Computing in Network Research Group
Internet-Draft
Intended status: Informational
Expires: January 13, 2021

P. Liu
L. Geng
China Mobile

July 12, 2020

Requirement of Computing in network
draft-liu-coinrg-requirement-03

Abstract

New technology such as IOT, edge computing,etc. propose the requirement of computing in network, so the convergence of network and computing has become a trend. It will bring some new directions and areas to be considered, such as the relationship between network and computing, the influence of integrating computing to the network, and so on.

This document points out the requirements of computing in network according to the development of new Industry, including the network and computing requirements.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 13, 2021.

Table of Contents

1. Overview

The new services' provider expects a user experience with lower latency and high reliability, which put forwards immense challenges to cloud computing and traditional network. Centralized computing requires a long transmission distance of traffic flow, and the existing network technology is to the best of its ability. Network operators start to think about how to meet the higher needs of service provider and users. Computing in the network may solve the

problems because it can provide a flexible network and computing integration system.

To integrate the computing resource to the network, it need to find suitable computing nodes to handle service's request, as well as a forwarding path to them. How much computing resources will affect the delay of service processing, which could also affect the whole network latency. Just as the measurement of network performance has one more dimension, it will interact and cooperate with others. So there are some requirments for both network and computing.

2. Requirements of Network

The network requirements includes precision, concurrent, addressing and information interaction.

2.1. Precision

Precision of the network refers to the deterministic of latency, packet loss rate and perception of computing resources.

* Latency: The traditional network's best-effort forwarding mode can no longer meet the demand of such services for network latency. The deterministic latency brings forward a new measure latitude for network, which changes from in-time to on-time.

* Packet loss rate: It is another factor to evaluate the precision of the network. Utilizing the ubiquitous computing capability of the network, network prediction and segment-by-segment path retransmitting are realized based on AI, network transmission can be optimized and service QoS can be ensured.

* Perception of computing resources: how to precisely obtain the status of computing resource to meet the requirements of business requests is also a challenge to the network. It considers the network status and the performance status of computing resources can be matched dynamicly. So the user experience, utilization rate of computing resources and the network efficiency can be optimum.

For the latency and packet loss rate, some technologies such as time-sensitive network TSN, deterministic network DetNet, etc., have proposed corresponding technical means to provide network bearers with deterministic latency(IEEE802.1Qbv, IEEE802.1Qbu) and packet loss rate and guarantee the user's business experience. However, it also needs to consider how to guarantee the service's end-to-end latency, packet loss rate and resource utilization rate.

For the perception of computing resources, we may consider about the OAM and telemetry to achieve it, however, the performance and information collection strategies are issues that need attention.

2.2. Concurrent

There will be number of computing nodes deployed in the network, or computing fucntions intergrated in the network device for network computing. A serivce's computing request may distributed in several computing nodes in order to response quickly to the client. So there may be a lot of parallel computing task, whitch cause too much connection among the nodes but consume less bandwidth. It will bring great challenges to the concurrent network connection including how to build and deploy these distributed computing nodes to ensure the processing capability of the network , as well as the storage, call of the database are worth studying.

2.3. Addressing

Traditional application-based addressing can not accurately grasp the network performance in real time. The comprehensive performance of addressing results based on application layer may not be the best. It is always to find the consistent host's address and go through a long distance internet, which results in poor business experience.

It need to find some new way to improve the addressing proccess. For example, in the function based addressing, the application deconstruction components on the server side are distributed on the cloud platform, and the business logic in the server is transferred to the client side. So the client only needs to care about the computing function itself, not about the computing resources such as server, virtual machine, container and so on.

2.4. Information interaction

The network needs to have the ability to sense application's requirments and expose network and computing status. For example, application can tell the network requirements including bandwidth, latency and jitter, as well as the computing requirements, such as CPU, storage and memory. The network also can have the capability to be aware of the application's requirements. Thus it can effectively support the network programming, which could meet the future business requirements.

3. Requirements of computing

The computing requirements includes computing resource deployment, discovery, reservation, scheduling and OAM.

