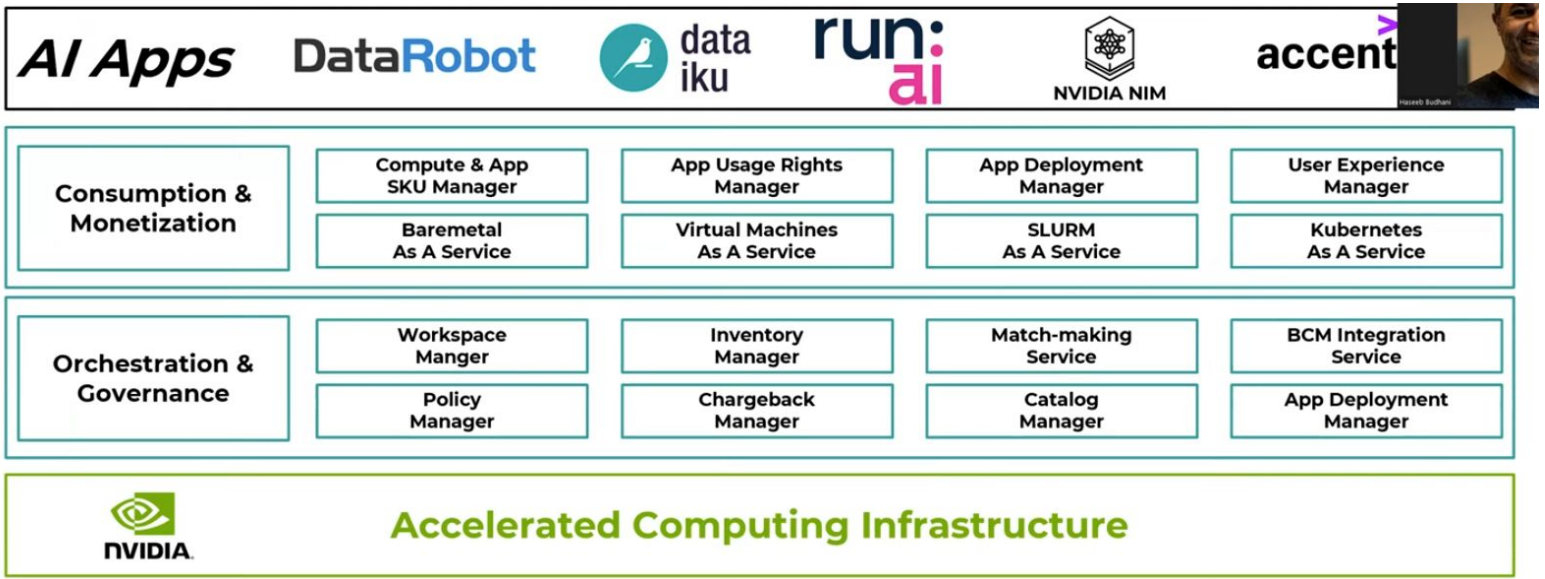

RAFAY

Rafay

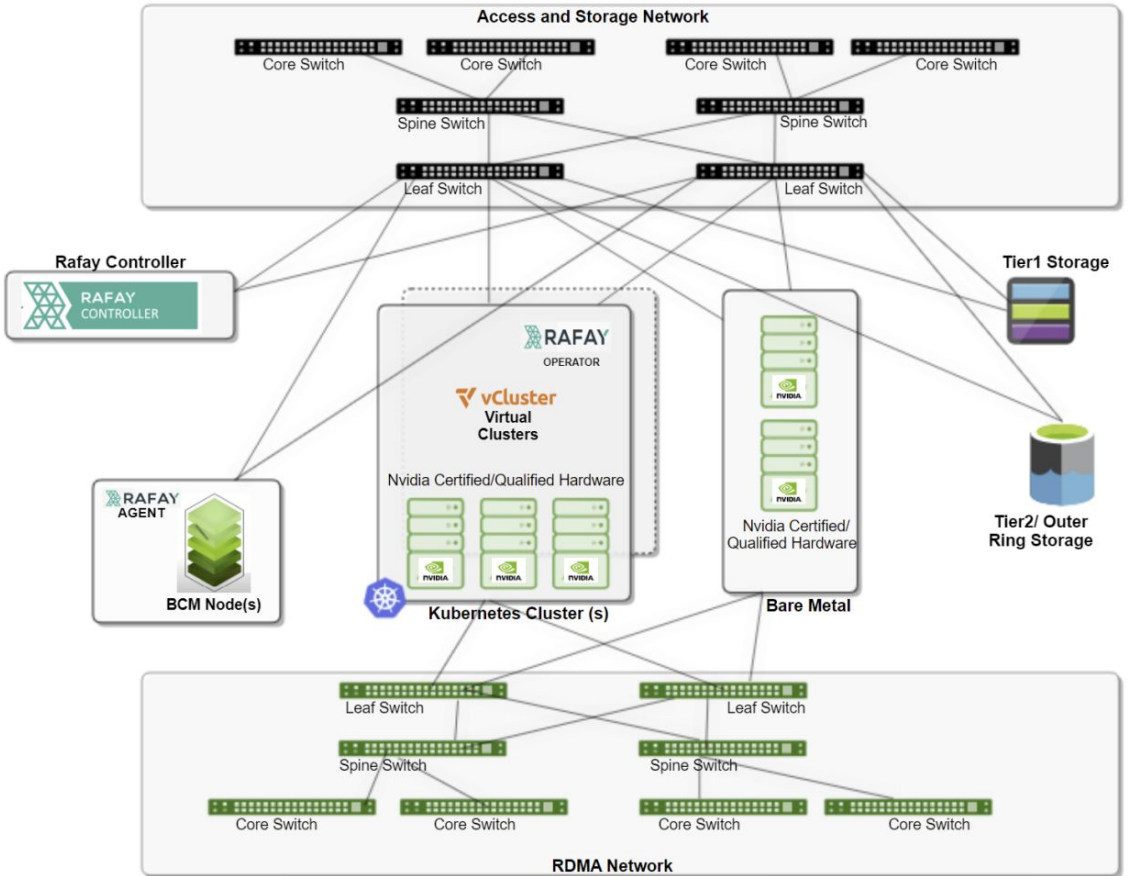
1 Rafay cung cấp dịch vụ K8s management. Rafay sẽ cài K8s trên top của BCM, không dùng K8s mặc định mà BCM cung cấp ⇒ Rafay sẽ nhúng K8s vào BCM. K8s management của Rafay có khả năng tích hợp với NIM, Run:AI, hoặc ứng dụng open source như Kubeflow, Jupiter notebook, …

