

Computing and Network Information Awareness system architecture for CATS

draft-yao-cats-awareness-architecture-00

Huijuan.Yao, China Mobile

Xuewei.Wang, Ruijie Networks

zhiqiang, Li, China mobile

Daniel Huang, ZTE Corporation

Goal

- Computing-Aware Traffic Steering (CATS)[I-D.Idbc-cats-framework]aims to solve the problem of how the network edge can steer traffic between clients of a service and sites offering the service.
- To enable the computing- and network-aware traffic steering decisions,awareness of computing service information and network information is the foundation.
- A comprehensive awareness architecture: introduce new comomponents and the corresponding interfaces and work flows are included,to facilitate the deployment of CATS.

Computing and Network Information Awareness system architecture

A control center component is additionally introduced to support fine-grained dynamic information awareness based on CATS framework.

- **CATS Computing information Base(C-CIB):** Maintain fine-grained computing information, such as service connections, CPU performance, which may be obtained from the routers or from the cloud management platform.
- **CATS Network Metric information Base(C-NIB) :** Maintain fine-grained network information, such as remaining bandwidth, delay, which could be obtained from the routers.
- **CATS Path Calculation Unit(C-PCE) :** Responsible for calculation optimal computing resource and network path based on C-CIB and C-NIB, and generate path policy and deliver to the CATS router .
- **CATS-SBI interface:** an extended interface based on the traditional controller southbound interface between the CATS-routers and the CATS-control center,

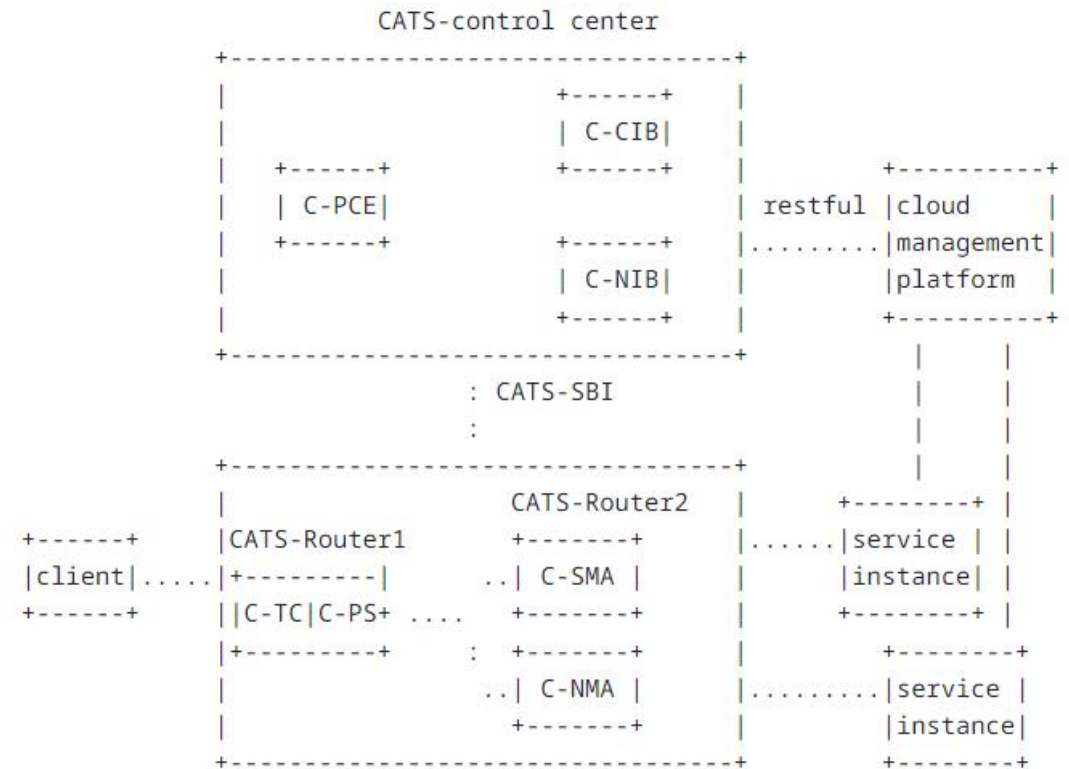


Figure 1: CNIA System Architecture

Given the comprehensive architecture described above, this document proposes a comprehensive perception system of the deployment location, real-time resource and service status, load information and requirements of computing resources and services, to and provides guarantee for computing-aware scheduling based on service requirement

Awareness Information Classification

In order to avoid introducing too much signaling overhead into the whole network advertisement, classify the content of the computing advertisement, according to the characteristics of the content and frequency of information announcement.

