



Rafay

- 1 Rafay cung cấp dịch vụ K8s management. Rafay sẽ cài K8s trên top của BCM, không dùng K8s mặc định mà BCM cung cấp ⇒ Rafay sẽ nhúng K8s vào BCM. K8s management của Rafay có khả năng tích hợp với NIM, Run:AI, hoặc ứng dụng open source như Kubeflow, Jupiter notebook, ...

Rafay



The slide features a grid of service components organized into three horizontal rows. The first row contains five boxes: Consumption & Monetization (with sub-options for Compute & App SKU Manager, Baremetal As A Service, App Usage Rights Manager, Virtual Machines As A Service, App Deployment Manager, SLURM As A Service, and User Experience Manager); Orchestration & Governance (with sub-options for Workspace Manger, Policy Manager, Inventory Manager, Chargeback Manager, Match-making Service, Catalog Manager, and BCM Integration Service); and Accelerated Computing Infrastructure (with sub-options for App Deployment Manager). The second row contains five boxes: Consumption & Monetization (with sub-options for Compute & App SKU Manager, Baremetal As A Service, App Usage Rights Manager, Virtual Machines As A Service, App Deployment Manager, SLURM As A Service, and User Experience Manager); Orchestration & Governance (with sub-options for Workspace Manger, Policy Manager, Inventory Manager, Chargeback Manager, Match-making Service, Catalog Manager, and BCM Integration Service); and Accelerated Computing Infrastructure (with sub-options for App Deployment Manager). The third row contains five boxes: Consumption & Monetization (with sub-options for Compute & App SKU Manager, Baremetal As A Service, App Usage Rights Manager, Virtual Machines As A Service, App Deployment Manager, SLURM As A Service, and User Experience Manager); Orchestration & Governance (with sub-options for Workspace Manger, Policy Manager, Inventory Manager, Chargeback Manager, Match-making Service, Catalog Manager, and BCM Integration Service); and Accelerated Computing Infrastructure (with sub-options for App Deployment Manager).

Consumption & Monetization	Compute & App SKU Manager Baremetal As A Service	App Usage Rights Manager Virtual Machines As A Service	App Deployment Manager SLURM As A Service	User Experience Manager Kubernetes As A Service
Orchestration & Governance	Workspace Manger Policy Manager	Inventory Manager Chargeback Manager	Match-making Service Catalog Manager	BCM Integration Service App Deployment Manager
Accelerated Computing Infrastructure				

AI Apps **DataRobot** **dataiku** **run:ai** **NVIDIA NIM** **accenture** 

Consumption & Monetization

- Compute & App SKU Manager
- Baremetal As A Service

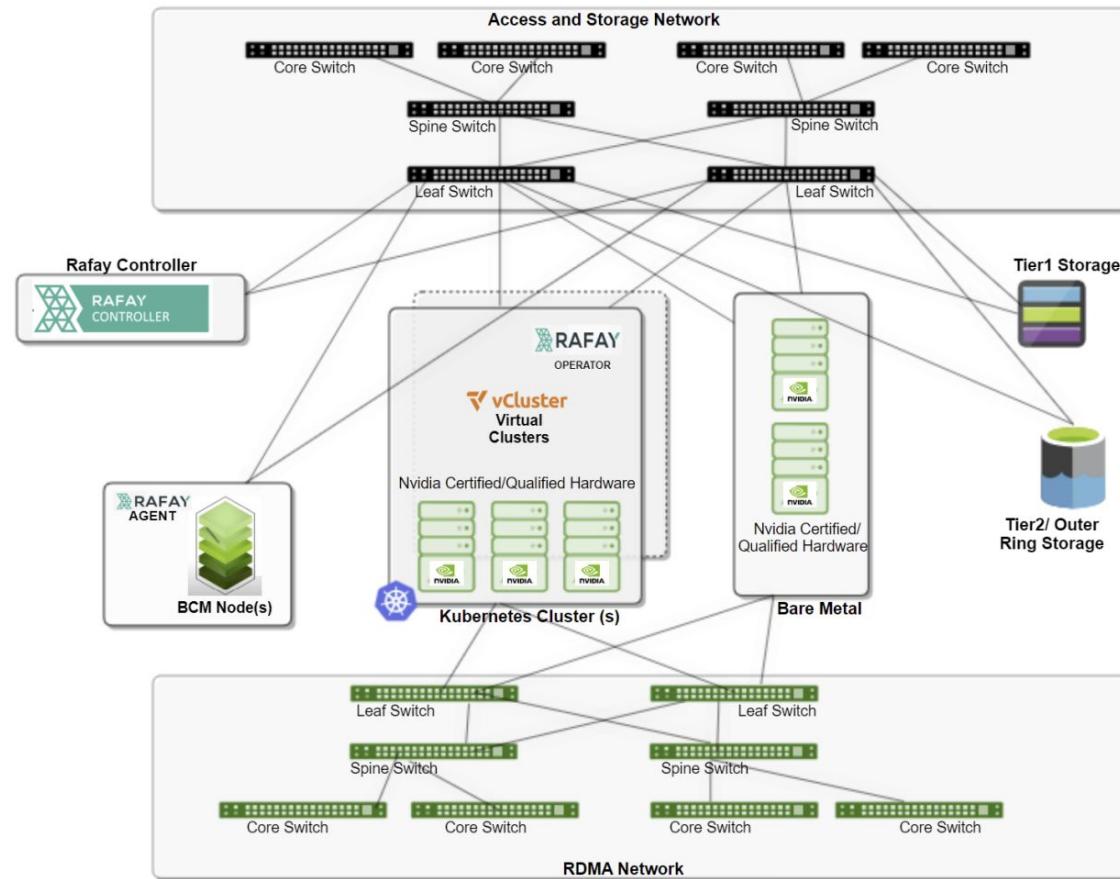
Orchestration & Governance

- Workspace Manger
- Policy Manager

Accelerated Computing Infrastructure

NVIDIA-Certified System Hardware and Network Topology

Rafay



Rafay

- 2 Rafay quản lý các **add-on** khi cài vào K8s, ví dụ add-on Nginx. Rafay quản lý version của add-on có phù hợp với version của K8s không. Rafay tự check sự phù hợp này khi muốn update version của add-on