Rafay


AI Apps
DataRobot
data iku
run:ai
NVIDIA NIM
accent
Consumption & Monetization
Compute & App SKU Manager
App Usage Rights Manager
App Deployment Manager
User Experience Manager
Baremetal As A Service
Virtual Machines As A Service
SLURM As A Service
Kubernetes As A Service
Orchestration & Governance
Workspace Manger
Inventory Manager
Match-making Service
BCM Integration Service
Policy Manager
Chargeback Manager
Catalog Manager
App Deployment Manager
NVIDIA
Accelerated Computing Infrastructure

Rafay

NVIDIA-Certified System Hardware and Network Topology



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

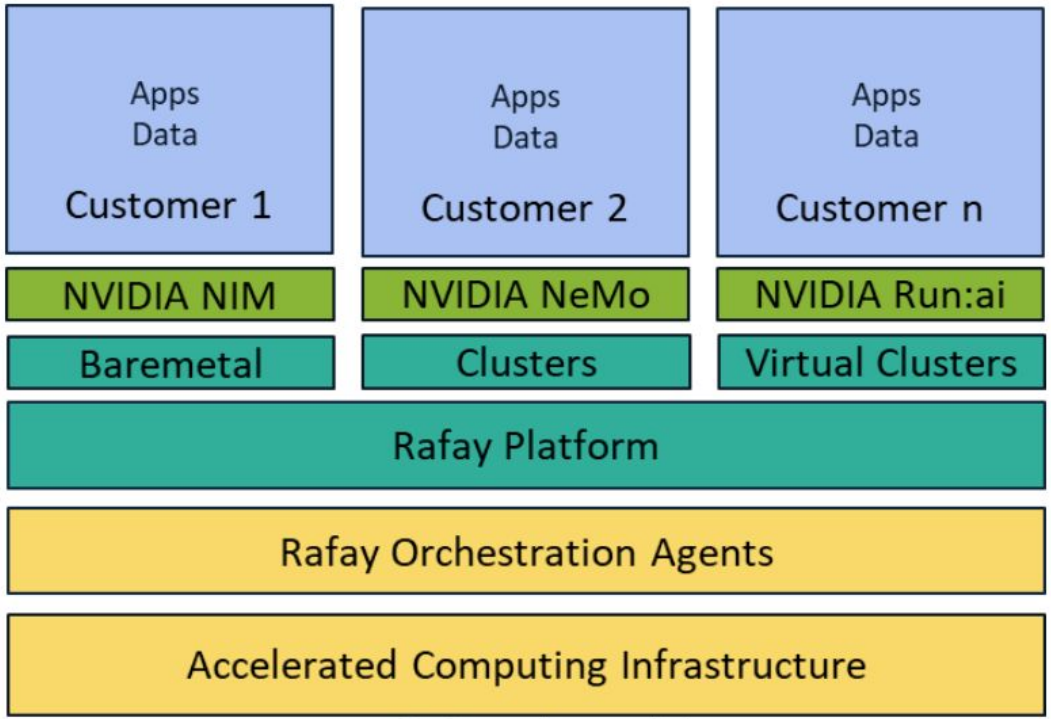
Rafay

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*).

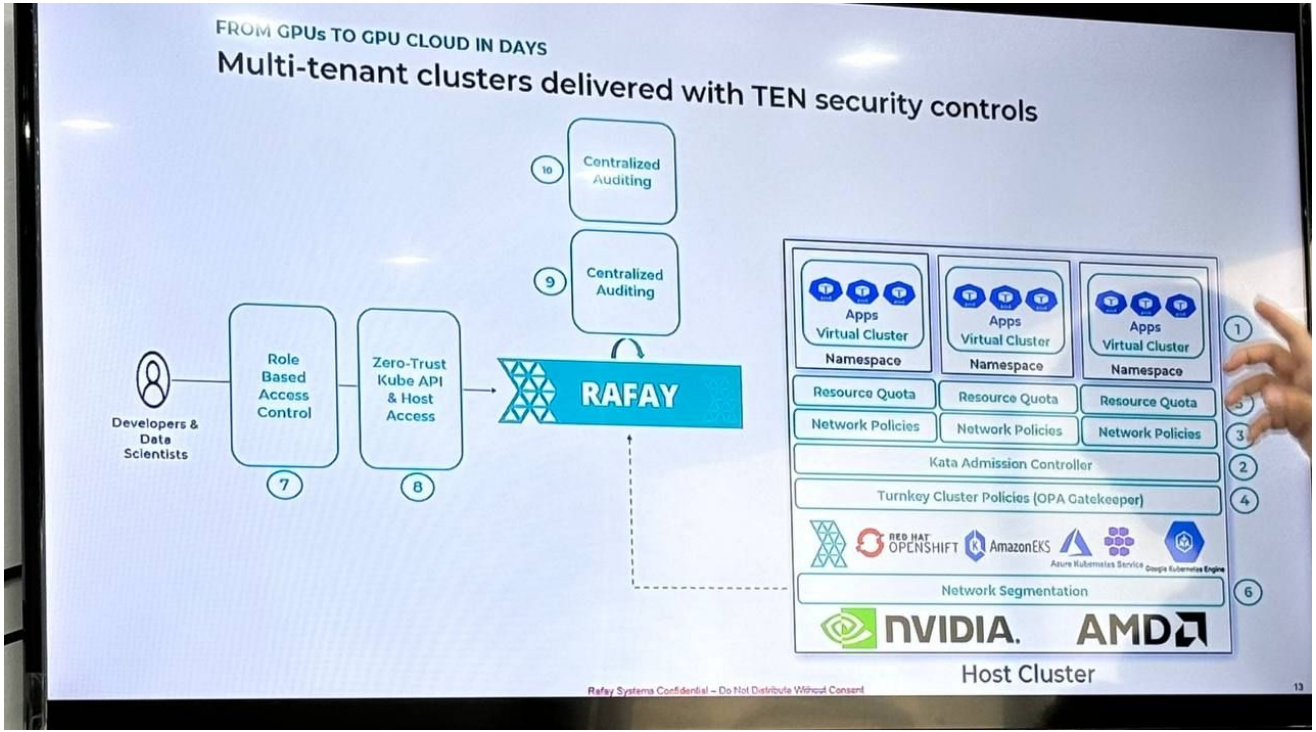
> Rafay có thể cung cấp giải pháp chạy nội bộ toàn bộ hệ thống trong private environment. Có thể chạy Rafay controller ở Data center nội bộ của công ty

