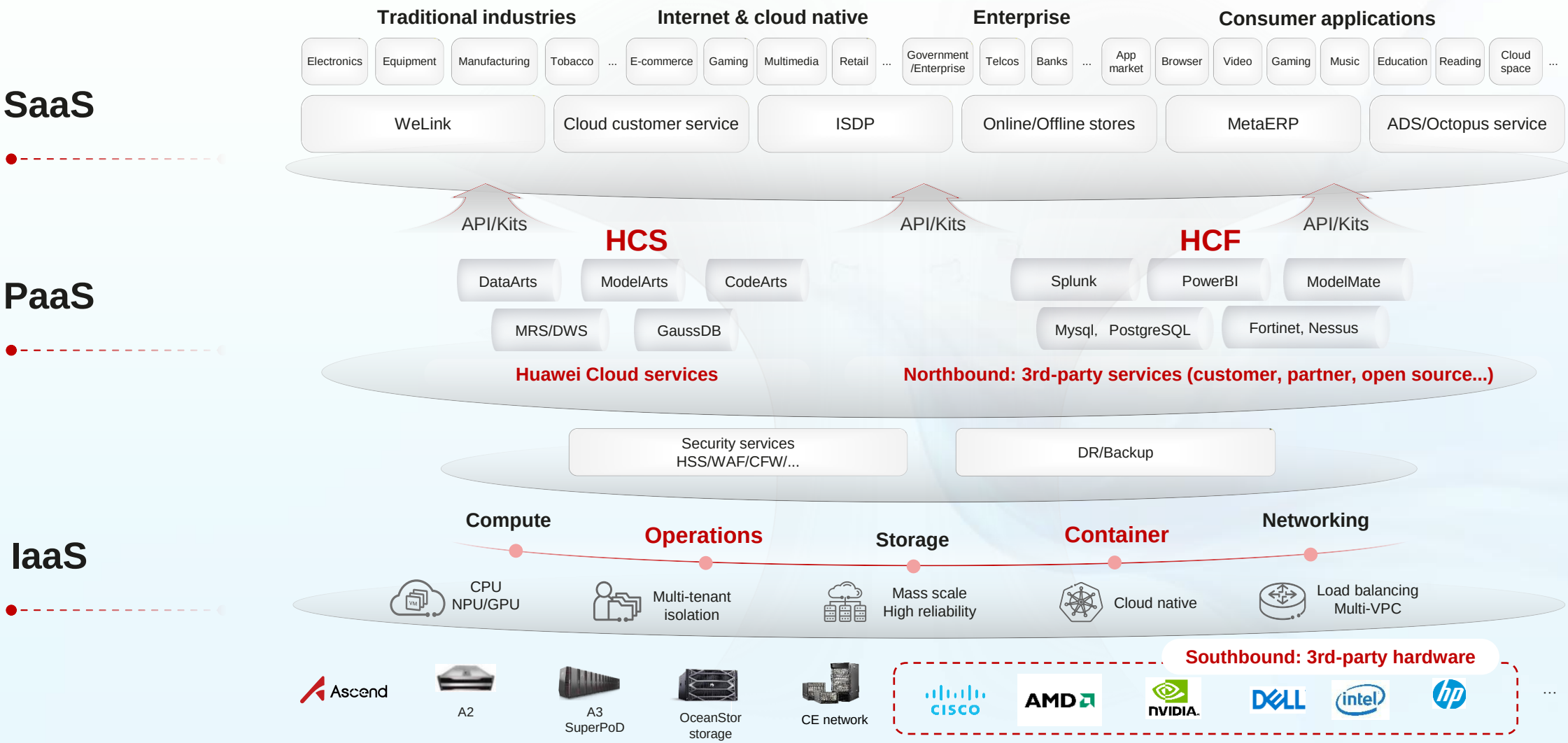
Outcomes

- **Robust compute, seamless services**: TPM up to 60 million, on-demand continuous scaling, stability in peak hours
- **Reduced latency, enhanced experience**: E2E latency down by 10-25% with CloudMatrix supernodes compared with peer vendors
- **Guaranteed compliance**: sensitive word matching + semantic interception

Use Case	RPM	Total Input Token Length	Total Output Token Length	E2E Latency Reduced
User interest evaluation	30,000	1500	50	23%
		4,800	53	24%
		6,700	50	15%
User sentiment analysis	30,000	3,000	44	12%
		3,600	43	10%
		5,400	52	13%
Greeting generation	30,000	700	18	10%
		1,900	14	12%

HUAWEI

HCF: An Open and Reliable Container-based Hybrid Cloud Platform, Built on a 9-Year Hybrid Cloud IaaS Foundation



HUAWEI

HCF's Core Competitiveness: A Cost-Effective Container Cloud Platform, Accelerating Enterprise Application Modernization

Leader in Gartner's Magic Quadrant for Container Management

Core competitiveness & customer benefits

Cost-effective containers

50%+ cheaper than competitors, accelerating application modernization for enterprises

Enterprise applications: Build once, Run Anywhere

Instant deployment of container images across on-premises and cloud different environments

Efficient resource utilization

Zero cross-container access overhead (**CCE Turbo**)

30% more efficient than K8s

HUAWEI

HCF 1.0.0: The Best Container Hybrid Cloud Foundation in the AI Era

50%+ ⬇

Basic package's per-core cost

Cost-Effective Container Cloud

Supports GPUs/NPUs

AI

8 Clou-native security

Security

DR/Backup | Safety

Database | Middleware | OS

Software

Open

Hardware

Intel | AMD | Nvidia

Dell | EMC | Cisco

Intelligent cloud management

A Unified Global Cloud

Simplified deployment

6 days > **3** days

Cost-effective

Cost reduction through flexible hardware choices

HUAWEI

Thank you.

HUAWEI