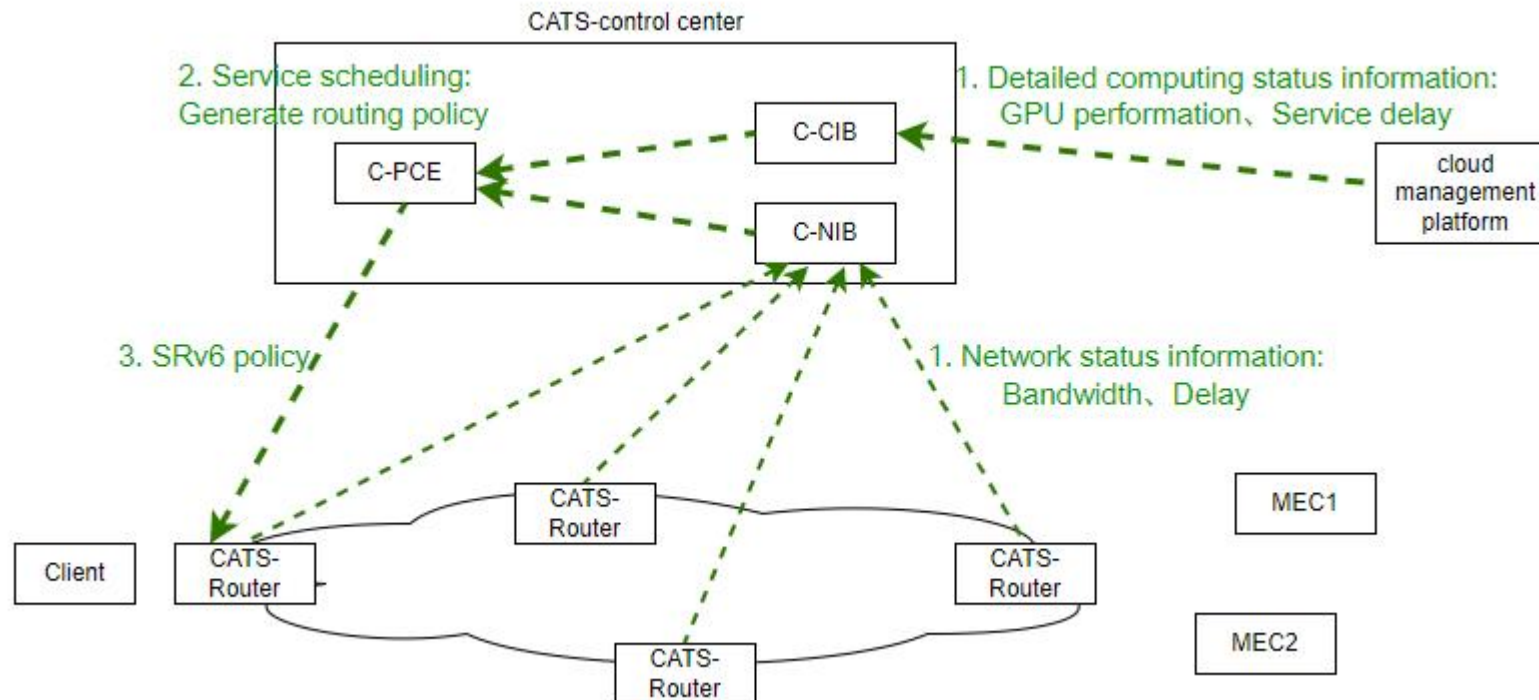
- **Capability information:** relatively static information with low update frequency, such as deployment location, identifier information, and so on.
- **Status information:** high update frequency, real-time status parameters, such as remaining bandwidth, delay, service connections, CPU performance.

