Computing in Network Research Group
Internet-Draft
Intended status: Informational
Expires: May 6, 2021

Y. Fu
P. Liu
L. Geng
China Mobile

November 2, 2020

Requirements of computing and network joint optimization and scheduling
draft-fu-coinrg-joint-optimization-req-00

Abstract

With the development of edge computing, there is a trend that computing is widely deployed in network rather than at other end of network, and provides services at nearer location. With the deep integration of network, traditional optimization and scheduling within network domain is not enough, the endpoint of the path matters a lot. So the relationship between computing and network are new and important topics to be studied. This document focus on the requirements of computing and network joint optimization and scheduling based on the newly arising service 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 .

Status of This Memo

Copyright Notice

Table of Contents

1. Overview

For traditional services without strict service requirements, the best-effort network can meet the requirements with traditional path optimization, which only consider the network condition. With new services arising, such as cloud AR/VR, cloud gaming, V2X, new and strict requirements towards network, also towards the service endpoint are proposed to meet the service requirements. So the computing and network joint optimization and scheduling are proposed to guarantee the service performance.

The computing and network joint optimization means that there is not only the path optimization in network, but also the endpoint joint optimization; also the two will affect each other. Based on the joint optimization, the service scheduling can be performed considering the network condition and also the endpoint condition, with the "optimal path+ optimal endpoint" policy. What's more, the computing and network resources joint reservation is required for services with strict performance requirements.

2. Requirements of awareness of computing

The service requirements arising include both network and computing requirements, which further require future network should perform the joint optimization according to service requirements. So the requirements towards the joint optimization are: the awareness of computing and network requirements, the awareness of computing resources and services in network, the computing-aware path optimization and the network-aware endpoint optimization.

2.1. Awareness of computing and network requirements

Awareness of computing and network requirements refers to consider the computing requirements in addition to the network requirements, including the awareness of computing requirements and the measurement of computing services.

Since network requirements can be measured with bandwidth, delay etc, it is also required to measure computing requirements in a unified way. On the one hand, there are many different computing services which are the "consumers" of computing resources, such as video processing, image classification etc, and they propose various requirements towards computing. It is required to firstly obtain the computing requirements and then model the requirements in a unified way, which then can be used as the constraint of joint optimization.

What’s more, the computing service modes are abundant compared to network services, which are the computing "producers", including there are heterogeneous hardware such as GPU, CPU, FPGA etc, and also various algorithms deployed in network, so it is also required to model the computing producers in a unified way, which is another important factor for joint optimization. As for the awareness of computing requirements, some technologies such as application-aware networking have proposed corresponding technical solutions to delivery computing requirements in the packet head, however, it needs further study on the security of application and also the efficiency of the information delivery. As for the measurement of computing services, there is no mature solution to model the computing requirements and the computing resources in a unified way, which is a challenge for the computing and network joint optimization.

2.2. Awareness of computing resources and services

With the development of edge computing, the computing resources and computing services will be distributed in network, since the limited physical conditions, each computing site will be small scale and with limited computing resources, so different from the cloud computing, which can finish the computing task within one site, the edge computing sites need the collaboration among many sites, and this collaboration can be done in network. To coordinate the computing sites, it is required for network to be aware of the computing status of edge sites, including the real-time status of computing resources and computing services. So how to generate the required information and then broadcast it to network brings new challenges.

3. Requirements of computing and network joint optimization

3.1 Computing-aware path optimization

With new services requiring computing and network resources, traditional network-based path optimization can not accurately guarantee the service requirements. The network-based path optimization only according to network conditions can only make sure the performance of network services, it can only find a best path towards a given endpoint, however, the given endpoint may be not optimal, causing the service requirements cannot be met.

So It is required to do the computing-aware path optimization to consider the status of endpoint. For example, before the path optimization, according to the awareness of computing resources and services in network, including the location and status, the network could firstly find a list of optimal computing nodes, then the network could do path optimization with different computing endpoints, which changes the traditional way to only do the path optimization with one destination.