5 Rafay có sử dụng MIG để chia GPU vào các pod

Rafay Multi-tenancy Architecture



Rafay Multi-tenancy with Virtual Cluster



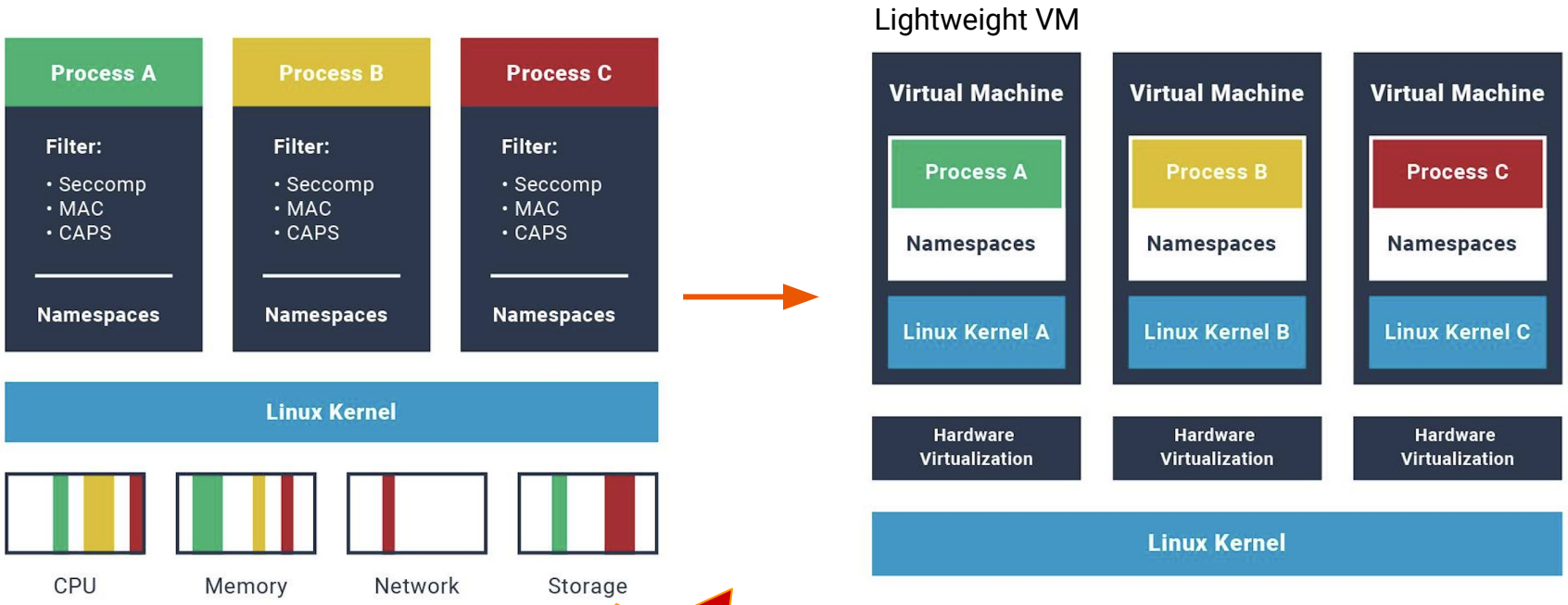
Rafay Multi-tenancy with Virtual Cluster

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



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

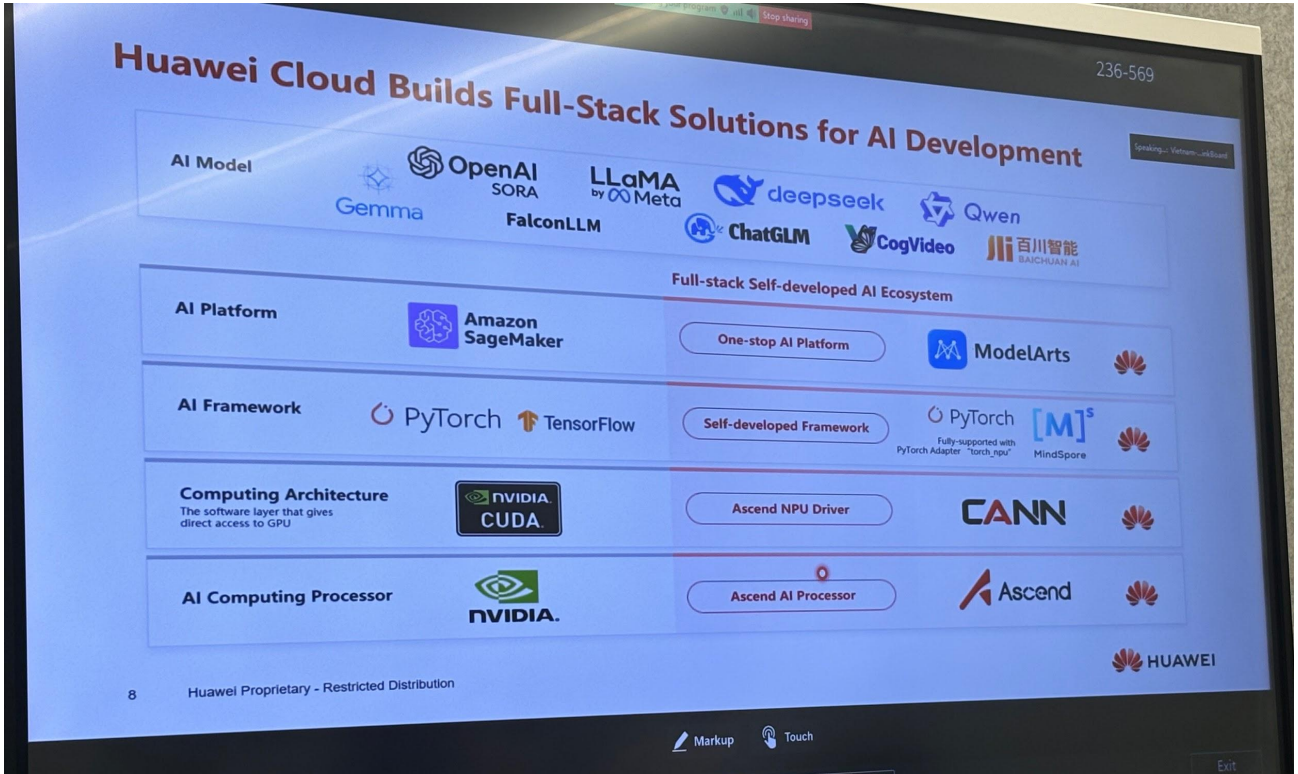
1. [Rafay Reefference Architecture]
   https://rafay.co/wp-content/uploads/2025/02/Rafay-Reference-Architecture_Feb25-Final.pdf
2. [Rafay Platform Multi Tenancy Controls]
   https://rafay.co/wp-content/uploads/2024/11/Rafay-Platform-Multi-Tenancy-Controls.pdf
3. [Rafay Platform Environment Manager Developer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


