China Mobile perspectives and practices on cloud native and AI native Evolution of Telco Cloud

China Mobile Research Institute

June, 2026

Current problems and challenges in NFV/SDN

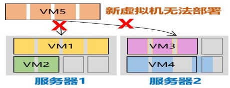
After more than a decade of development, NFV/SDN has successfully achieved large-scale commercial deployment. It is now entering a new phase focused on improving quality and efficiency. There is still room for optimization in areas such as network agility, resource utilization, and the intelligentization of OAM. These challenges need to be gradually addressed by introducing new concepts, technologies, and operational models.

Network Agility

- **Tightly coupled architecture:** Stateful modules of NF pose risks during upgrades or migration,limiting dynamic network maintenance.
- **Weak innovation support:** Some ToB business require agile innovation, but platforms lack of modular services, hindering fast development and deployment.
- **Slow scaling response:** Long preparation time for demands arising from holidys, events and sudden bsiness suges, hard to response in real time.

Resource Utilization

- **Resource fragmentation:** NFs have multiple VM specifications, and parts are of large granularity, which leads to resource fragments.



- **Resource scheduling imbalance:** Different VM workloads have distinct resource requirements (e.g., CPU, network), unbalanced scheduling of a single resource type lead to resource allocation bottlenecks.

- **Underused AI computing resource:** Due to insufficient virtualization slicing and fine-grained scheduling capabilities in AI accelerators from vendors, AI computing resources can’ t be fully utilized.

Intelligentization of OAM

- **Complex tasks relies on human labor:** Manual alarm handling, root-cause analysis, and service provisioning are slow and costly. With fast growing network complexity and number of devices, relying on manpower is no logner sustainable.
- **Constraints on CI/CT/CD Application Scenarios:** Current automation capability is limited to simple and well-defined application scenarios. Once any changes occur in the environment, human intervention becomes necessary, and the efficiency advantage becomes insignificant.

Future network service trends

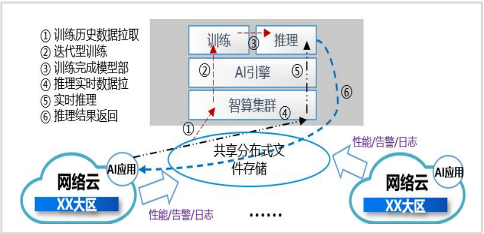
Intelligence will be a defining feature of future 6G networks. Regardless of whether it is AI for network or network for AI scenarios, a holistic layout for computility deployment—particularly AI computing resources—is required to achieve globally optimal delivery of computing services.

AI for Network

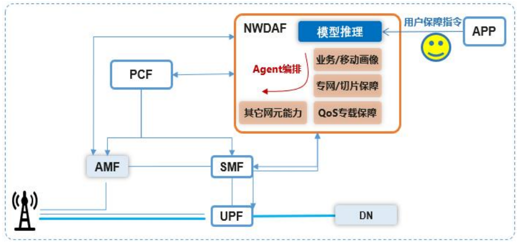
➢ Driven by growing network complexity, escalating performance demands, and a rising number of customized service scenarios, telco operators must harness AI computility & services to boost operational efficiency, lower costs, and elevate the user experience.

Typical use cases

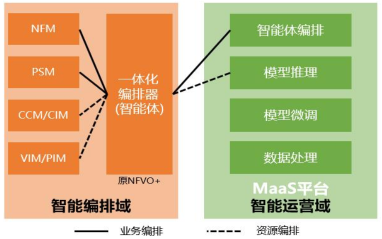
AI-Driven OAM



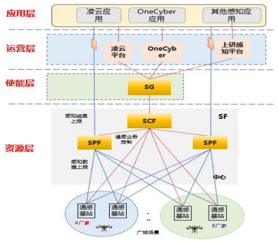
AI-Driven Operation



AI-Driven Orchestration



Low-altitude monitoring



New Calling AI Services

AI速记 智能翻译/通话字幕 数字人

AI4N places distinct requirements on AI computility of telco clouds. Business growth of network O&M agents and New Calling AI services drives higher inference computing needs.

Network for AI

➢ With AI evolving towards the era of AGI, the proliferation of new services and end devices has created a demand for AI access anytime and anywhere. This necessitates the support of ubiquitous intelligent computing power, network connections, and data.

Typical use cases

Embodied AI Home Robots

Embodied AI Industrial Robots

Intelligent Connected Vehicles

Low-altitude Economy

AI Glasses/XR Devices

Multi-Agent Collaboration

NET4AI introduces potential requirements for cloud-edge-end collaboration, with network latency emerging as a critical factor in deciding where to deploy computing resources.

Future network technology trends

Driven by the rapid development of ICT, the virtualization technologies, Operation & Maintenance mode, and Orchestration & Management methods of current NFV/SDN are undergoing profound changes.

Infra from general-purpose computing to AI computing

Computing	CPU centralized	Integrated development of general and AI computility
Storage	Block storage	File and object storage, multi-protocol storage integration
Network	Legacy DC network	High-speed DC network with ultra-low packet loss rate

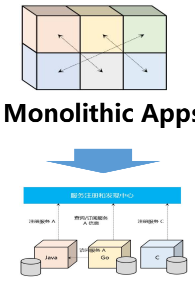
Network from simple virtualization to cloud native

Cloud Platform

Virtual Machine

Container

NF Architecture



Monolithic Apps

Microservice based

Process Framwork

Manpower

Automation

Technology development

Network O&M from automation to Intelligence

Automation

Perform tasks based on preset fixed rules or workflows, only applicable to scenarios where both process and outcome are highly predictable.

Development and debugging primarily relying on specialized automation tools and pipelines.

Automated		
Integrate	Config	Test

Intelligence

Featuring deep autonomous learning and dynamic decision-making to manage complex, ever-changing, and unpredictable tasks.

Agent-centric, achieving autonomous decision-making and automated execution for complex O&M tasks.

Deep learning; NLP; Model inference; Tool calling

Service provisioning; Fault analysis; Network optimization

Orchestration from resources to AI capabilities

Resource Orchestration

Based on pre-configured VNFD and NSD files, providing basic orchestration services for generic computing and network connectivity resources for NF

MANO

APP / VM / CPU — APP / VM / CPU

Network resource

Integrated Orches of resource & AI capabilities

Driven by agents as primary interface, it works in tandem with orchestration systems to manage and coordinate AI capabilities (including computing, network resources, and model services) enabling E2E deployment and LCM of emerging agent-based applications.

Agent — Orchestration & Operation

APP / Container / CPU — Agent app/AI model / Container / CPU, GPU

Network resource

Core demand on Telco Cloud infra driven by development trends

5G+ 中国移动 China Mobile

Network development trends have posed explicit requirements on Telco Cloud infrastructure in terms of both technical capabilities and resource deployment.