Ref: https://docs.rafay.co/blueprints/managed_addons/

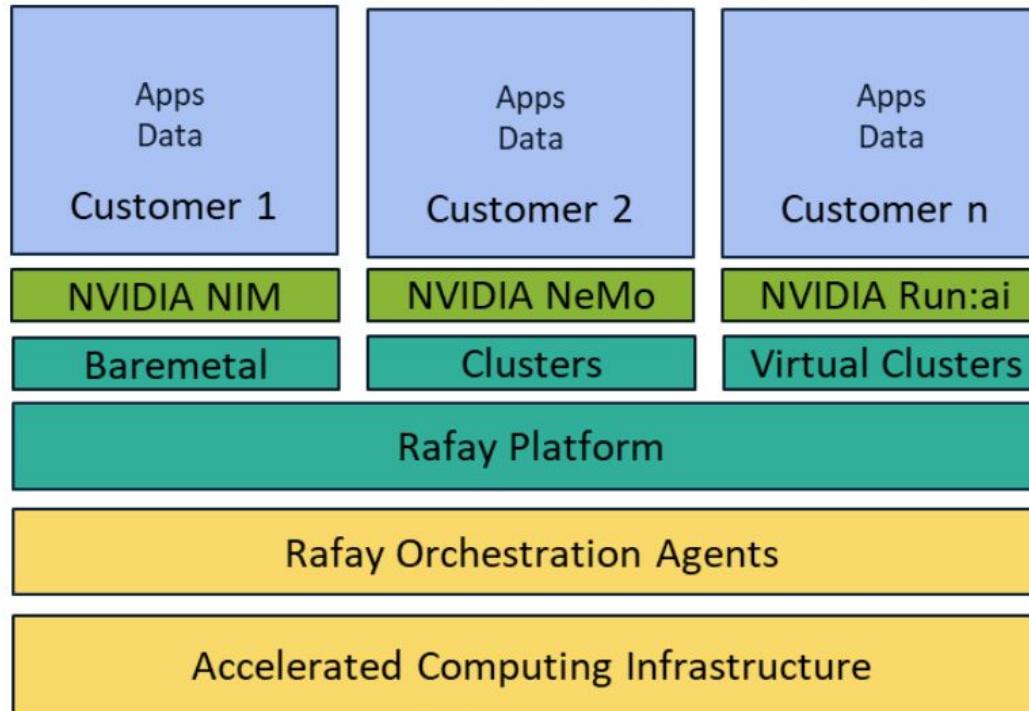
Rafay

- 3** Rafay có “**blueprint**” là tập hợp các cấu hình, add-on trên 1 project.
- Có thể share blueprint này và tái sử dụng blueprint này để apply vào các project về sau. Không cần cấu hình thủ công lại từ đầu
 - Khi add-on K8s có version mới (vd: nginx có version mới), Rafay sẽ kiểm tra tính compatible với version K8s hiện tại. Nếu phù hợp, sau khi add-on được cập nhật version, tất cả blueprint chứa add-on đó sẽ có thông báo để cập nhật version

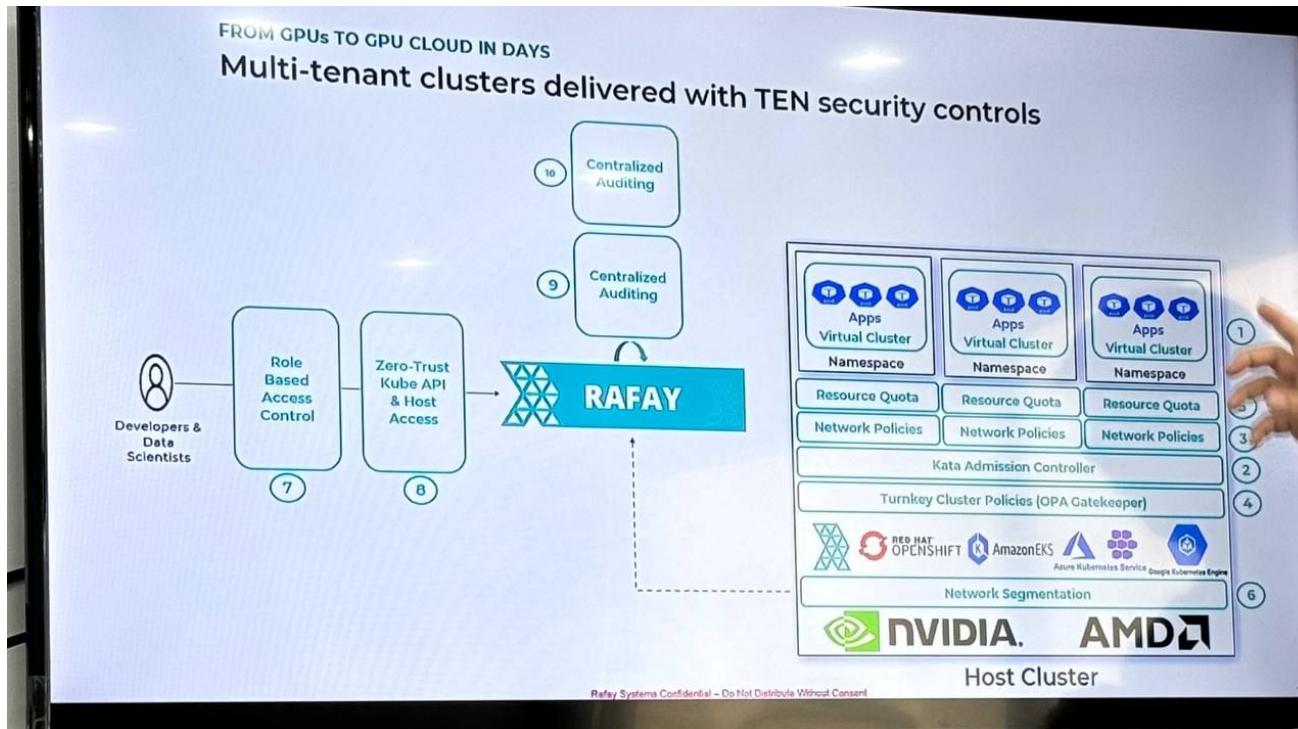
Rafay

- 4 With Rafay, CSPs can offer various flavors and sizes of GPU and compute resources (*e.g. bare metal nodes, VMs, Kubernetes Clusters, Virtual Kubernetes clusters with one or more GPUs, and fractional GPUs*).
 - Rafay có thể cung cấp giải pháp chạy nội bộ toàn bộ hệ thống trong private environment. Có thể chạy Rafay controller ở Data center nội bộ của công ty
-
- 5 Rafay có sử dụng MIG để chia GPU vào các pod

