Huawei Technologies Co., Ltd
Bantian Longgang District
Shenzhen 518129, P.R. China
Tel: +86-755-28780808
www.huawei.com

Speeding Up AI Adoption to Reshape Various Industries

Huawei AI Training/Inference HCI Appliance (DeepSeek)

Built-in DCS AI E2E Software Stack

Need more information?
Email: enquiry@asl.com.hk
Tel: +852 2608 6399

01 Architecture Overview

Huawei DCS AI Solution offers FusionCube A3000 training/inference HCI appliances for the local deployment of DeepSeek, including V3, R1, and distilled models, in various industries. With a data engineering tool (ModelEngine) and standardized model and application APIs, the solution enhances the efficiency and cost-effectiveness of training, inference, and fine-tuning of industry expert models. This lightweight, professional solution supports local private deployment and ensures high-quality delivery and fast rollout, driving the development and adoption of industry-specific large AI models.

Smart office | Content creation | Intelligent customer service | Healthcare | Education & Scientific research | Operation data query

deepseek

DeepSeek-V3/R1 | DeepSeek-R1-Distill-Qwen | DeepSeek-R1-Distill-Llama

Huawei AI Training/Inference HCI Appliance (DeepSeek)

Software

ModelEngine
AI enablement platform (for data, models, and apps)

DCS eContainer
Container software

DCS FusionCompute
Virtualization software

DME
Unified O&M

Resource provisioning

Resource scheduling

Full-stack monitoring

Intelligent O&M

Hardware

FusionCube A3000

AI compute node | Ascend

Management node | HCI

Storage node | OceanStor

Network node | CloudEngine

02 Component Description

Software components

Data enablement

The AI data enablement tool enables data integration, development, analysis, insights, and sharing.

Application enablement

The one-stop tool helps develop and deploy large AI model applications with low-code orchestration, simplifying development and enabling fast implementation.

Model enablement

The AI model enablement tool provides full-process service capabilities such as model development, training, fine-tuning, inference, deployment, and management.

DCS FusionCompute

The virtualization software virtualizes servers, storage, and network devices to form elastic VM resource pools, automating resource scheduling and management.

DCS eContainer

The container software offers bare-metal and VM containers, xPU pooling, task scheduling, cluster lifecycle management, app lifecycle management, a container image repository, and O&M and monitoring.

DME

The full-stack data center management software offers IT admins with panoramic monitoring, fault self-healing, and O&M, while providing tenants with unified operation management.

Hardware components

- ✓ Storage: all-flash (high-density NVMe) + hybrid flash storage, balancing performance and costs
- ✓ Servers: Atlas cluster with high-density computing power, providing an optimal choice for large AI models
- ✓ Switches: high-performance RoCE, zero packet loss in a 8K network

03 Technical Advantages

One-stop solution

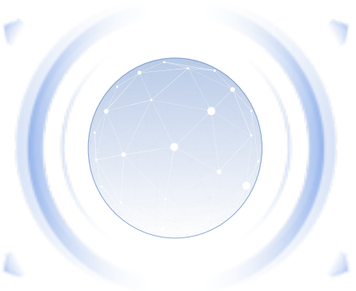
Lightweight and flexible: Supports single-node startup and on-demand smooth expansion.

Full-stack integration: Supports full integration of AI computing, general-purpose computing, storage, and network resources, meeting AI application requirements.

Fast application rollout

Efficient model training: Supports the efficient and cost-effective training and fine-tuning of general models into industry expert models.

Standardized application APIs: Visualized application orchestration and open ecosystem enable application rollout in hours.



Strong capabilities

Tool-based data engineering: Built-in E2E toolchain enables 10x faster knowledge base building.

Standardized model APIs: Automatically generated chain-of-thought (CoT) question answering (QA) data enables rapid distillation of industry-specific models.

High-quality delivery

One-stop deployment: Pre-adaptation to DeepSeek enables one-stop inference service setup.

Expert support: Huawei's expert teams, with extensive AI project delivery experience, ensure seamless implementation.

04 Typical Configuration

Product Model		FusionCube A3000 Ultra (Full-Powered Version)	FusionCube A3000 Pro (Distilled Version)	FusionCube A3000 Lite (Lite Version)
Model		DeepSeek-R1 (671B) DeepSeek-V3 (671B)	DeepSeek-R1-Distill-Qwen-32B DeepSeek-R1-Distill-Llama-70B	DeepSeek-R1-Distill-Qwen-1.5B/7B/14B DeepSeek-R1-Distill-Llama-8B
Capability		Inference (INT8)	Fine-tuning/Inference	Inference
Application Scenario		Scientific research High-end enterprise services (Complex logical reasoning)	Enterprise Q&A and data query Content creation (Cost-performance balance)	Personalized recommendation Smart office (Low costs and latency)
Throughput (Token/s)		671B: 1,911	32B: 4,940 70B: 3,300	14B: 2,920 7B/8B: 3,824
Hardware	AI compute node	2 × Atlas 800I A2 1 × Atlas 800 3000	1 × Atlas 800I A2	1 × Atlas 800 3000 (4 × Atlas 300I Duo)
	Storage node	OceanStor Dorado 5000	OceanStor Dorado 2100	-
	Management node	HCI node	HCI node	-
	Network node	1 × XH9210 1 × CE6885 1 × CE5855	1 × CE6885 1 × CE5855	1 × CE6885 1 × CE5855
Software	AI software stack	DCS ModelEngine, DCS eContainer, DCS FusionCompute, and DCS DME		
Service	Implementation	Hardware design, planning, and implementation, and software deployment (including DeepSeek model deployment)		
	AI enablement	ModelEngine		