Full-stack intelligence

Intelligence is a core network requirement for future telecom services and a key enabler for the self-evolution of Telco Cloud infra:

- Build a converged computing-network foundation to drive the leap from general comoputing to AI computing.
- Open AI capabilities (e.g., model inference & fine-tuning) to diverse services and AI agents.
- As networks grow complex and devices multiply, intelligent O&M replaces manual management to cut costs and boost efficiency.

Network-wide Agility

Agile telecom cloud-native tech and mgt. systems are essential for flexible network adjustments and business innovation:

- Containerized infra provides finer-grained resource management, scheduling capabilities, and faster response speeds.
- Bare metal containers, as the de facto standard for AI-era infra, are highly suited for the intelligent evolution of the core network.
- Stateless and microservices designs enable agile releases like canary upgrades, comprehensively boosting service reliability.

Ubiquitous Deployment

Diverse services demand varying compute and latency, requiring a balanced deployment of computing & network resources based on performance, cost, and location:

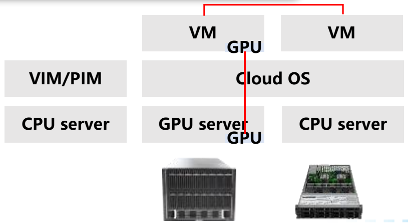
- Resources must be deployed locally for extreme low-latency or data-sovereignty scenarios.
- Edge nodes near end devices can be leveraged for end-edge collaboration, ensuring latency SLAs.
- Edge provides network connectivity on demand for device-to-device and device-to-internet scenarios.

Integrated general & AI computing: A converged foundation

Driven by the synergy of general and AI computing, the 6G era will usher in a new computing and sotrage architecture featuring the deep integration and unified deployment of communication and intelligence.

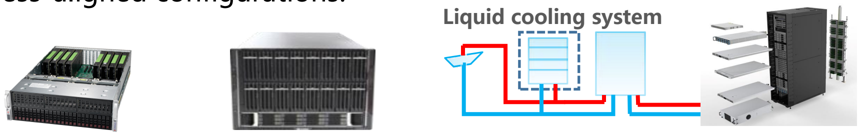
Integrated general & AI computing

- Leveraging a unified resource pool, a unified scheduling platform, and a unified O&M system to achieve E2E visibility and manageability of the next-generation resource pool.



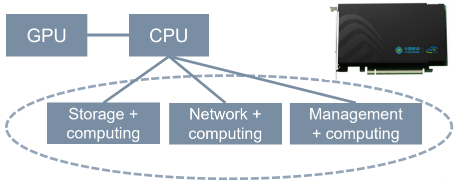
AI computing devices

- Evolving from 8-GPU servers to 16-GPU/64-GPU supernodes to deliver diverse form factors and high-performance computing, enabling fine-grained, business-aligned configurations.



Data-centric

- Breaking CPU boundaries to move computing closer to data, leverage DPU/SNIC to empower 6G infra.



High Density & Efficiency

- Doubling processor core density and network I/O performance, alongside upgraded memory BW.

Advance Key Storage Technologies Tailored for AI

Tailored to the intelligent business characteristics of future network, this initiative focuses on the critical data for AI training and inference. It involves in-depth research into core technologies such as hot/cold data tiering, protocol convergence, parallel clients, xDS, KV interfaces, and multi-tier KVCache storage.

Inference Data			Training Data		
Moel Registry	KVCache	RAG Vector Database	Data Set	Model Registry	Checkpoint

Upgrade Legacy NFV Storage Solutions

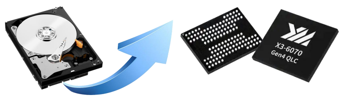
Cloud-native evolution drives the demand for shared storage

VM + C

Develop smart storage O&M capabilities, and strengthen fault prediction, diagnostics, and automated remediation.

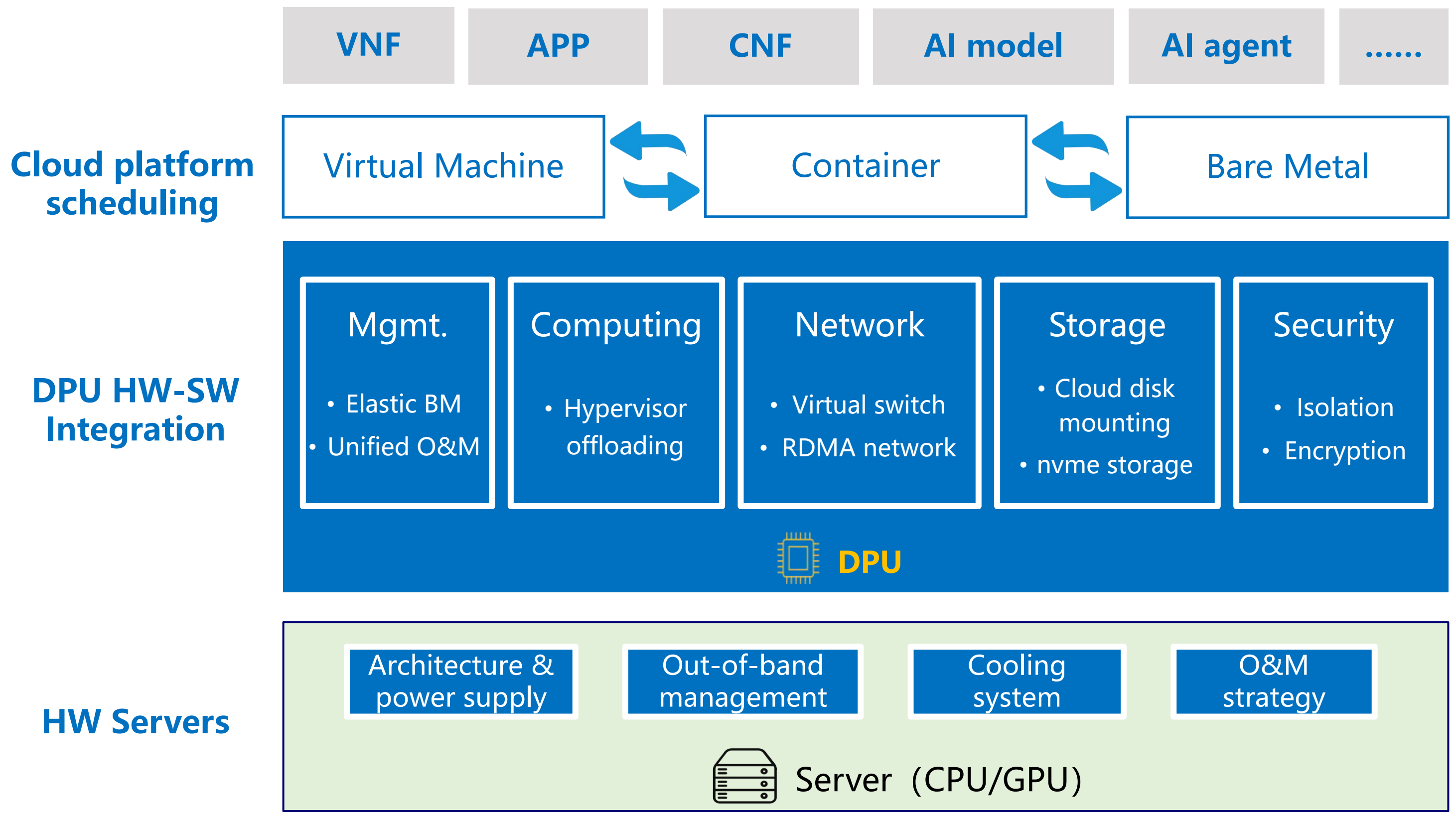
Ongoing Evolution of Storage Media

Accelerate the transition to all-flash storage architectures. Build performance-tier storage with TLC SSDs, and introduce QLC SSDs to accommodate large-capacity warm and cold data.