Rafay Multi-tenancy Architecture



Rafay Multi-tenancy with Virtual Cluster



Rafay Multi-tenancy with Virtual Cluster

1. Dev/test environment

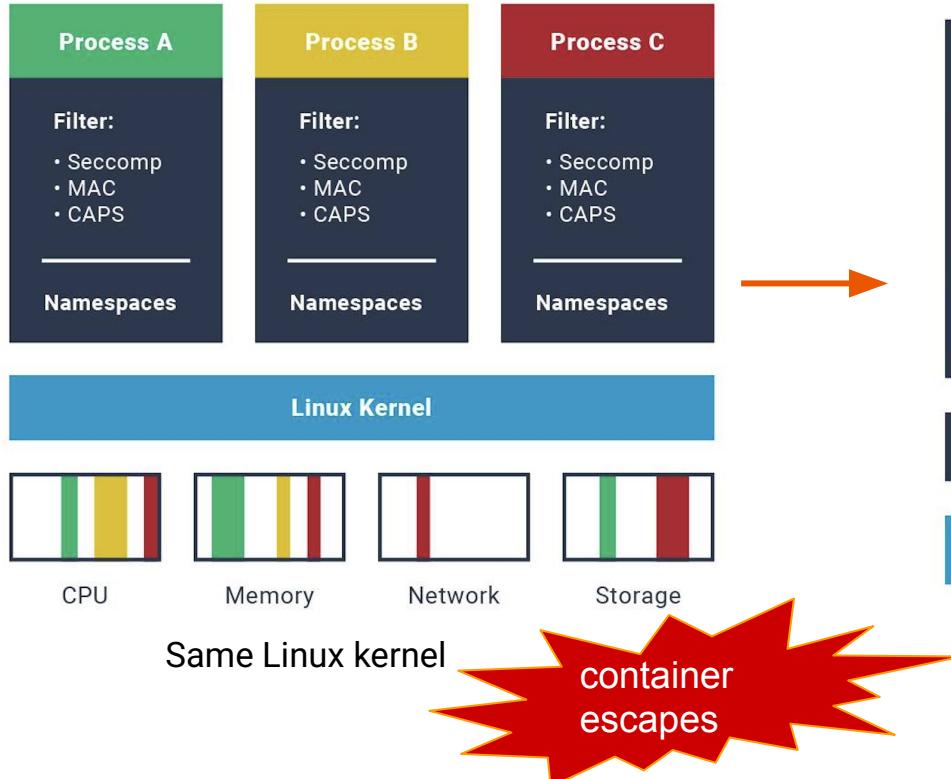
Developers can spin up lightweight vClusters quickly, test changes, and discard them when done.

2. Multiple Kubernetes Versions

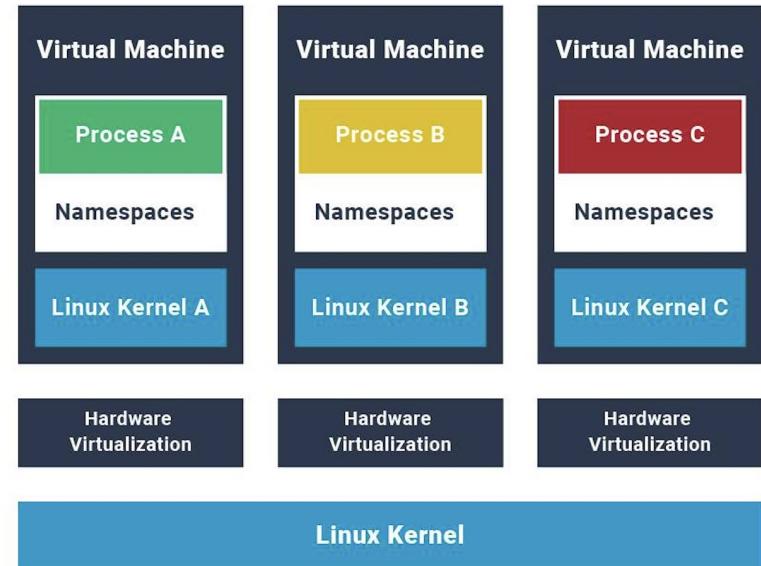
Each vCluster has its own API server

3. Isolation for Third-Party Applications

Rafay Multi-tenancy Runtime Isolation



Lightweight VM



Kata Container

container
escapes

Rafay OPA Gatekeeper

- Tích hợp OPA Gatekeeper là công cụ thi hành policy mã nguồn mở cho Kubernetes, hoạt động như một webhook kiểm tra **trước khi tài nguyên được chấp nhận vào cluster**
- Cung cấp bộ chính sách được xây dựng sẵn (turnkey) để đáp ứng các yêu cầu bảo mật và tuân thủ phổ biến như:
 - ❖ Ngăn chặn container chạy với đặc quyền root
 - ❖ Đảm bảo giới hạn tài nguyên được thiết lập
 - ❖ Thực thi sử dụng registry container đáng tin cậy
 - ❖ Ngăn chặn gắn kết volume nhạy cảm từ máy chủ

Rafay triển khai các pod dịch vụ trên vCluster

- 6 Rafay cho tạo Slurm dạng pod trong mỗi project (vCluster). Có cả slurm master, slurm daemon, ... trong vCluster đó.

Rafay gợi ý các dịch vụ cho Cloud Provider

1. GPU Bare Metal Server as Service
2. GPU Kubernetes Cluster as Service
3. GPU Virtual Cluster as a Service
4. AI Workload Deployment & Mgmt as Service
5. Inference as a Service (NIM)

Rafay References

1. [Rafay Reference Architecture]
https://rafay.co/wp-content/uploads/2025/02/Rafay-Reference-Architecture_Feb25-Final.pdf
2. [Rafay Platform Multi Tenancy Controls]
<https://rafay.co/wp-content/uploads/2024/11/Rafay-Platform-Multi-Tenancy-Controls.pdf>
3. [Rafay Platform Environment Manager Developer Guide]
<https://rafay.co/wp-content/uploads/2024/11/Rafay-Platform-Environment-Manager-Developer-Guide.pdf>



NPU - Neural Processing Unit

NPU (Neural Processing Unit) là một loại chip được thiết kế đặc biệt để xử lý các tác vụ AI và machine learning. Cho phép Inference thời gian thực và độ trễ thấp



Huawei AI Full-Stack

Huawei Cloud Builds Full-Stack Solutions for AI Development

AI Model

- Gemma
- OpenAI SORA
- LLaMA by Meta
- deepseek
- FalconLLM
- ChatGLM
- CogVideo
- Qwen
- 百川智能 BAICHUAN AI