HUAWEI

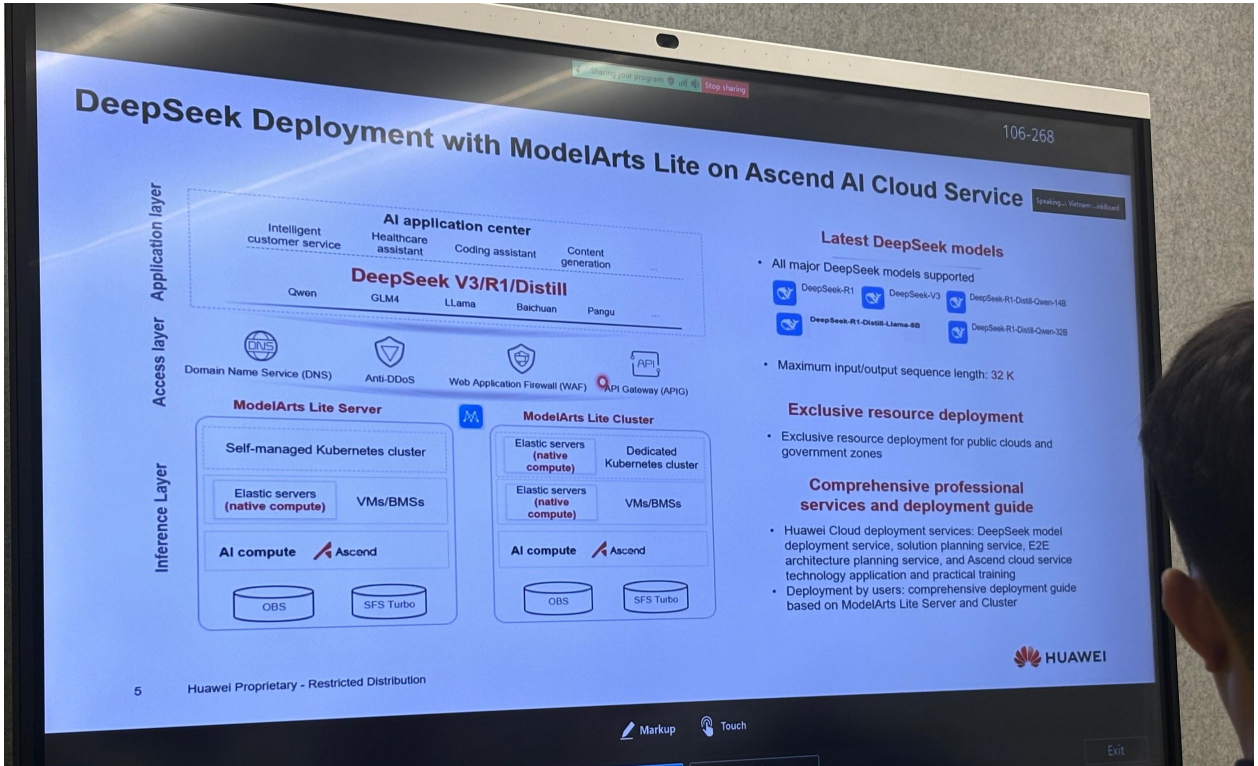
NPU - Neural Processing Unit

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

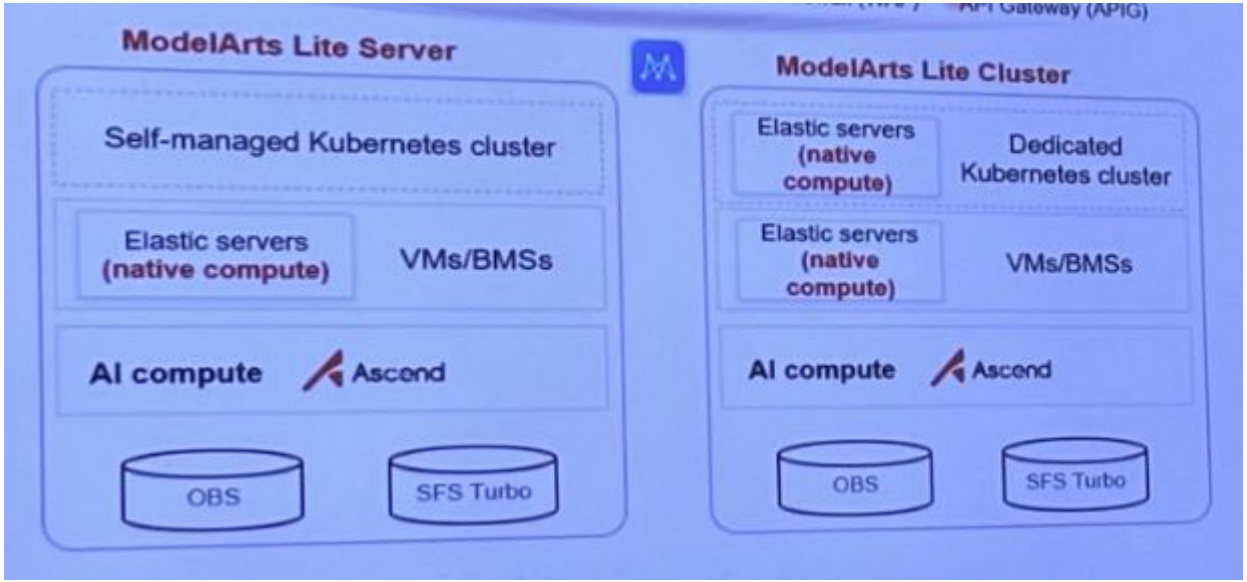
Huawei AI Full-Stack



Huawei triển khai ModelsArt



Huawei triển khai ModelsArt



ModelsArt 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.”*