Integrated general & AI computing: DPU-driven unified computing foundation

Powered by server hardware, DPU hardware-software synergy, and cloud platform scheduling, the DPU computing base integrates five core systems: compute, network, storage, management, and security. It delivers advanced capabilities—including elastic bare metal, high-speed lossless networking, and high-performance storage—to provide unified hosting and management across virtual machines, containers, and bare metal.



- **Empower diverse application scenarios:** Powered by DPU, this platform features a unified dual-stack architecture for both containers and VMs. Designed to support cloud native and AI native NFs and apps, while seamlessly accommodating legacy VM and BM workloads.

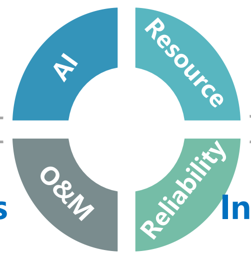
- **Dynamic Adjustment of Computing Forms:** Powered by a unified DPU computing base, Hypervisor, CIE, or Bare Metal nodes can be flexibly provisioned on demand. It also facilitates dynamic transitions among these three forms, maximizing hardware resource sharing and operational efficiency.

Cloud-native network functions: Values and challenges

Driven by the continuous shift of 5G-A and 6G networks towards intelligent and agentic paradigms, the need for agile innovation and rapid capability iteration is more critical than ever. CNFs can enable atomic service decoupling, elastic scaling, and fine-grained resource scheduling, which are essential for meeting extreme demands of autonomous networks—specifically on-demand generation and dynamic orchestration—while simultaneously driving further improvements in utilization efficiency and reliability.

As the de facto infrastructure standard in the AI era, Bare Metal Container technology is ideally positioned to drive the future AI-native evolution of cloudified core networks.

Compared to virtual machines, BM containers combined with DPU incur significantly lower resource overhead, thereby improving the resource utilization of compute servers.



CNFs offer superior concurrency during upgrade operations compared to VNFs, thereby improving the efficiency of software upgrades for NFs in live network.

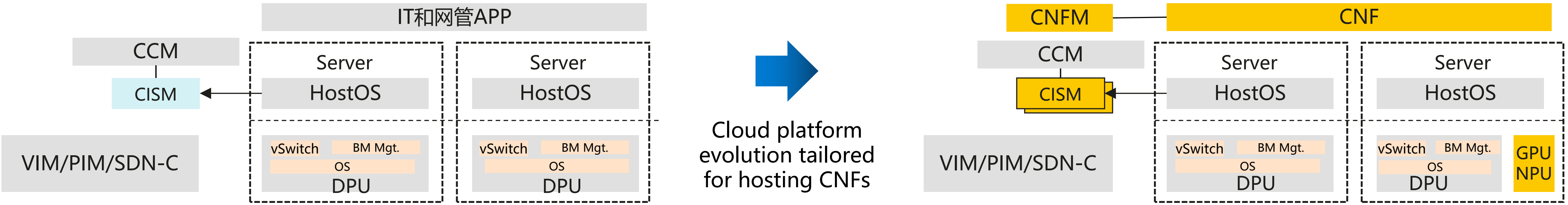
In disaster recovery failovers, CNF architecture enables rapid provisioning of network element instances, mitigating signaling surges and guaranteeing smooth service restoration.

Integrating cloud-native capabilities into telecom service network elements is a highly complex systemic undertaking. It demands a fundamental architectural overhaul and upgrade of the core functional modules within existing NFV frameworks.

Cloud Platform	Network Function	Orchestration	Security & Isolation
Carrier-grade container infrastructure	Microservices architecture design	Independent orchestration architecture for CNFs	Comprehensive Security Hardening

Cloud-native network functions: Cloud infrastructure

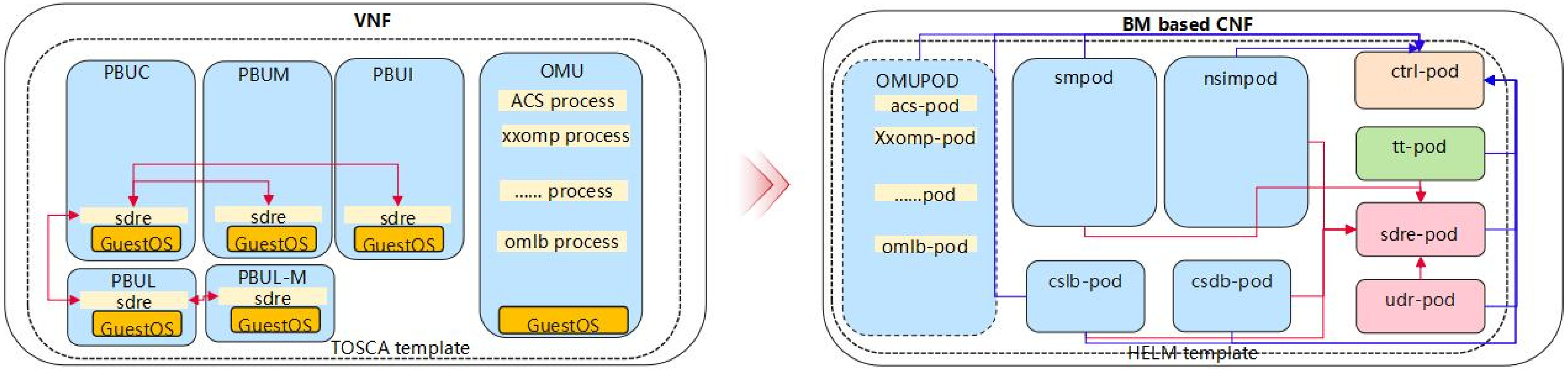
For future large-scale deployment scenarios of CT-oriented CNFs in 6G, cloud infrastructures require systematic capability enhancements and optimizations to meet the stringent performance and reliability requirements of telecom services.