AI Platform

- Amazon SageMaker
- ModelArts

AI Framework

- PyTorch
- TensorFlow
- PyTorch [M]^s MindSpore

Computing Architecture
The software layer that gives direct access to GPU

- NVIDIA CUDA
- Ascend NPU Driver
- CANN

AI Computing Processor

- NVIDIA
- Ascend AI Processor
- Ascend

Full-stack Self-developed AI Ecosystem

One-stop AI Platform

Self-developed Framework

Fully-supported with PyTorch Adapter "torch_npu"

Ascend NPU Driver

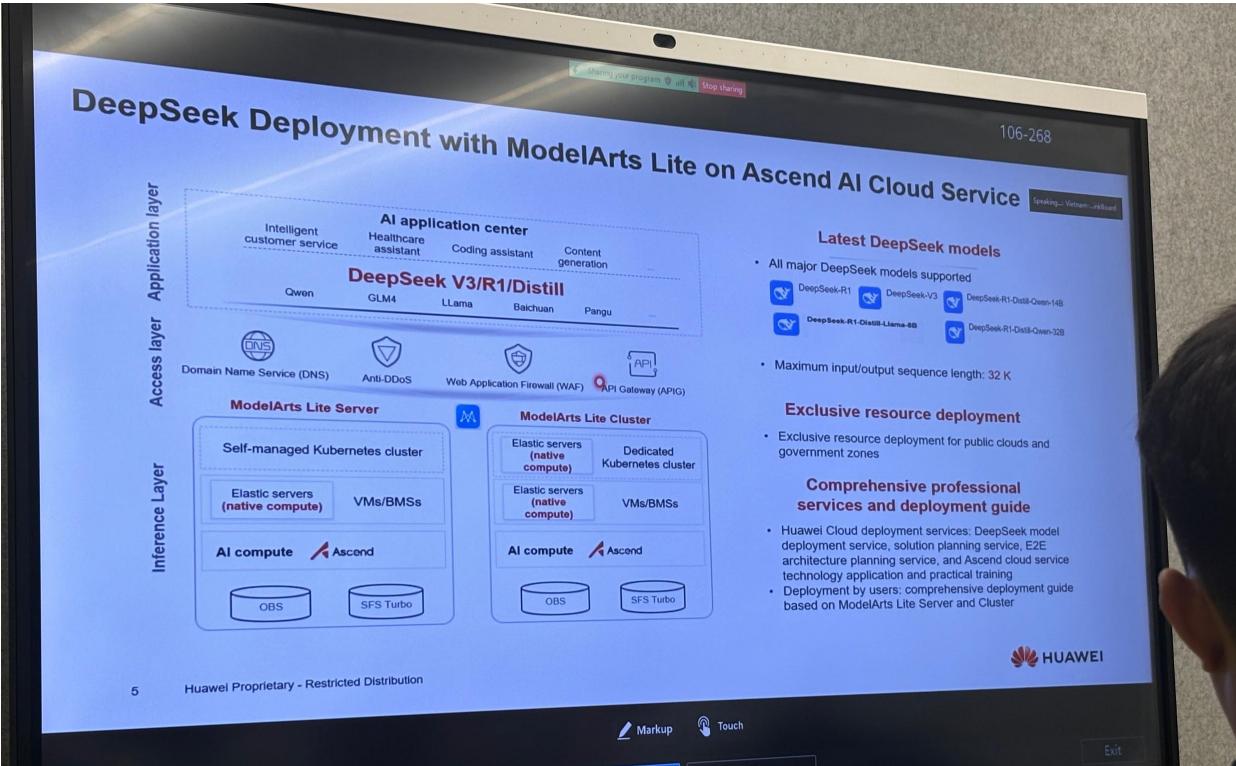
Ascend AI Processor

HUAWEI

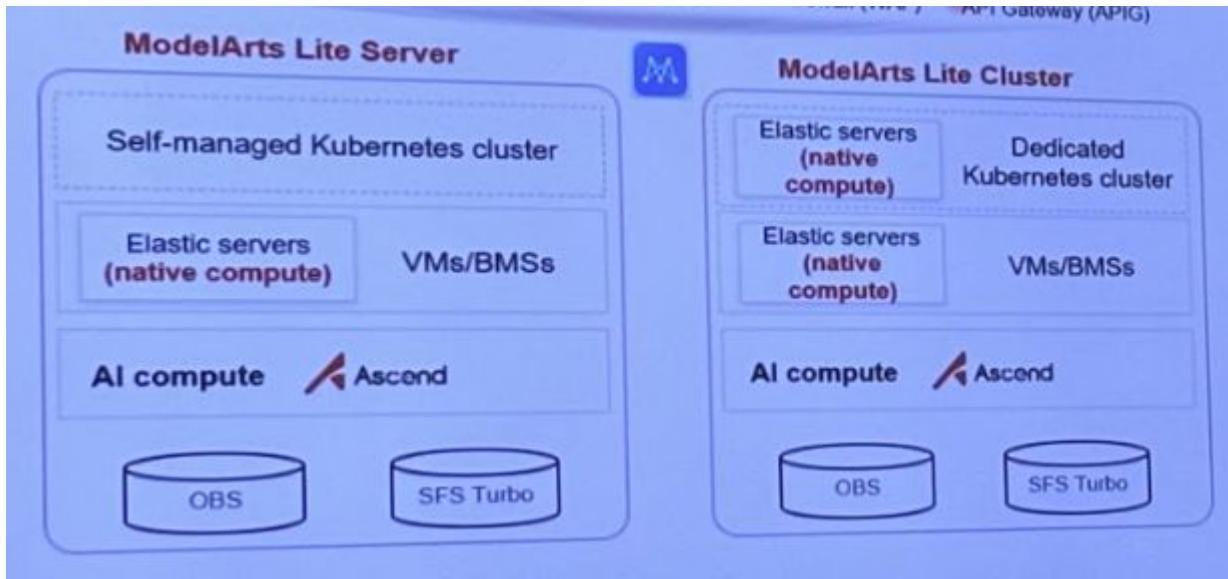
8 Huawei Proprietary - Restricted Distribution

