Cloud-aware Network Operation for AI Services

Qiong Sun

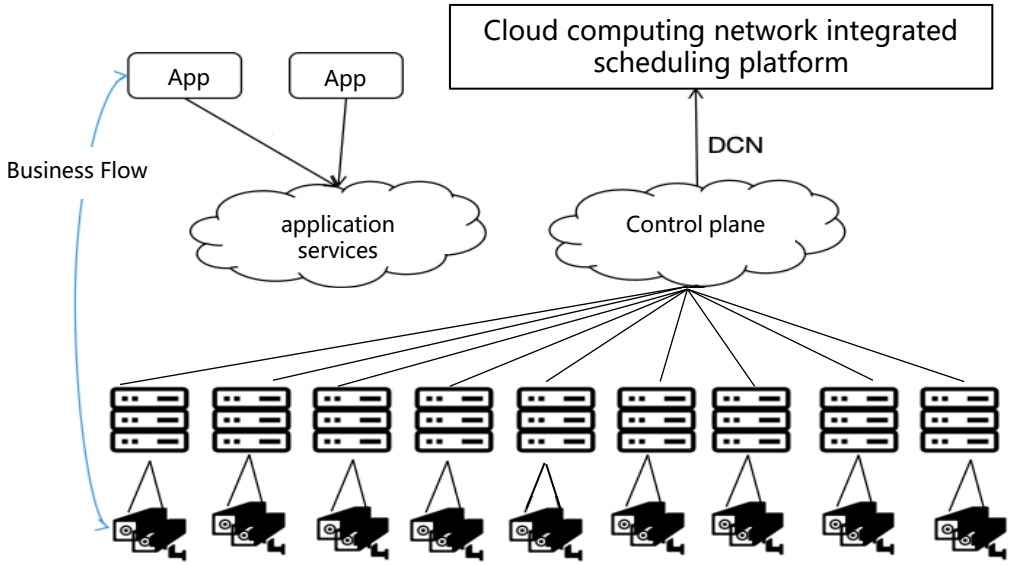
China Telecom

Content

- Overview of the Use Case
- Architecture and Procedures
- New Interface Definition
- Summary

AI-based Video Recognition for City Management

- **Key Requirements:** When deploying AI algorithms for city management, how does operator achieve flexible scaling and dynamic scheduling of cloud and network resources?
- **Objectives:** AI services can be deployed edge clouds of different cities, with shortened service loading time to minutes, elastic response to faults or changes in traffic by using nearby computing nodes.



Road blocking

Illegal Parking

Scattered sundries on the ground

Garbage classification

Challenges

- **Real-time and dynamic resource scheduling:** Traditional network scheduling cannot adapt to sudden network traffic surges or elastic scaling of computing power.
- **Contradictions among different objectives:** Computing power utilization vs. network redundancy, computing power savings vs. network overhead
- **Scheduling effectiveness evaluation:** Existing methods do not cover the joint scheduling of "computing + network," difficult to quantify and verify scheduling effectiveness.
- **Security and strategy fragmentation:** Lacking of a unified model between cloud security groups and network policies increases the risk of cross-domain attack surface exposure.

Content

- Overview of the Use Case
- Architecture and Procedures
- New Interface Definition
- Summary

Architecture of Integrated Network Operation

Network Service Orchestration and Scheduling System (NS-OSS)