Dimensions	Key capabilities
Performance	• **Dynamic Service Network Deployment:** CNFs possess far more intricate network topologies than generic IT container applications, characterized by a massive number of network planes and ports. This complexity renders large-scale pre-provisioning unfeasible. Consequently, a dynamic network orchestration mechanism is essential to synchronously provision and reclaim resources on demand in alignment with the CNF lifecycle.
	• **Hardware Passthrough Forwarding Mechanism:** Certain CNFs demand ultra-high forwarding throughput and strong network isolation. Traditional open-source CNIs, heavily reliant on software, often hit performance ceilings. Therefor, HW-SW co-design technologies like DPU passthrough can be deployed to scale network device density and maximize forwarding performance.
Reliability	• **Storage Bypass:** In BM container deployments, a synergistic approach integrating the container layer, servers, DPUs, and backend storage is required to establish robust Storage Bypass capabilities for CNFs.
O&M	• **General O&M:** By leveraging and enhancing open-source de facto standards like Prometheus and OpenTelemetry, establish standardized, cloud-native observability and manageability capabilities specifically designed for CNFs.
Service Mgt.	• **Service Management:** By integrating CNF microservices strategies, it is possible to build public service capabilities tailored to CNF on demand, utilizing K8S Service/Ingress, CoreDNS, and Service Mesh. This approach effectively reduces development workload and improves resource utilization efficiency.

Cloud-native network functions: NF architecture

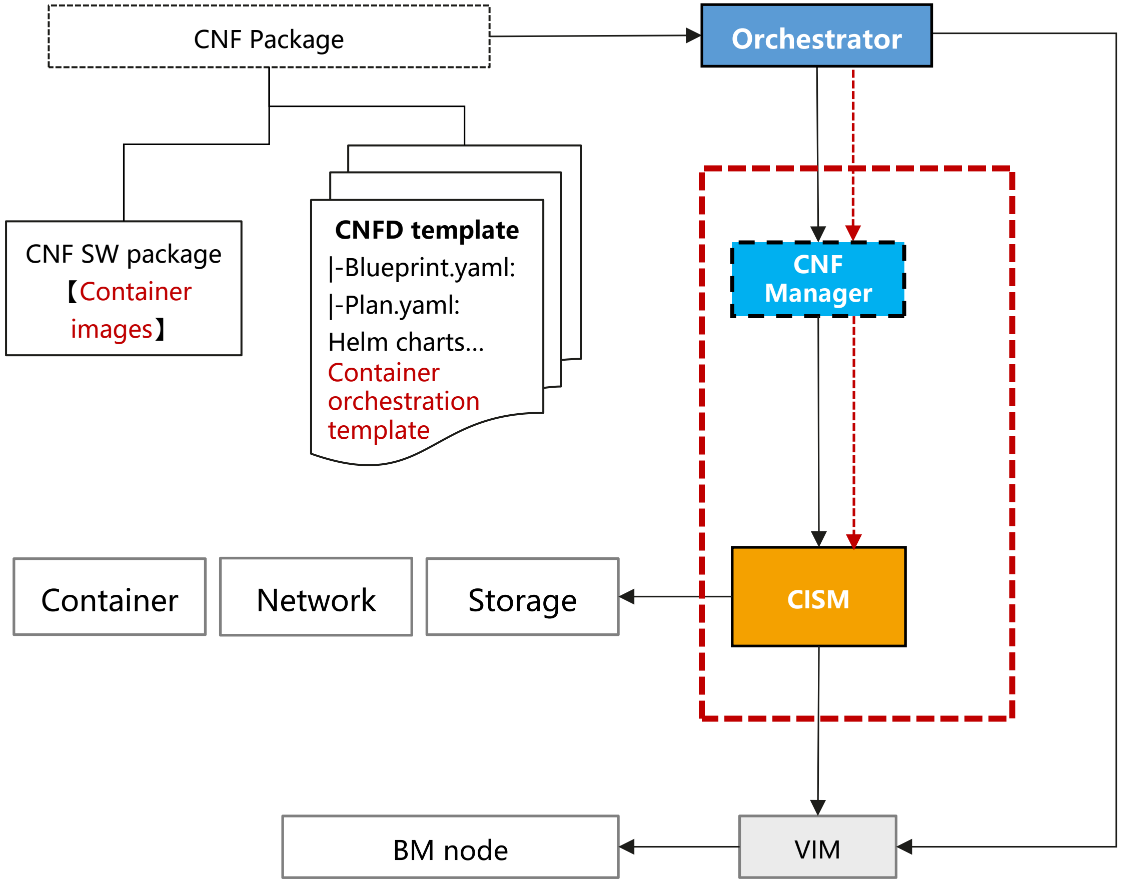
As 5G-A and 6G networks evolve towards cloud-native, in addition to shifting resource deployment from VM to BM containers, network functions need to reconstruct their business logic based on microservices. This involves modifications to functional sub-modules, templates, and scheduling models, as well as the introduction of new capabilities for container and microservice management.



Dimensions	Key capabilities
LCM Architecture	• **Reconstructing the LCM architecture:** Utilizing the combination of Operator and Helm to achieve deployment, upgrading, scaling out/in, and other lifecycle operations.
Service Architecture	• **Service Scheduling Model:** Extracting shared business capabilities (e.g., UDR, SBI, SDRE) to drive a container-centric microservice scheduling paradigm.
Communication Architecture	• **Tier-1 Communication Architecture:** Restructuring the microservice communication topology with an externalized service bus and distributed POD-to-POD forwarding, thereby empowering seamless horizontal scaling.

Cloud-native network functions: Management & orchestration

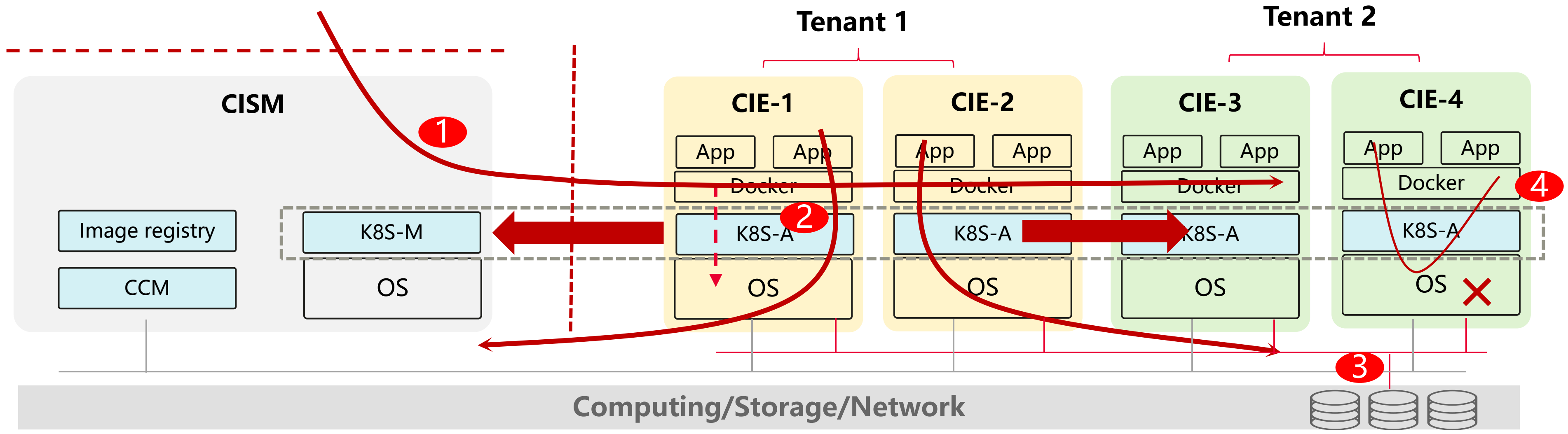
BM container-based CNFs and VM-based VNFs differ significantly in resource forms and resource description templates, and rely on entirely different technology stacks. Evolving by extending existing VNF package management, resource orchestration templates, and interfaces is detrimental to leveraging the advantages of new technologies. Therefore, it is recommended to build a new, more concise and efficient CNF management and orchestration architecture.