Markup Touch Exit

Huawei triển khai ModelsArt



Huawei triển khai ModelArt



ModelArt Lite Server

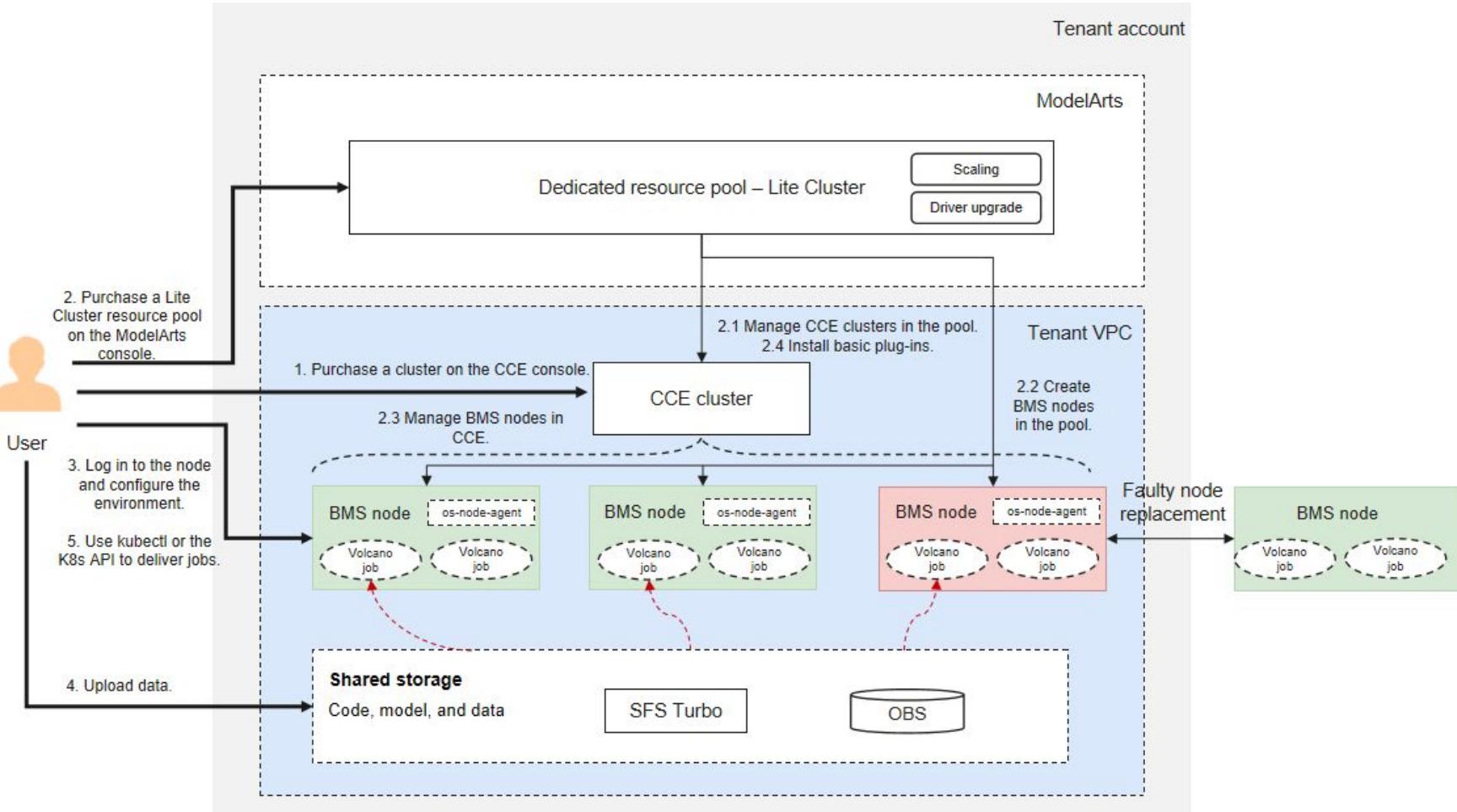
“ModelArts Lite Server provides various GPU, NPU bare metal servers (BMSs), allowing you to install and deploy third-party software such as AI frameworks and applications as user root. To create a BMS, you only need to configure the specifications, image, network, and key pairs.”

It is designed for users who have built their own AI development platforms and require only computing power. It provides cost-effective AI computing power, mainstream AI development suites, and Huawei's acceleration plug-ins. ModelArts Lite Server offers cloud servers running on bare metal servers, accessible via EIPs.