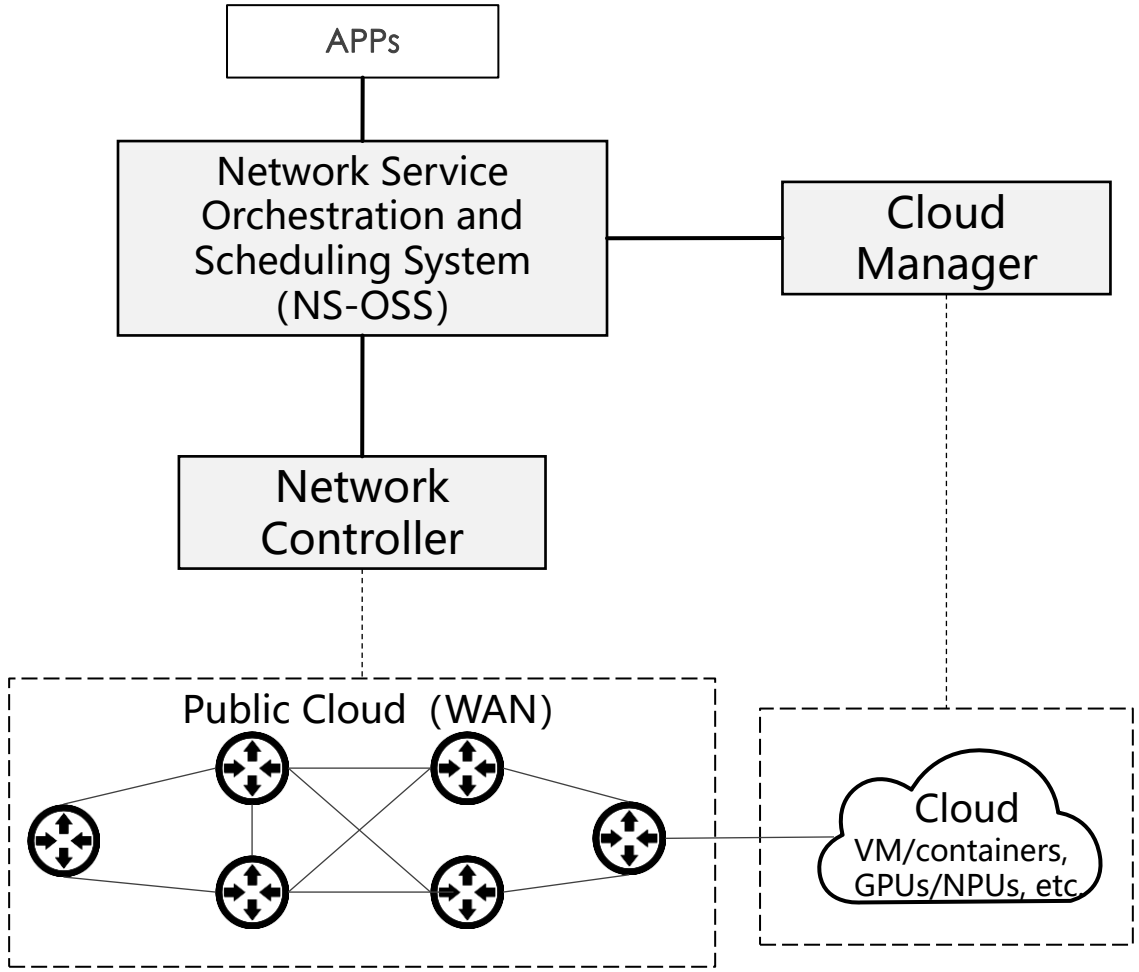
- Orchestration and scheduling of cloud &network resources, cross-domain policy collaboration, monitoring and maintenance.

Cloud Manager

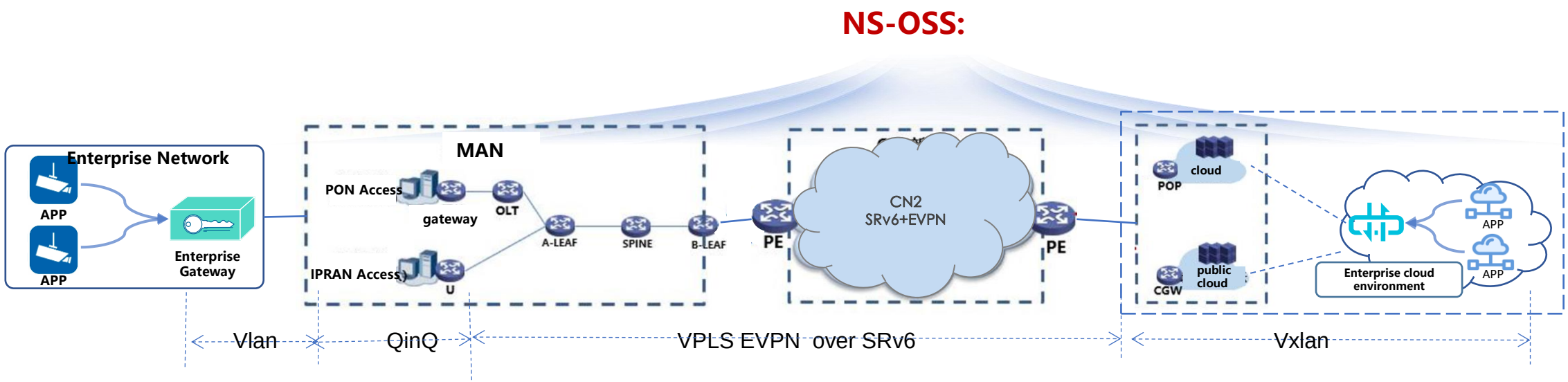
- Cloud resources management: collection, configuration, and monitoring.
- Exposing information to network