Dimensions	Key Capabilities
Orchestration	• **Cluster Orchestration:** MANO supports the provisioning of BM container clusters for CNFs, re-architecting the tenant models of the Orchestrator, VIM, and CISM.
	• **CNF Orchestration:** The Orchestrator and CNF Manager support the CNF orchestration, lifecycle management, package management, and container template parsing of NF.
	• **Monitoring:** The Orchestrator enhances performance in container management and monitoring.
O&M	• **Management Objects:** Shifting from VMs to BM containers exponentially increases both the volume of managed entities and the complexity of operations.

Cloud-native network functions: Security isolation

The shared host OS architecture of BM containers inherently compromises multi-tenant isolation. To address this, It is necessary to fortify platform-level capabilities—including escape detection, affinity scheduling, and DPU offloading—and align them with optimized cluster planning to proactively minimize security risks.



Issue	1: Control plane attack	2: Risk of container escape	3: Risk of storage network	4: OS Fault Propagation
Description	Under a shared cluster model, a successful breach of the control plane exposes all tenant CIE nodes to potential attacks.	BM containers escape to the host, attack other apps on the same node, and further compromise the control plane and other tenants' CIE nodes.	After escaping to the host, attackers can exploit the storage network to attack other nodes and remote storage systems.	Due to shared OS kernel architecture, a kernel crash triggered by SW bugs or malicious attacks can compromise all co-located containers.
Actions	Security domain partitioning and container cluster isolation	Container security hardening, container escape detection, and container cluster isolation	DPU bare metal (The storage network is invisible to tenants)	Container security hardening and host anti-affinity strategies

Network Intelligence: MaaS Platform

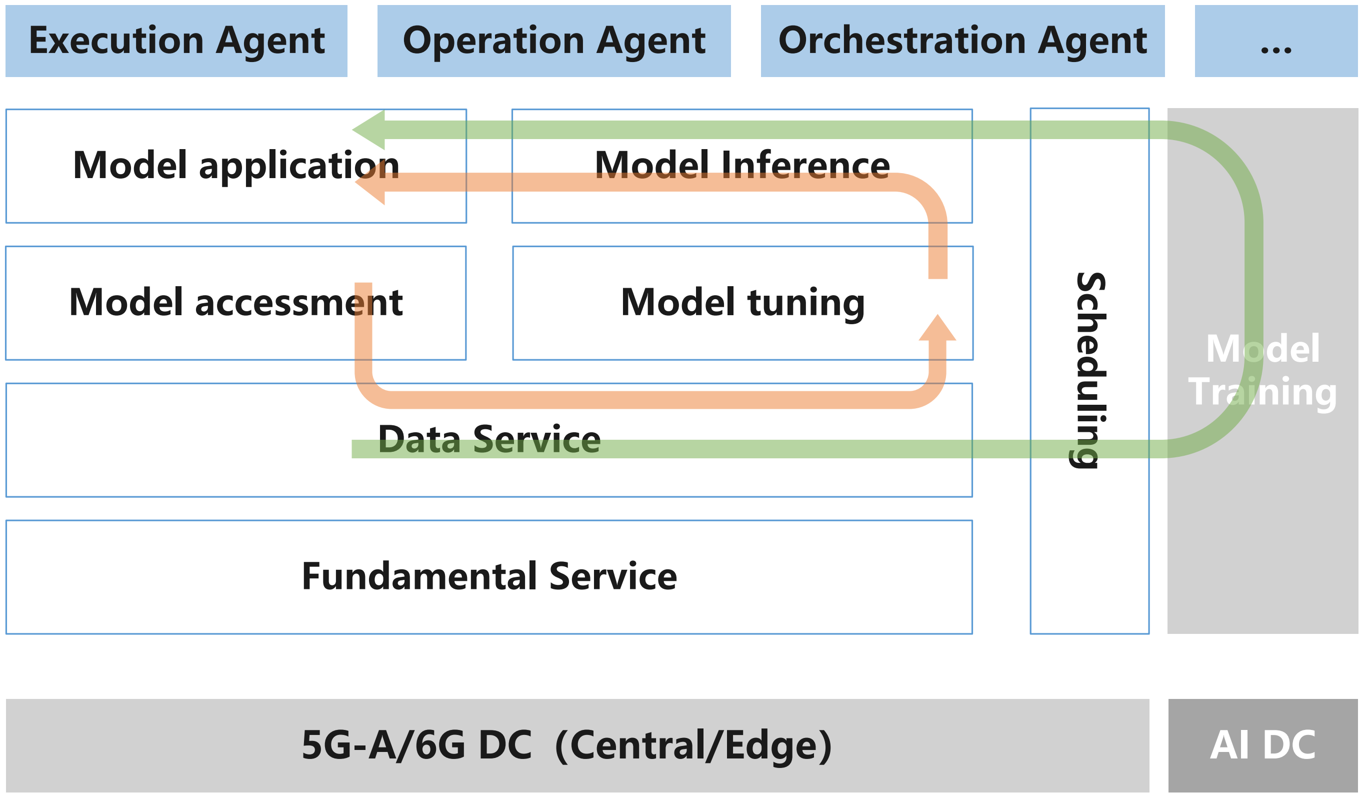
Telco cloud infrastructure requires AI capabilities that emphasize inference over training, utilizing AI to realize integrated sensing and communication, digital twins, and other core objectives. By leveraging open-source or centrally trained models, the MaaS platform develops tailored applications to deliver AI-native model services to business operations.

Different scenarios, such as network operations, O&M, and third-party applications, have specific requirements for model training, fine-tuning, and inference.