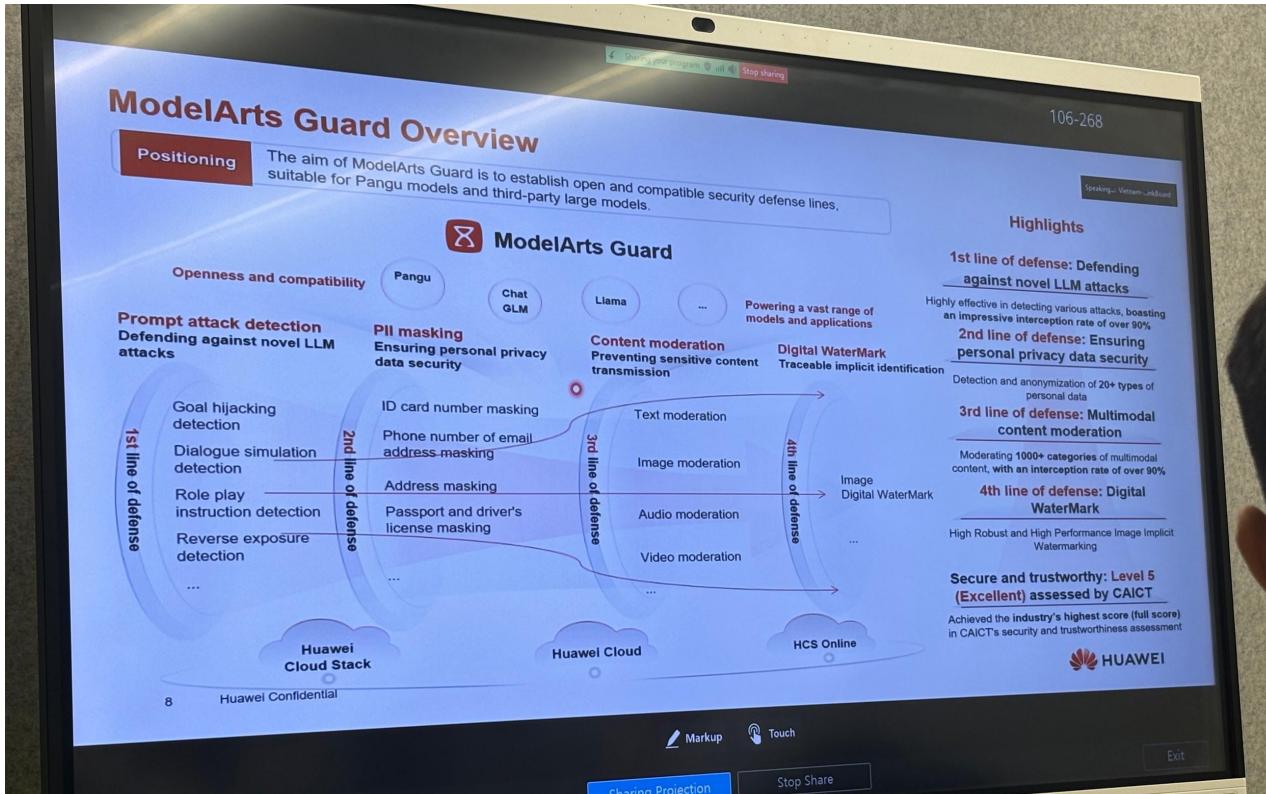
ModelsArt Lite Cluster

“ModelArts Lite Cluster offers hosted Kubernetes clusters with pre-installed AI development and acceleration plug-ins. These elastic clusters allow you to access AI resources and tasks in a cloud-native environment. You can directly manage nodes and Kubernetes clusters within the resource pools. This document shows how to get started.”

ModelsArt Lite Cluster



Các layer Guardrail trong AI Platform ModelArts





ZTE

The ZTE logo consists of the letters "ZTE" in a bold, blue, sans-serif font. The letters are thick and rounded at the top and bottom. The "Z" is slanted to the left, while the "T" and "E" are vertical. The logo is centered on a white background.

Nội dung chính

- Sản phẩm All in one AI - AI Cube
- Ứng dụng AI vào các thành phần trong hệ thống mạng
- Ứng dụng AI vào việc vận hành
- Các xu hướng chuyển dịch công nghệ trong tương lai

AI Cube

AiCube Helps to Customize a Large Model Easily

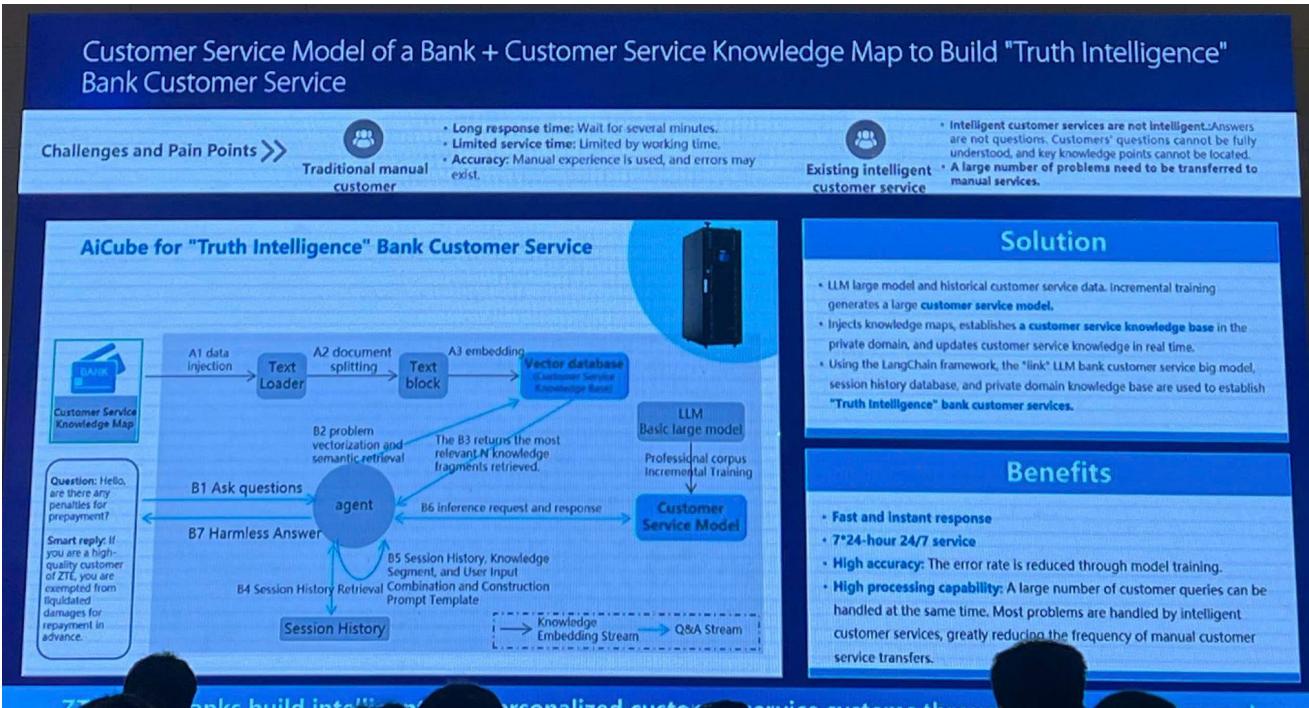
The diagram illustrates the AiCube system architecture and its integration into various AI delivery models. On the left, a physical AiCube unit is shown next to a schematic diagram. The schematic is divided into three main sections: "Multiple Models Available" (listing DeepSeek, Llama, ChatGPT, ResNet50, Stable Diffusion, YOLOv8), "Training & Inference Platform" (listing Labeling, Training, Inference, Computing MyL, Cluster, xCL), and "Hardware" (listing AI Server, Storage, and Linker Switch). Arrows point from this central schematic to four different AI delivery models on the right:

- Rapid Delivery**: Full-stack solution, hour-level delivery.
- Easy-of-Use**: End-to-end tool chain, Zero-Barrier development.
- High Security**: Local training & inference, 0-disclosure of enterprise data.
- High Utilization**: Dynamic allocation of training and inference resources.

Each model is represented by a blue box containing a server icon and a detailed description:

- AI Integrated Machine + APP**: AI infrastructure + AI platform + LLM repository. Options: Standard (16 GPU cards), Enhanced (32 GPU cards), Flagship (64 GPU cards).
- Inference Integrated Machine + APP**: AI infrastructure + AI inference platform. Options: Single server: 1 GPU server; Multi-Server: 2~8 GPU server.
- AI Application Integrated Machine**: AI infrastructure + AI platform + APP. Options: Knowledge Assistant (KAI), Coding Assistant.