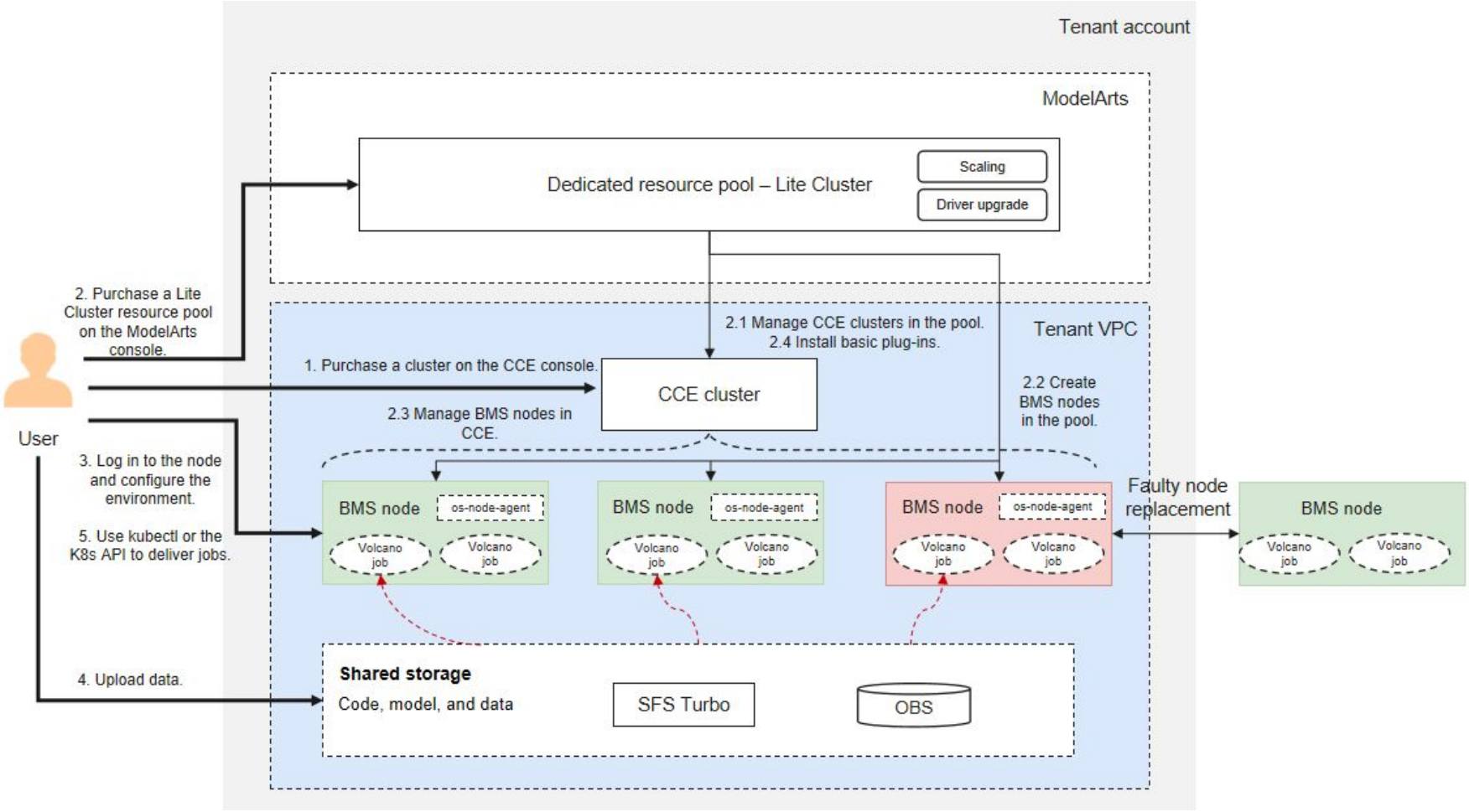
It is designed for users who have built their own AI development platforms and require only computing power. It provides cost-effective AI computing power, mainstream AI development suites, and Huawei's acceleration plug-ins

ModelArts Lite Server offers cloud servers running on bare metal servers, accessible via EIPs.