The MaaS platform provides two core service capabilities:

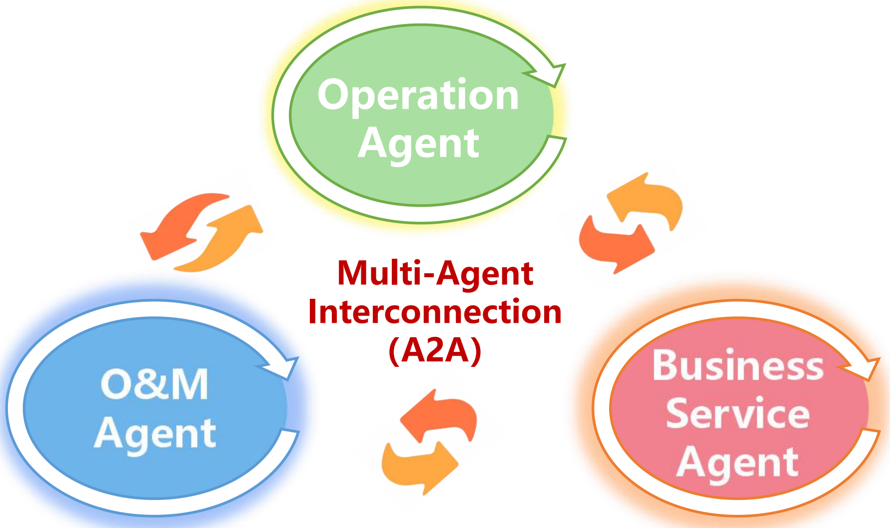
- **Open-source Model Fine-tuning:** Leverage pre-trained open-source LLMs (e.g., LLaMA, DeepSeek) and live network O&M datasets, through processes like model compression, weight adjustment, and knowledge transfer to optimize model performance and maximize alignment with business demands.

- **Proprietary Model Training:** Based on self-developed models and live network O&M datasets, complete model training via large-scale AI data center clusters, thereby delivering proprietary model capabilities for business operations.



Network Intelligence: Agent-Centric Network

Future communication networks will be fundamentally agent-centric. Externally, they need to seamlessly support massive agent-to-agent communications. Internally, by integrating agents across network operations, maintenance, and business domains, networks will be upgraded from static pipelines that passively execute commands into adaptive organisms capable of proactively perceiving their environment and making independent decisions for optimization.



Operations Agent: Capable of real-time perception of network status, user experience, and service quality, **making autonomous decisions** and optimizes network resource allocation based on user demands.

O&M Agent: Driven by data, it **autonomously completes** fault analysis. Through unified natural language configuration, enables large-scale and simplified application of equipment across entire network and all domains.

Business Service Agent: Centered on **emerging forms of intelligent communication**, constructs an intent-driven network to realize smart services like immersive calling, AI translation, and intelligent note-taking.

Serving as the physical and logical substrate for diverse network agents, telecom computing-network infrastructure must deliver holistic, systematic support in resource provisioning, runtime environments, and security governance, ultimately acting as the "native soil" for agent deployment.

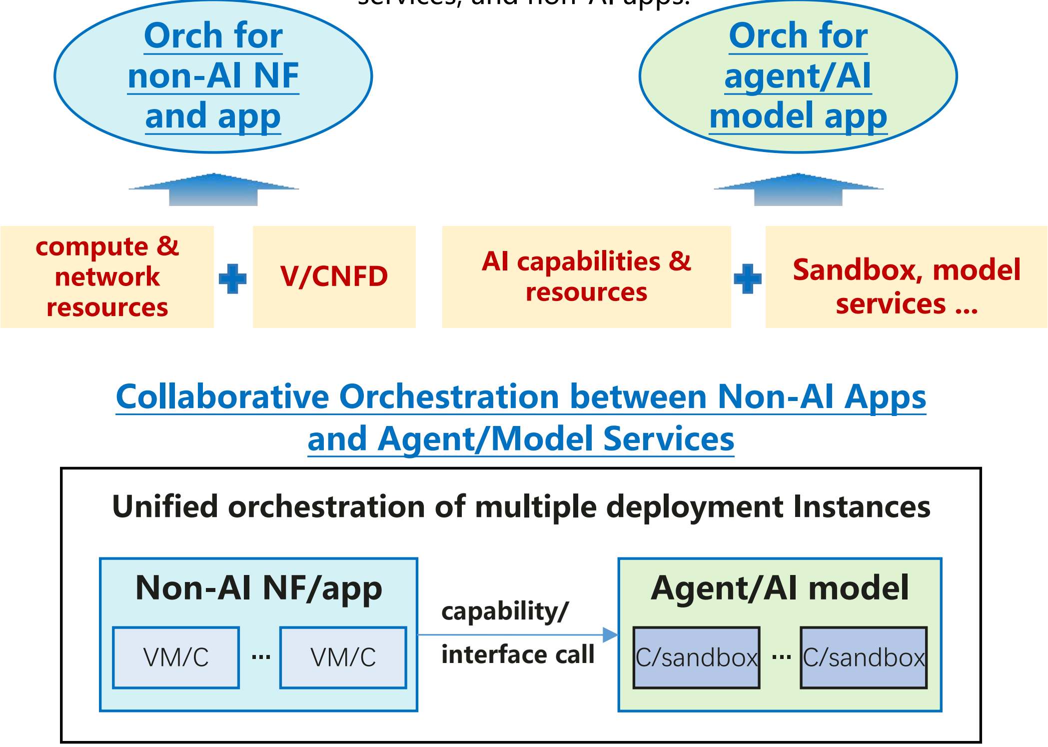
Agentic Runtime	Security Sandbox	Intelligent Resource Scheduling	Memory Management

Network Intelligence: Orchestration for Agent/Model Services

With the rapid intelligent evolution of telecom services, it is necessary to systematically plan and build an E2E orchestration framework that encompasses diverse business entities, including agents, model applications, and legacy non-AI NFs as well. Furthermore, agent technologies could be leveraged to achieve a comprehensive upgrade of orchestration capabilities.