Network Controller

- Collection of network information: topology, network load and status,
- Dynamic optimization of cloud traffic
- Cloud-aware configuration of devices



Dynamic Scheduling of Network Resources for Services

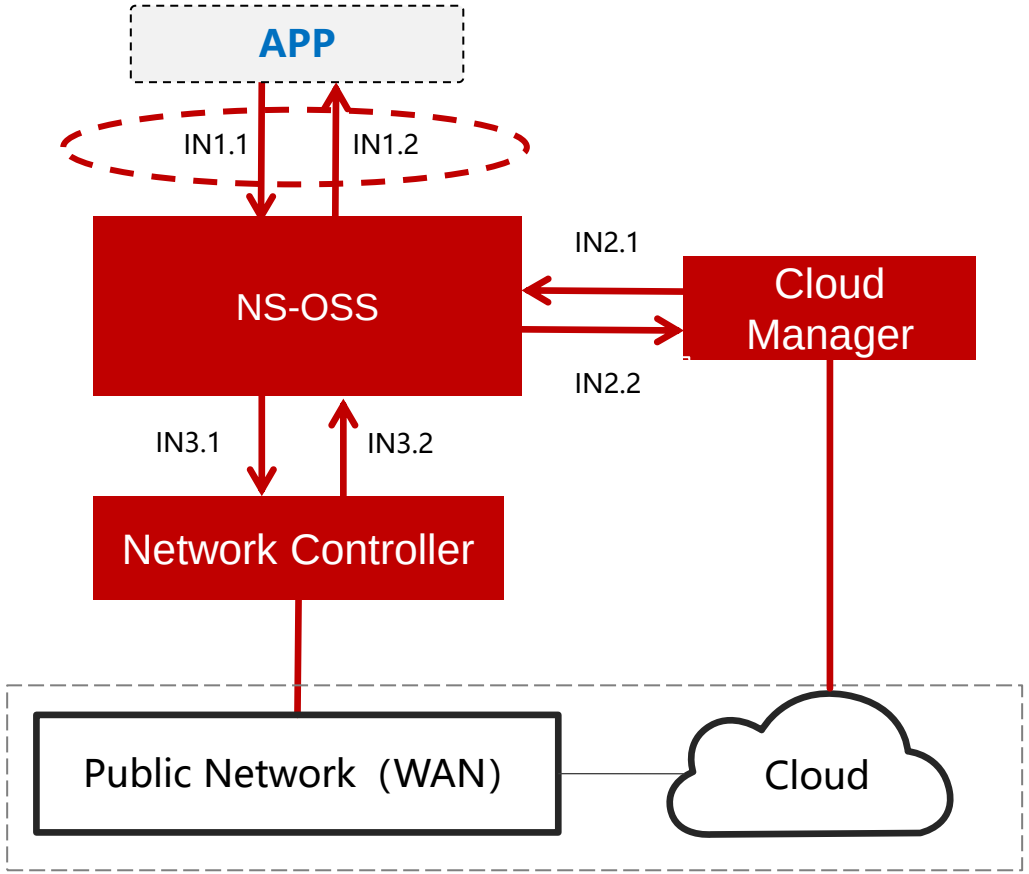


- **Real-time Resource Collection,** Computing resources (GPU, CPU, storage, etc.), network sources(bandwidth, latency, topology, and other parameters).
- **Resource Allocation,** Based on AI service characteristics, edge nodes and network resources are selected from the resource pool (with specific multi-factor selection algorithm).
- **Service Deployment,** when AI models being deployed in edge clouds, the network synchronously issues policies to the network devices through control layer.
- **Cloud-aware Network Scheduling,** In responses to services scaling and faults events, dynamic scheduling of network resources and intelligent routing migration are enabled.