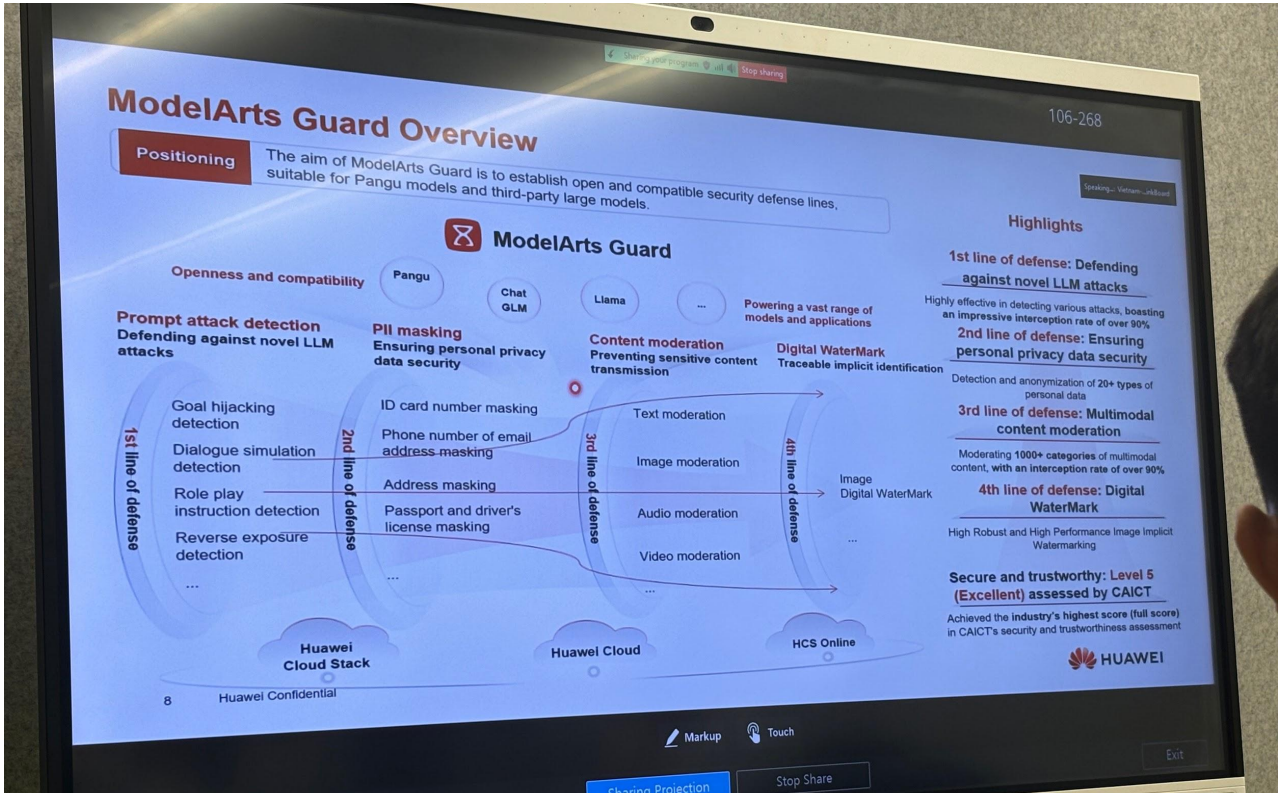
ModelsArt Lite Cluster

“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

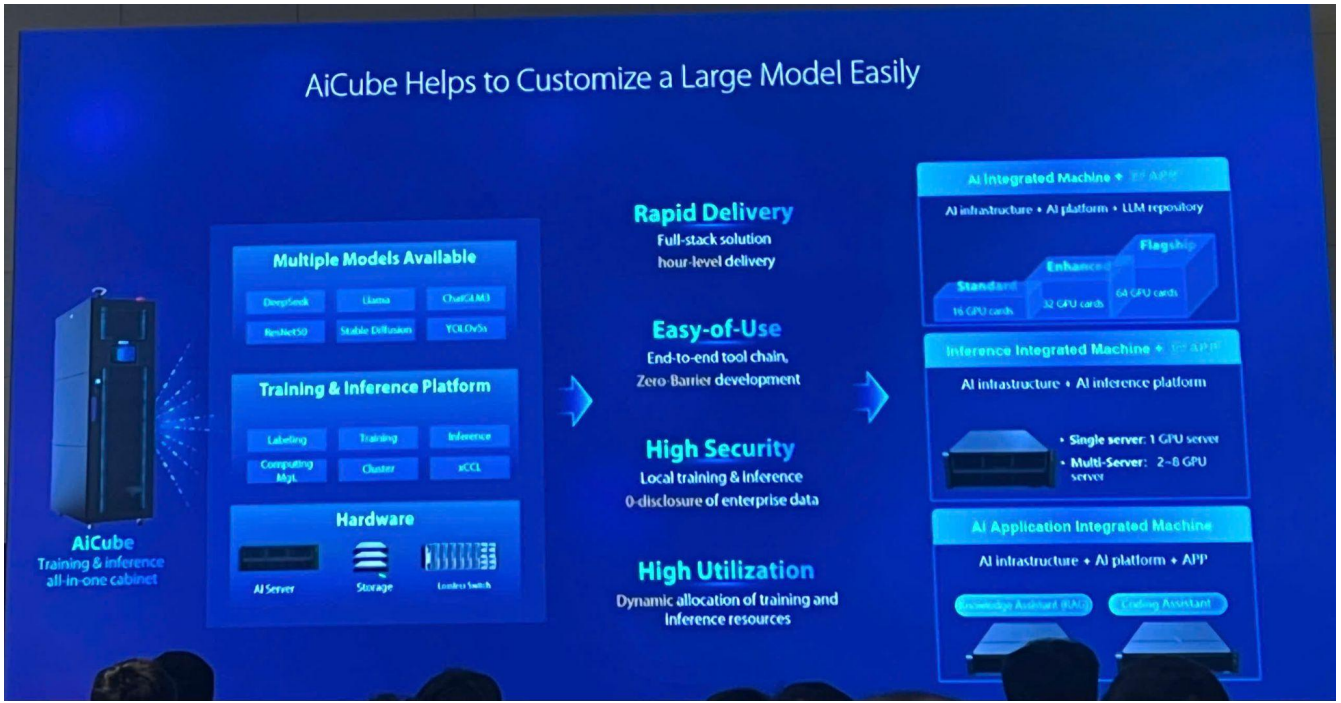



ZTE

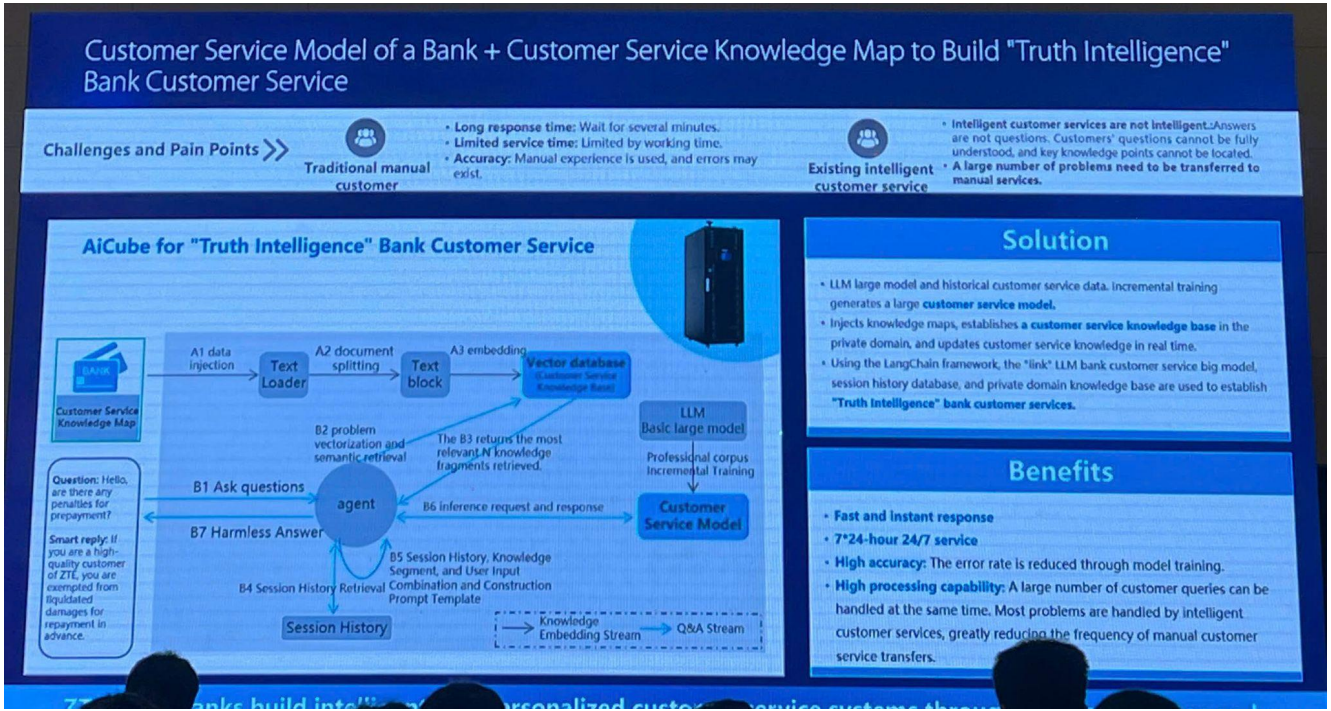
Nội dung chính

- 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

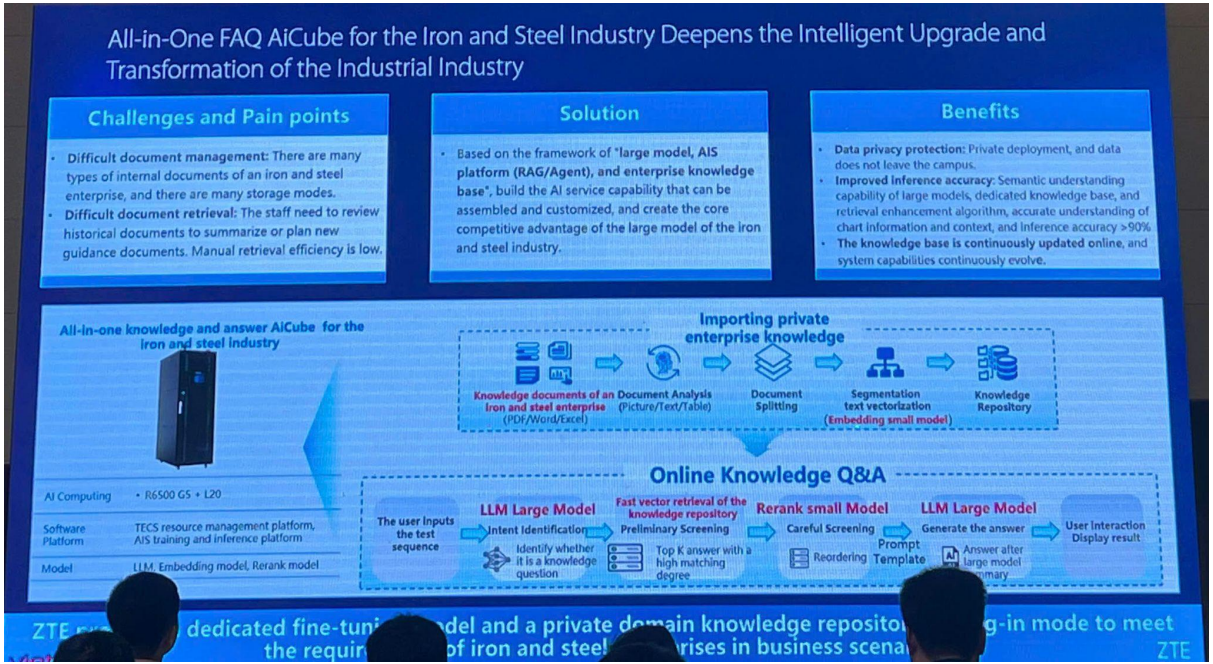
AI Cube


AiCube Helps to Customize a Large Model Easily
AiCube
Training & Inference
all-in-one cabinet
Multiple Models Available
DeepSeek
Llama
ChatGLM3
ResNet50
Stable Diffusion
YOLOv5s
Training & Inference Platform
Labeling
Training
Inference