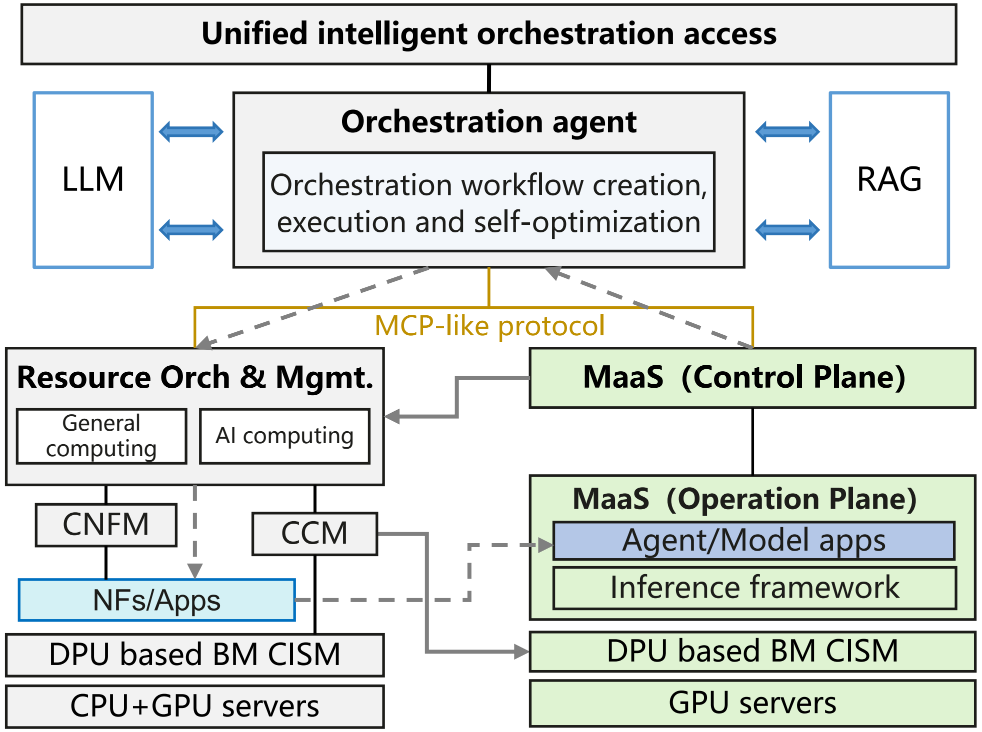
Orchestration for AI

Legacy NFV-MANO primarily addresses the automated deployment of homogeneous VNFs. In contrast, future networks requires the orchestration & dynamic coordination of three types of heterogeneous entities: agents, model services, and non-AI apps.



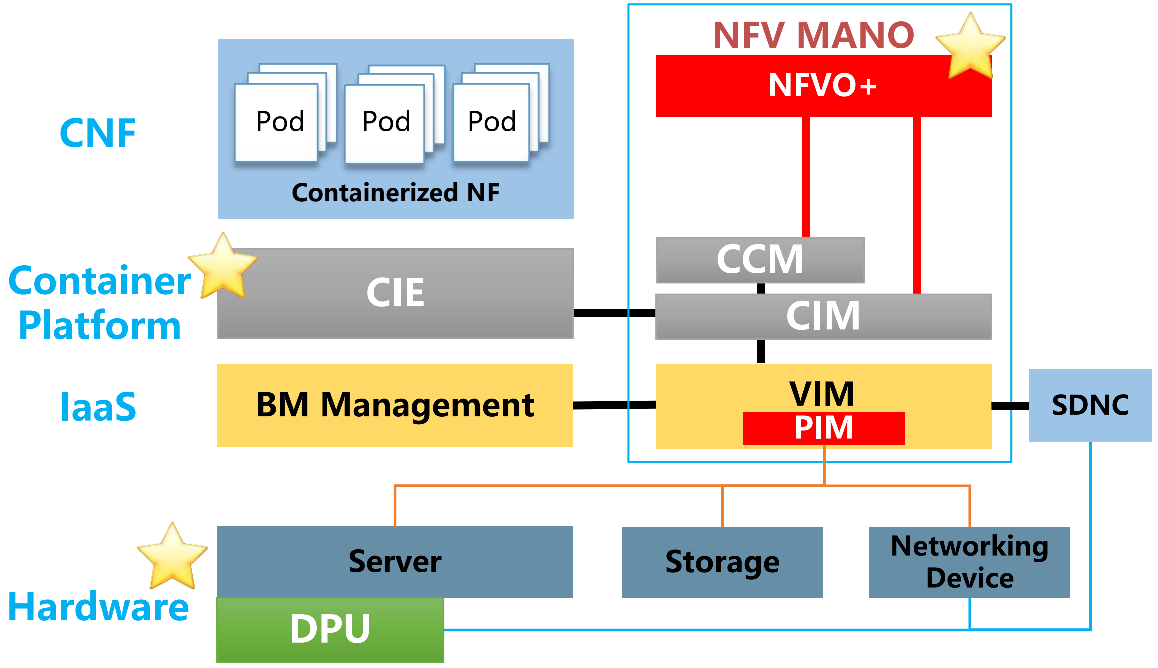
AI for Orchestration

Due to the distinct lifecycles and real-time intent-driven collaborations of agents, model services, and non-AI apps, **integrating agent tech** can upgrade current MANO from a static orchestrator to an intelligent hub with intent-awareness, dynamic planning, and cross-domain coordination.



China Mobile practices on cloud-native for telco cloud

To promote evolution of operators' network, China Mobile introduces two major IT innovative technologies, Cloud Native and DPU, and integrates them with telco-grade enhancements on hardware, container management, MANO, etc. to build an end-to-end cloud-native solution based on DPU smart NICs.



Enhanced based on NFV Classical "Three-layers & one-domain“ Architecture

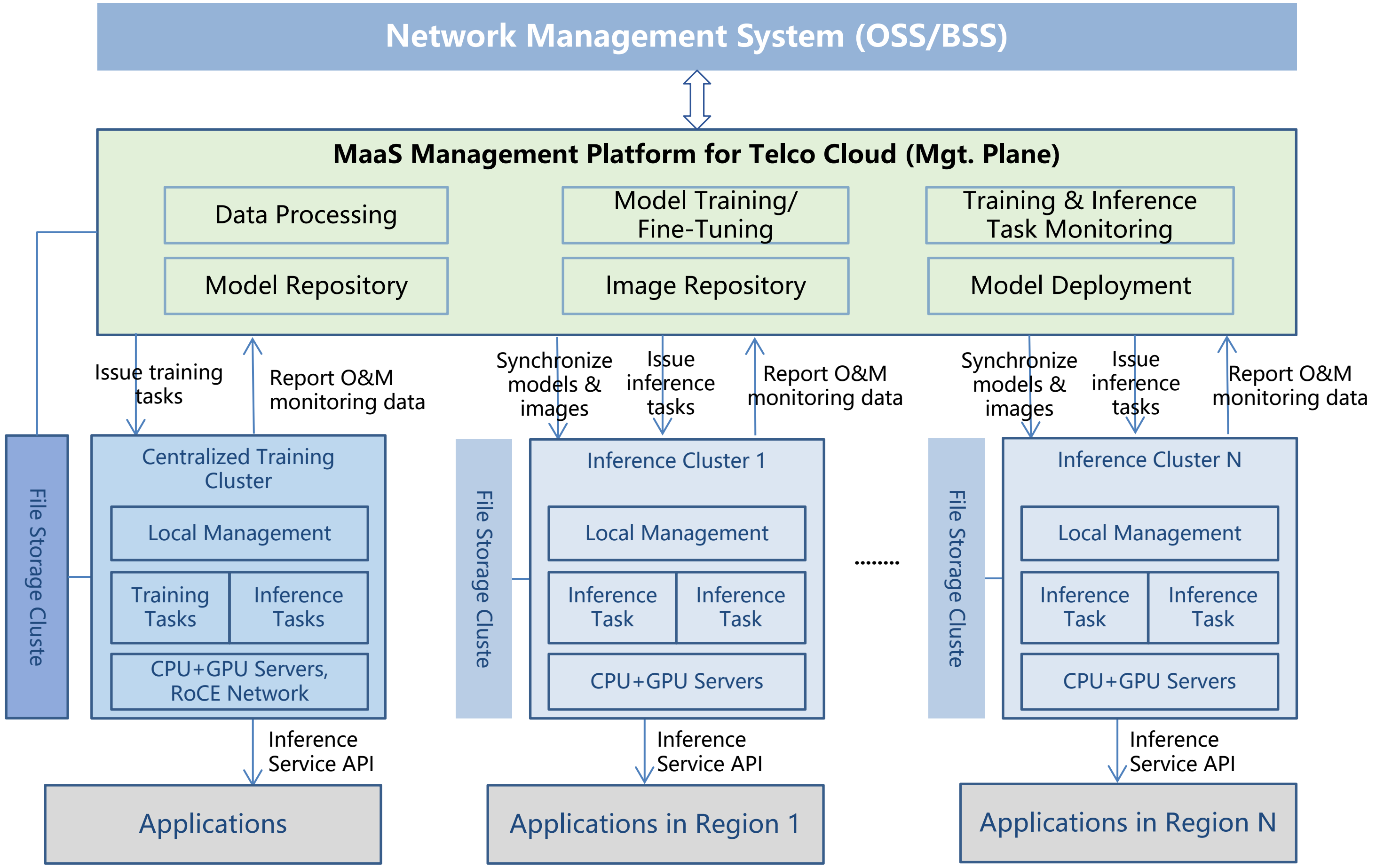
Hardware: Introducing DPU (Data Processing Unit) smart NICs to offload virtualization tasks such as management, network, and storage, and cooperate with the virtualization layer to achieve flexible provisioning of bare metal, virtual network acceleration, and storage network security.