Awareness information	Network information	Computing information
Capability parameters	Device location; Device type; Topology information	Service ID; Service-domain name; Computing energy consumption; Computing cost; Peak value of available computing
Status parameters	Service policy information; Traffic information (bandwidth, delay, packet loss rate, delay jitter)	Number of available service connections; Available resources; CPU/GPU/NPU performance; Storage capacity; Service delay

Table 1: Awareness information content examples

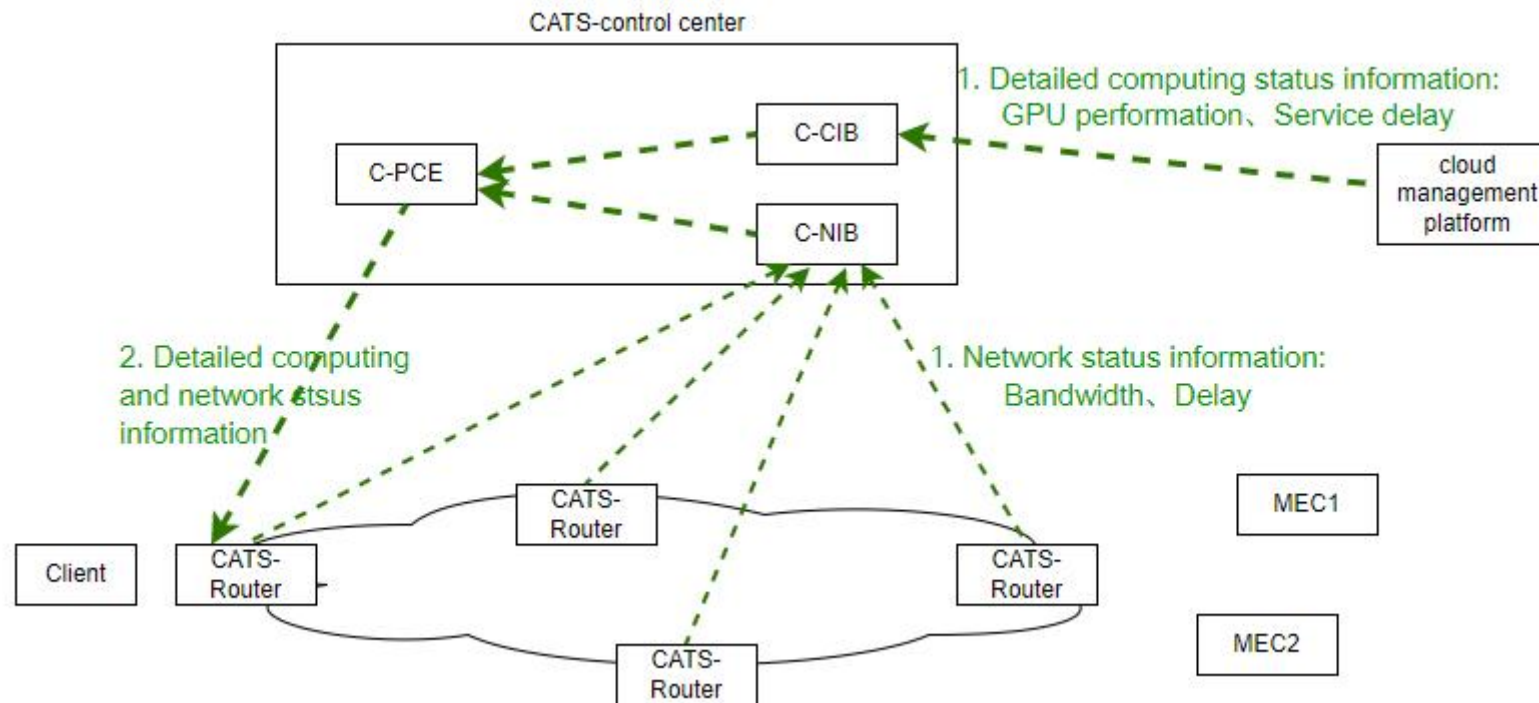
Workflow – A centralized model

- Computing information: aware by the CATS-control center by restful interface from cloud management platform.
- Network information: aware by the BGP-LS or telemetry interface from routers.
- CATS-control center performs service scheduling according to the detailed computing information and network information, then generates routing policy and sends to CATS ingress router.