Case study for Intelligence Bank Customer Service



Case study for Iron and Steel Industry

All-in-One FAQ AiCube for the Iron and Steel Industry Deepens the Intelligent Upgrade and Transformation of the Industrial Industry

Challenges and Pain points	Solution	Benefits
<ul style="list-style-type: none">Difficult document management: There are many types of internal documents of an iron and steel enterprise, and there are many storage modes.Difficult document retrieval: The staff need to review historical documents to summarize or plan new guidance documents. Manual retrieval efficiency is low.	<ul style="list-style-type: none">Based on the framework of "large model, AIS platform (RAG/Agent), and enterprise knowledge base", build the AI service capability that can be assembled and customized, and create the core competitive advantage of the large model of the iron and steel industry.	<ul style="list-style-type: none">Data privacy protection: Private deployment, and data does not leave the campus.Improved inference accuracy: Semantic understanding capability of large models, dedicated knowledge base, and retrieval enhancement algorithm, accurate understanding of chart information and context, and inference accuracy >90%.The knowledge base is continuously updated online, and system capabilities continuously evolve.

All-in-one knowledge and answer AiCube for the Iron and Steel Industry

```
graph LR
    subgraph Top [All-in-one knowledge and answer AiCube for the Iron and Steel Industry]
        direction TB
        A[Knowledge documents of an Iron and steel enterprise (PDF/Word/Excel)] --> B[Document Analysis (Picture/Text/Table)]
        B --> C[Segmentation text vectorization (Embedding small model)]
        C --> D[Knowledge Repository]
        style A fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style B fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style C fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style D fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
    end
    subgraph Bottom [Online Knowledge Q&A]
        direction TB
        E[The user inputs the test sequence] --> F[LLM Large Model]
        F --> G[Intent Identification]
        G --> H[Identify whether it is a knowledge question]
        H --> I[Fast vector retrieval of the knowledge repository]
        I --> J[Preliminary Screening]
        J --> K[Top K answer with a matching degree]
        K --> L[Rerank small Model]
        L --> M[Careful Screening]
        M --> N[Prompt]
        N --> O[Generate the answer]
        O --> P[Answer after large model summary]
        P --> Q[User Interaction Display result]
        style E fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style F fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style G fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style H fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style I fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style J fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style K fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style L fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style M fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style N fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style O fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style P fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
        style Q fill:#e0f2ff,stroke:#0072bc,stroke-width:1px
    end
    A -.-> E
    D -.-> P

```

AI Computing: i9-6500 G5 + L20

Software Platform: TECS resource management platform, AIS training and inference platform

Model: LLM, Embedding model, Rerank model

ZTE has developed a dedicated fine-tuning model and a private domain knowledge repository, running in a closed-loop mode to meet the requirements of iron and steel enterprises in business scenarios.

Case study for Automotive Industry

Working with Dongfeng Auto to Build an All-in-One AiCube for Vehicle Design

Background: Dongfeng Auto, Hubei Mobile, and ZTE jointly built an all-in-one AiCube for automobile design to simplify the design process of the automobile industry, improve design efficiency and quality, and meet the diversified requirements of the modern automobile industry.

AiCube for Automotive Design (Single Version)

Industry Application

Dongfeng Automobile Design Platform

Large model
Large Model of Dongfeng Auto Wensheng
(LoRA fine-tuning: Open source model and automobile texts and pictures)

AI platform
AIS+ TECS Resource Management Platform (Basic Version)

AI hardware
GPU Server R6500 G5
• High-performance domestic GPU * 8
• PCIe Gen5 high-speed interconnection
• bandwidth

Domestic Intelligent Computing Service Dongfeng Vehicle Model Design to Improve Vehicle Design Efficiency

Model Design → Model sketch → Plane Rendering → Review screening → Modeling

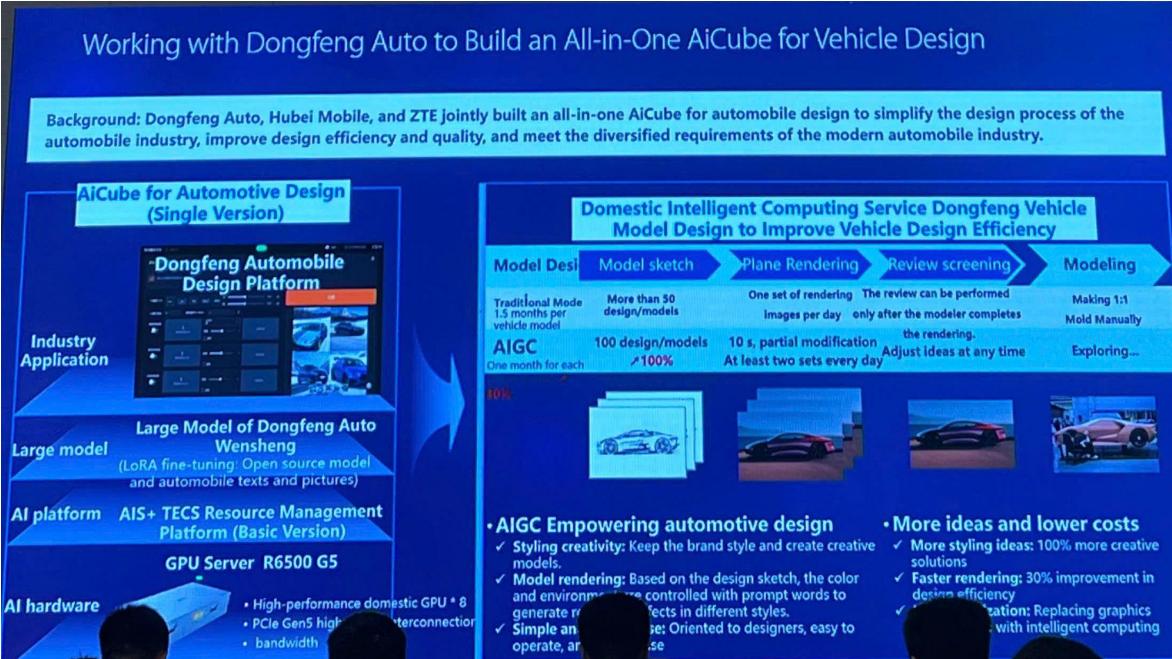
Traditional Mode	AIGC
More than 50 design/models 1.5 months per vehicle model	100 design/models One month for each
More than 50 images per day	>100% 10 s. partial modification
The review can be performed only after the modeler completes the rendering.	At least two sets every day
Making 1:1 Mold Manually	Adjust ideas at any time
Exploring...	

AIGC Empowering automotive design

- ✓ Styling creativity: Keep the brand style and create creative models.
- ✓ Model rendering: Based on the design sketch, the color and environment can be controlled with prompt words to generate rendering effects in different styles.
- ✓ Simple and fast operation: Oriented to designers, easy to operate, and quick results.