3.1. Computing resource deployment

If some computing tasks in the network is planned to be implemented, it needs to consider about what kinds of chips and where should them be deployed. On the one hands, different kinds of computing require different kinds of chips, such as CPU, GPU and memory chip. On the other hands, those chips may be put into router, switch, server or some dedicated machines, which are connected by the network.

There is an example about AI algorithm which might be discussed before. The AI algorithm has several steps including training and matching, and they also have different requirements of chips. In network computing, those steps could be distributed in different computing nodes.

3.2. Computing resource discovery

The network needs to have the ability to discover computing resources. when the computing nodes are deployed in the network, it need to be registered to the network management system, and the information of computing resource or routing can update. In this way, when there are computing tasks to be executed, the network can reasonably allocate resources according to the needs of the application.

3.3. Computing resource reservation

There might be serial distributed computing model of computing in the network, and different resources need to be reserved for different nodes. For example, AI algorithm now has a model of step-by-step iteration at multiple nodes. The previous iteration will affect the next calculation results, and the computing resources required for each iteration are not the same. From the perspective of network standard, we hope to regard computing resources as the dimensions to measure network performance, such as the same bandwidth, path, etc., while the traditional technologies of resource reservation have not considered the reservation of computing resources, and have not considered the differentiated resource reservation model. Therefore, new protocol or extension of existing protocol is needed to meet the requirement.

3.4. Computing aware scheduling

Computing in network needs a reasonable scheduling strategy, which means computing aware scheduling. According to the business requests, dynamicly computing power matching is carried out based on network status and performance of computing resources to achieve optimal user experience, optimal utilization of computing resources and optimal network efficiency. In computing aware scheduling, computing is seen as "link state" and the computing resource information should be exposed.

3.5. Computing resource OAM

The ability of OAM can be used to continuously update the current computing power resources, and perform some troubleshooting tasks. However, OAM of computing resource is more complex than network. Network monitoring is relatively simple, like bandwidth, latency, jitter, while computing can be divided into many categories, different application need different kinds of computing. So it need to implement fine-grained OAM of computing resource.

4. Requirements of management

The management requirements includes cross domain management, joint optimisation and multi user access.

4.1. Cross domain management

The computing in the network should ensure the end-to-end network management to meet the needs of different network topology, performance and function, which involves cross domain network arrangement. In the process of network data transmission, different services will forward in different ways or different network protocols, and computing resources may be distributed in different network domains. Effective cross domain management will enhance the performance of network and computing.

4.2. Joint optimisation

As computing resources are integrated in the network, and may be used as one of the measurement dimensions of network performance, joint optimization is also a very important part. Network and computing resources will affect each other, including performance, scheduling and so on. So It need a good joint optimization scheduling strategy.

4.3. Multi user access

Many existing applications, such as games, remote video conferencing, are usually multi--accessed and interacted by several users at the same time. This brings about the problem of service consistency, that is, users accessing to the same game or video need the consistency of SLA, otherwise it will seriously affect the experience of other users. Service consistency can be achieved through network management or application layer control.

5. Conclusion

Based on the requirements of new business, this document puts forward the requirements of computing in network, and gives some reference technologies and use cases. Computing in network is a new direction, some details need more in-depth discussion and research.

6. Security Considerations

Computing In network has brought the trend of network convergence in different regions. For example, 5g network of operators can go deep into the vertical industry user site to provide users with higher quality network services, which will bring the convergence of operator's network and user site network. Besides, industrial Internet brings the trend of integration of industrial OT network and IT network to further improve the production efficiency of the industry. It need to ensure the security of the network, including the mutual trust and non aggression of information among regions, which may require further protection and detection measures.

7. IANA Considerations

TBD.

8. Normative References

[I-D.kutscher-coinrg-dir]
Kutscher, D., Karkkainen, T., and J. Ott, "Directions for Computing in the Network", draft-kutscher-coinrg-dir-01 (work in progress), November 2019.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

Authors' Addresses

Peng Liu
China Mobile
Beijing 100053
China

Email: liupengyjy@chinamobile.com

Liang Geng
China Mobile
Beijing 100053
China

Email: gengliang@chinamobile.com