ZETACUBE™
The Solution Leader of DePIN & AI

ZetaCube DeepSeek All-in-One Server Solution

One-Stop Delivery, Ready-to-Use

ZetaCube DeepSeek All-in-One Server is designed with the core concept of "Ready-to-Use, One- Stop Delivery." Through deep vertical optimization of hardware and software, it seamlessly adapts to various DeepSeek model specifications. The system's inference efficiency has increased by 35%, achieving breakthroughs in both efficiency and accuracy. It accelerates the efficient deploy- ment of industry-specific models and enterprise small models, driving AI accessibility and provid- ing strong support for industry-wide intelligent transformation.

One-Stop Delivery, Ready-to-Use

Pre-installed with DeepSeek's highly optimized large models, integrated hardware and software delivery. It eliminates the need for complex processes like environment configura- tion, driver installation, and algorithm adapta- tion, allowing for "zero-barrier" quick setup.

Professional Service, On-Site Deployment

A professional team provides one-on-one remote model fine-tun- ing guidance tailored to the user's scenario. On-site services are also available for full deployment, allowing users to effortlessly reap the benefits.

Comprehensive Enhancement, Ultimate Cost Efficiency

Deep hardware and model adaptation optimization improves inference efficiency by 35% and reduces latency by 40%, meeting the needs of various high-reliability scenarios. With an investment of under one million RMB, the deployment of models with hundreds of billions of parameters can be completed.

Secure and Reliable, Model Privatization

Local privatized deployment and training ensure that enterprise data stays within the domain, eliminating the risk of data leakage.

Multiple Model Options, Flexible Deployment

Three model specifica- tions (Industry-level / Enterprise-level / Department-level), supporting single-ma- chine single-card, single-machine multi- card, and multi-machine multi-card deployments, offering flexibility to meet diverse scenario requirements.

DeepSeek-671B All-in-One Server (Industry-Level)

Video Memory	Computing Power
768GB	10560 TFLOPS
Memory	**Throughput** (Total tokens/s)
1TB	1400
Concurrent Users (Simultaneous Access)	
100	

Supported Models	DeepSeek R1-671B DeepSeek V3-671B
Product Solution	8U All-in-One Deployment
CPU	4x 32-core Processors
GPU	16x NVIDIA RTX 4090
Network Interface	Dual 25GbE Ports
Software Package	DeepSeek V3/R1-671B, vLLM
Service Package	8-hour one-on-one on-site model fine-tuning guidance service

DeepSeek-70B All-in-One Server (Enterprise-Level)

Video Memory	Computing Power
192GB	1320 TFLOPS
Memory	**Throughput** (Total tokens/s)
512GB	2100
Concurrent Users (Simultaneous Access)	
260	

Supported Models	DeepSeek R1-Distill-Llama-70B
Product Solution	4U All-in-One Deployment
CPU	2x 32-core Processors
GPU	4x NVIDIA RTX 4090
Network Interface	Dual 25GbE Ports
Software Package	DeepSeek R1-Distill-Llama-70B, vLLM
Service Package	8-hour one-on-one on-site model fine-tuning guidance service

DeepSeek-32B All-in-One Server (Department-Level)

Video Memory	Computing Power
48GB	660 TFLOPS
Memory	**Throughput** (Total tokens/s)
256GB	1500
Concurrent Users (Simultaneous Access)	
50	

Supported Models	DeepSeek R1-Distill-Owen-32B
Product Solution	4U All-in-One Deployment
CPU	1x 12-core Processor
GPU	1x NVIDIA RTX4090
Network Interface	Dual 25GbE Ports
Software Package	DeepSeek R1-Distill-Owen-32B, vLLM
Service Package	8-hour one-on-one on-site model fine-tuning guidance service

ZETACUBE The Solution Leader of DePIN & AI
TEL: +82-2-1644-3096 Website: zetacube.net Email : allen@zetacube.net
zetacube.net