More ideas and lower costs

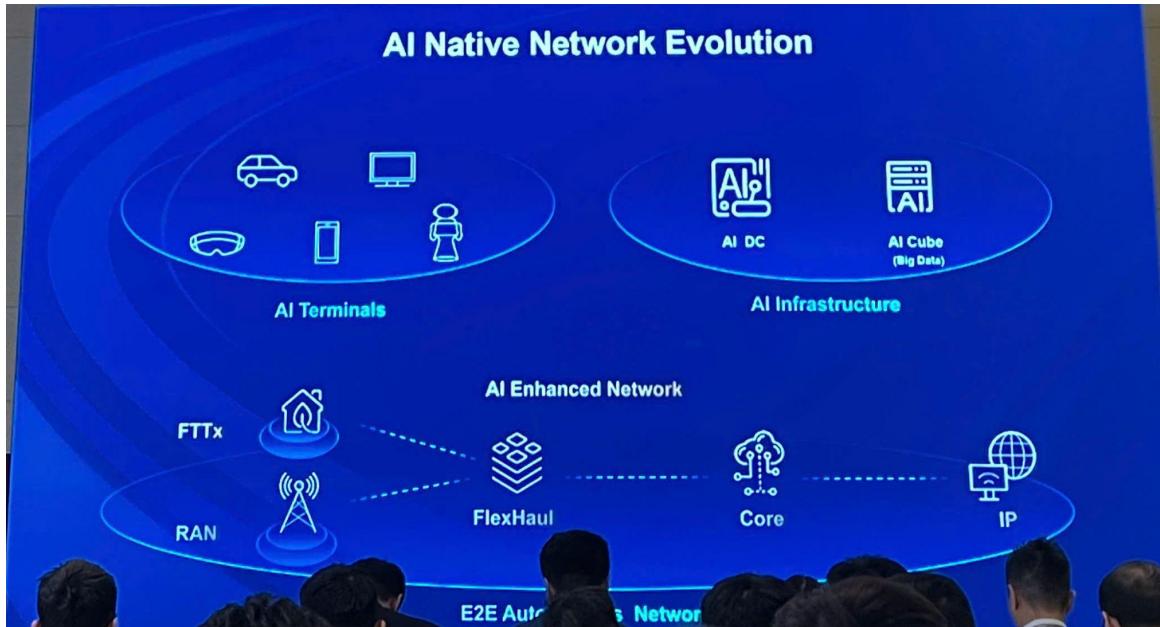
- ✓ More styling ideas: 100% more creative solutions
- ✓ Faster rendering: 30% improvement in design efficiency
- ✓ Cost reduction: Replacing graphics cards with intelligent computing





Xu hướng và ứng dụng AI cho Network & Operation

AI Native Network



Telco to Techco Cloud



Risk prediction

Proactive O&M to Improve Reliability and Reduce Optical Layer Risks

Ultra-Broadband **Smart**

Challenges

- Poor user experience: Services should be impacted if a fault occurs unexpectedly and then repair it
- High cost: Unexpected faults cause excessive repair costs

Solution

Algorithm: LSTM+ARMA

中国移动 @Xinjiang
@Zhejiang

```
graph TD
    DC[Data Collection] --> DP[Data Preprocessing]
    subgraph DP
        FE[Feature extraction]
        C[Classification]
    end
    DP --> ST[Sample Test]
    ST --> DS[Data Sample]
    DS --> AS[Algorithm Set]
    AS --> AO[Algorithm Optimization]
    AS --> BA[Best Algorithm]
    BA --> EF[Execution Forecast]
    EF --> PR[Prediction Result]
    PR --> AO
```

Highlights

- Ensure service quality: Warning risks more than 24 hours in advance, and pre-active O&M
- Reduce cost: Orderly network repair reduces costs by 20%

Result

- ✓ Provide a view of the health evaluation of whole optical fibers and optical channels

Statistics of fiber/Och

Status	Count
Normal	37
Sub-health	3
Warning	1

Health Score

- ✓ Analyze the trend of optical fiber health and predict optical fiber faults

AI-based Root Cause Analysis

AI-RCA Reduces Interference to Improves Fault location Efficiency

Ultra-Broadband

Challenges

- **Alarm interference:** Fault locating and troubleshooting are affected by a large number of derived alarms
- **Rules based on expert:** Fixed rules based on expert experience

Solution

Algorithm: Pearson Correlation Coefficient. Alarms location, correlative services and FA should be considered.

```
graph TD; A[History Alarms] --> B[Data Cleaning and clustering]; B --> C[Data modeling]; C --> D[Rule mining]; D --> E[AI-based rule library]; E --> F[Real Time Data]; F --> G[Rule mining AI-based]; G --> H[Compress Alarm]; H --> I[Root Alarm]; I --> J[New Rule]; J --> K[Rule Application AI-based]
```

The diagram illustrates the AI-RCA process flow. It starts with 'History Alarms' which undergo 'Data Cleaning and clustering', 'Data modeling', and 'Rule mining' to create an 'AI-based rule library'. This library is then applied to 'Real Time Data' via 'Rule mining AI-based' steps, leading to 'Compress Alarm', 'Root Alarm', and 'New Rule' generation. Finally, 'Rule Application AI-based' leads to the 'Root Alarm' and 'New Rule' stage.

Highlights

- **Depress derived alarms and locate root alarms:** More than 90% alarms should be depressed and show root alarms
- **Rules based on AI:** AI can find new rules and experts make decisions

Result

✓ improves O&M efficiency by 35%.

1. Alarm Monitoring and Statistics
2. Alarm compression
3. AI-Based rule mining

Category	Value
1. Alarm Monitoring and Statistics	35%
2. Alarm compression	27% 16%
3. AI-Based rule mining	Optimal Compression Ratio Find new rules

This section displays three main results of the AI-based rule mining process. Box 1 shows 'Alarm Monitoring and Statistics' with a 35% improvement. Box 2 shows 'Alarm compression' with 27% and 16% reductions. Box 3 shows 'AI-Based rule mining' with 'Optimal Compression Ratio' and 'Find new rules'.

Category	Value
Depressing alarms	40 K+
Depressing alarm Work-order	800
Depressing 88.5% alarms	4,323 items
Depressing 88.5% alarms	804 items
Depressing 88.5% alarms	40,203 items

This summary table provides a high-level overview of the AI-RCA results. It includes metrics for depressing alarms (40K+), depressing alarm work-orders (800), depressing 88.5% alarms (4,323 items, 804 items, 40,203 items), and depressing 88.5% alarms (40,203 items).

中国移动 @Xiniiang

@Zhejia

nan @114-1