Container Platform: A container-based virtualization layer added, including Container Cluster Manager (CCM, responsible for LCM of container cluster), Container Infrastructure Manager (CIM, responsible for LCM of container apps, networks, and storage), and Container Infrastructure Engine (CIE, providing & managing container operating environment).

MANO: Enhanced telco-grade container cluster, container application orchestration management solution and operation & maintenance management solution based on DPU smart NICs, with unified container northbound O&M interface and standardized container application package format and process.

China Mobile practices on AI-native for telco cloud (1/2)

To meet the demands of intelligent network O&M and intelligent business innovation, China Mobile Telco Cloud introduces MaaS platform. Driven by model fine-tuning and model inference capabilities, it accelerates the transformation of O&M management from "passive response" to "proactive intelligence", and empowers business innovation to shift from a "rule-driven" paradigm to an "intent-driven" paradigm.



Core Capabilities

- Provides capabilities including data preprocessing, model fine-tuning, inference, and evaluation. Collaborates with MANO to complete deployment of training and inference clusters, alongside other lifecycle management and O&M monitoring tasks.
- Supports full lifecycle management of models required by network management applications such as CDAF and MDAF.

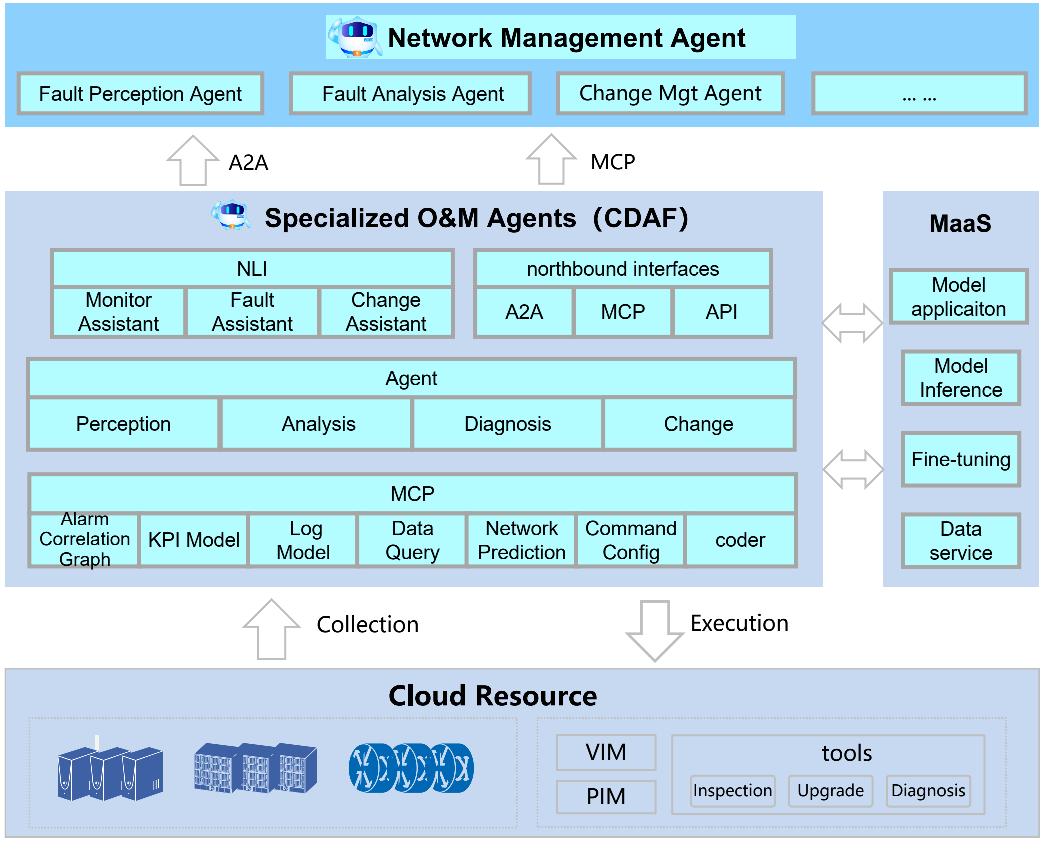
Deployment Scheme

- A single set of MaaS management platform is deployed across the entire network with NBI connecting to upper-layer network management systems (e.g., OSS/BSS), while SBI connecting to and manages all regional business clusters.
- Adopts a layout strategy of "1 centralized training node (cluster) + multiple inference nodes (clusters)". Models fine-tuned in the training cluster are synchronized via the MaaS management platform to inference clusters in all regions for business use.

China Mobile practices on AI-native for telco cloud (2/2)

To address operation and maintenance problems, including lengthy troubleshooting cycles and predominantly manual change operations, we build a two-tier AI agent collaborative architecture which enables timely detection and localization of hidden risks and faults, automatically generates recommended handling suggestions and disposal plans, driving the evolution of telco cloud toward the AN L4 maturity target.

Architecture



Capabilities

Two-tier collaborative architecture of AI agents consisting of a "centralized network management agent" & "distributed regional specialized O&M agents":

- **Network Management Agent:** Serving as the interaction portal for O&M administrators, it focuses on command and scheduling, and invokes appropriate secondary-level agents or tools according to task
- **Specialized O&M Agents:** deliver core atomic capabilities while ensuring open data access and secure execution of actions and instructions. Including perception, analysis, diagnosis, and change agents, which provide O&M functions such as fault perception & prediction, fault localization, and change verification.
- Specialized O&M agents expose northbound interfaces and support protocols such as **A2A** and **MCP** for invocation by upper-layer consumers.

Thank you for listening!