Workflow – A hybrid model

- the **CATS-control center** obtains computing and network information through SBI or restful interfaces, the details of this information are directly transmitted to the CATS ingress routers. The CATS ingress routers perform accurate resource matching and continuous experience detection after receiving service traffic.
- For some high-value customers, hybrid awareness can be deployed to accurately match customer requirements.



Workflow – A distributed model

- The **ingress CATS router** responsible for collecting computing and network information and scheduling service.
- When receives the service demand from the client **the ingress CATS router makes decision** of the service instance to access independently according to the service instances status and network status and maintains instance affinity.
- The detailed workflow can be seen in [I-D.ldbc-cats-framework].

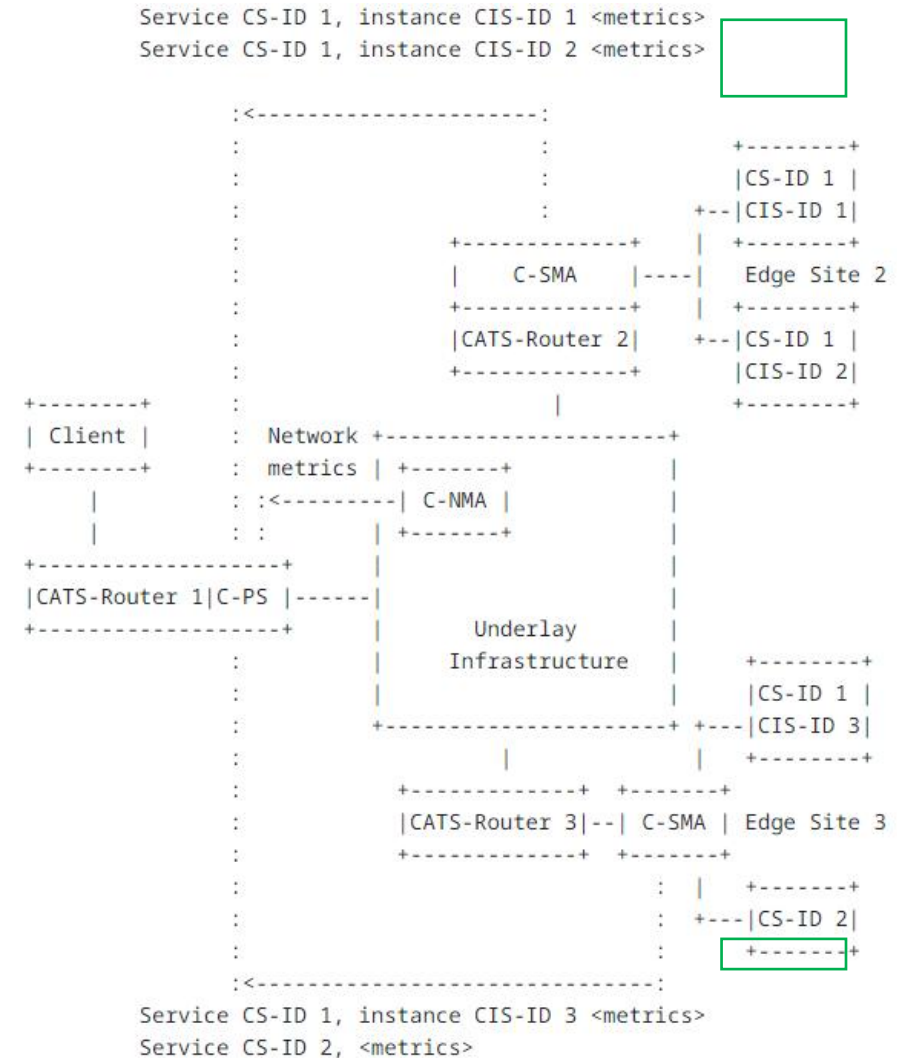


Figure 2: Example CATS Metric Distribution

Comments are welcome

Any questions or comments?

Thank you!