Computing Mgt.
Cluster
xCCL
Hardware
AI Server
Storage
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
AI Integrated Machine + APP
AI infrastructure + AI platform + LLM repository
Standard
Enhanced
Flagship
16 GPU cards
32 GPU cards
64 GPU cards
Inference Integrated Machine + APP
AI infrastructure + AI inference platform
• Single server: 1 GPU server
• Multi-Server: 2~8 GPU server
AI Application Integrated Machine
AI infrastructure + AI platform + APP
Coding Assistant

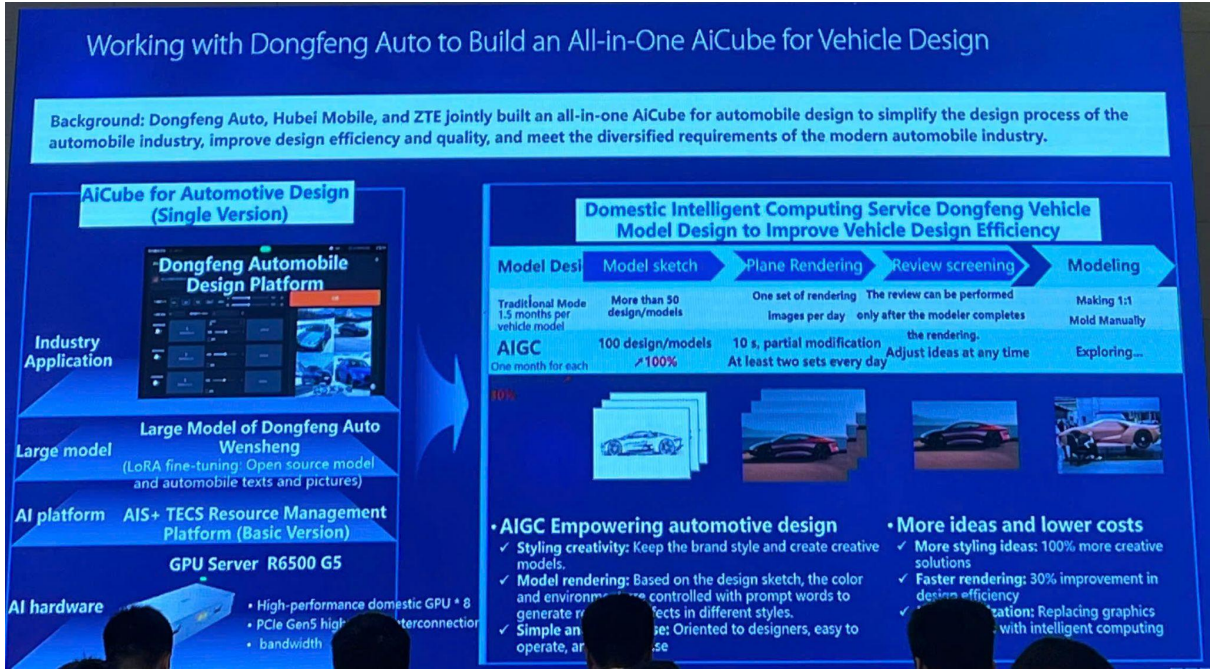
Case study for Intelligence Bank Customer Service