Content

- Overview of the Use Case
- Architecture and Procedures
- New Interface Definition
- Summary

Interface Requirements - IN1



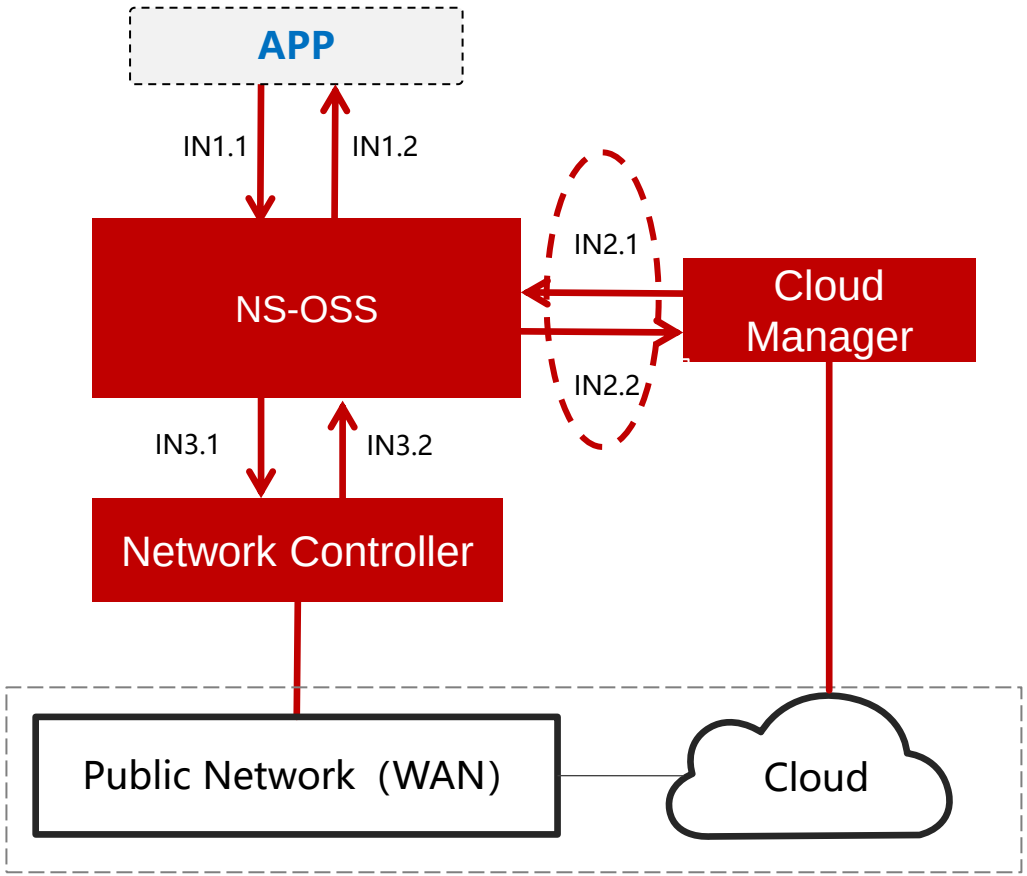
IN1.1：Service Deployment Request

- Cloud Computing Resource：CPU、RAM, etc.
- Storage：Type，size, etc.
- Network：Source、destination location and SLA requirements

IN1.2：Resource Allocation Result

- Cloud/computing resource node information
- Storage location information
- Network slicing information

Interface Requirements – IN2



IN2.1：Exposure of Computing Metric（Cloud Manager →NS-OSS)