To better optimize the computing-aware path, we need to consider different weights of computing and network metrics when calculating the optimal path. For traditional path optimization, there are only network metrics as the parameters of algorithm; it is required to add computing metrics also as the calculation metrics of the algorithm and to combine the computing and network metrics.

What’s more, based on the awareness of service requirements, for different services, there will be different requirements towards computing and network. For some computing-intensive services, computing counts more on the whole process of services, so they will require more on computing than network; and for communication-intensive services, the computing is less during the service process, while there will be frequent communication, which will propose higher requirement towards network than computing. So it can be inferred that computing and network matters differently during the service process for various services.

Based on what discussed above, it is required to adaptively define different weights of computing and network metrics for different services, adapting to various service requirements. For example, for the computing-intensive services, it is required to put more weights on computing metrics than network metrics, which could be based on the percentage of predicted computing time in whole time; as for communication-intensive services, more weights could be put to adapt to the service requirements.

3.2 Network-aware endpoint optimization

Based on the computing-aware path optimization, there will be the optimal "path + endpoint" pair, combing the computing and network status. But there will also be inner scheduling in computing node, which may also influence the computing time. With proper task assignment, the computing time could be less to make sure that endpoint provides the promised services. So it is also required for endpoint to know the service requirements precisely, otherwise the endpoint will just do the usual scheduling without considering the service requirements.

With the network-awareness, the endpoint will know the performance of network, such as the endpoint will know the transmission time in network and then calculate the rest of required time, and then it will do the inner scheduling accordingly.

4. Requirements of computing and network joint resource reservation

For services with strict computing requirements, the resource reservation should include network reservation and computing reservation, also, the two will affect each other.

There is network resource reservation in traditional QoS guarantee mechanism based on the network resources reservation calculation to reserve specific resources for specific services. With new services arising, the network resources reservation is not enough, since the completion of services include not only network transmission but also endpoint calculation, only reserving the network resources cannot make sure the required computing resources are available during the required time for specific service.

So facing the trend of computing and network convergence, it is also required to reserve the computing resources together with the network resources. Based on the awareness of service requirements and the joint path optimization, it is required to map the computing requirements into the corresponding computing resources reservation, for example, to map the services type into the computing resources type, and translate the computing latency requirements towards the required amount computing resources.

On the other hand, the reservation of network and computing resources are closely linked, there will be different network resource reservation policy considering the computing resources reservation. For example, the order of the two resources reservation requires to be considered since they are relative independent usually.

What’s more, it is also required to dynamically adjust the resources reservation according to real-time status. One scenario is that the computing resource reservation could be adjusted based on the information from network domain, including the reservation time and also the reservation amount. Another scenario is the co-adjust of the two resources reservation, in network domain, the path and the relative reservation could be adjusted, and then the computing domain is required to adjust on-demand.

5. Conclusion

Based on the new services’ requirements on computing and network, this document puts forward requirements of computing and network joint optimization, and also proposes requirements of computing and network joint resource reservation. Computing in network is a new direction, how to collaborate computing and network need further study.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. Acknowledgements

TBD.

9. Informative References

[I-D.li-apn6-problem-statement-usecases]
Li, Z., Peng, S., Voyer, D., Xie, C., Liu, P., Liu, C., Ebisawa, K., Previdi, S., and J. Guichard, "Problem Statement and Use Cases of Application-aware IPv6 Networking (APN6)", draft-li-apn6-problem-statement-usecases-01 (work in progress), November 2019.

Authors' Addresses

Yuexia Fu
China Mobile
No.32 XuanWuMen West Street
Beijing 100053
China

Email: fuyuexia@chinamobile.com

Peng Liu
China Mobile
No.32 XuanWuMen West Street
Beijing 100053
China

Email: liupengyjy@chinamobile.com

Liang Geng
China Mobile
No.32 XuanWuMen West Street
Beijing 100053
China

Email: gengliang@chinamobile.com