Case study for Iron and Steel Industry

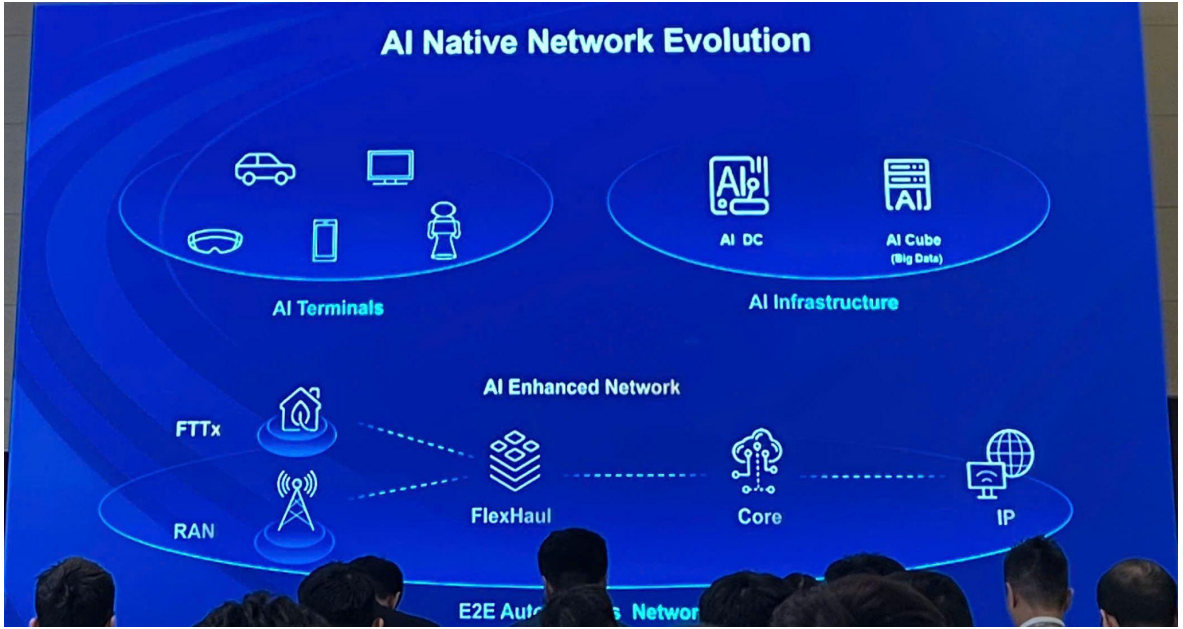


Case study for Automotive Industry



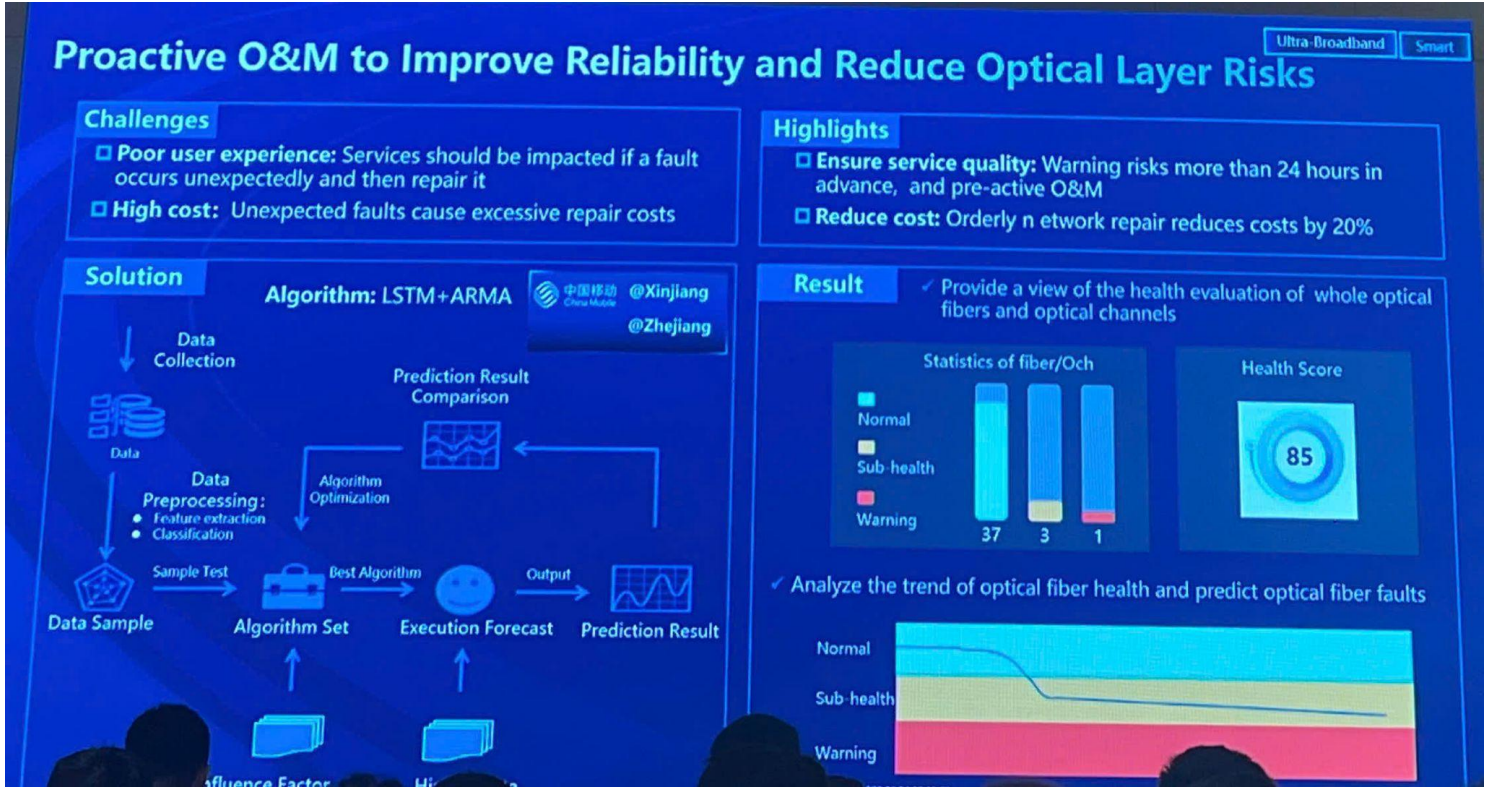
Xu hướng và ứng dụng AI cho Network & Operation

AI Native Network



Telco to Techco Cloud

Risk prediction



AI-based Root Cause Analysis