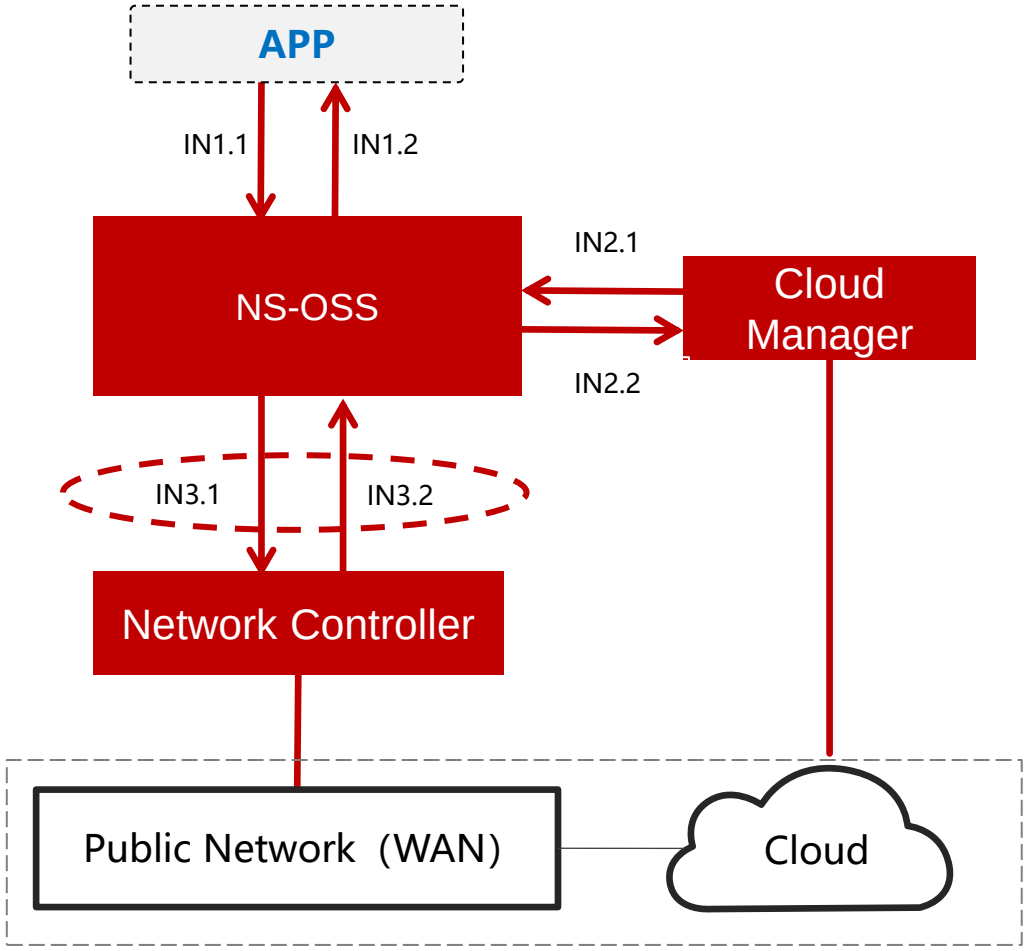
Resource Identification: VM ID/Container Group/Storage Volume ID

- Indicator type: CPU utilization/memory usage/disk IOPS/GPU load
- Sampling period: seconds/minutes/event triggered
- Related service tags: Service/Tenant/SLA level

IN2.2：Computing Resource Scheduling and Control（NS-OSS→Cloud Manager)

- Computing power requirements: computing power types (CPU/GPU/FPGA), Resource quantity (number of CPU cores/memory/GPU model and quantity), Scenarios (training/inference/storage/high-performance computing)
- Network status: topology, bandwidth, latency and other information
- Deployment configuration: availability data center, image identification (operating system/preset image ID), network configuration (VPC ID/subnet ID/security group rule summary)
- Resource pre occupation: resource pool type (public cloud/private cloud/hybrid cloud), pre occupation mode (on-demand/reserved instance), storage configuration (type/capacity/IOPS)

Interface Requirements – IN3



IN3.1：Issuing of Network Control Policy（NS-OSS→Network Controller)

Link identifier: source/destination node ID, logical link name

- Cloud Serivce instance ID
- Target bandwidth required (Mbps/Gbps)
- Effective method: immediate effect/smooth transition (rate gradient time window)

IN3.2：Report of Network Status（ Network Controller→ NS-OSS)

Link ID: Logical link globally unique identifier

- Real-time bandwidth utilization: current traffic percentage (%)
- Delay and packet loss: Avg/Max delay (ms) and packet loss rate (%) in the most recent sampling period
- Timestamp: Data collection time

Content

- Overview of the Use Case
- Architecture and Procedures
- New Interface Definition
- Summary

Summary

- Cloud computing facilities has become an essential part of infrastructure of opertors, requiring for integrated cloud-network resource scheduling and end-to-end security.
- To meet the needs of cloud-based AI services deployment, it is necessary to incorporate cloud related information into network control policies to achieve dynamic resource management and scheduling.
- Lacking of key standardization hinders cross-domain collaboration between cloud resources and wide area network resources.

Standardization Request

- **Service Model:** Standardized expression of cloud service requirements, unified cloud and network resource view，integrated orchestration of resource for optimal deployment.
- **Cloud Manager → NS-OSS：** Exposure of cloud related information to make the whole network operation be cloud-aware, thus achieving the best network resource scheduling policy.
- **NS-OSS—Network Controllers:** Incorporation of cloud metric and service status into network scheduling and configuration policies, achieving dynamic network resource adjustments and services SLA guarantee.
- **Other：** A platform is needed for the community to discuss of the network-cloud operation.

Thank you