alto Working Group ZJ. Sun, H.Yang, C.Li
Internet-Draft Beijing University of Posts and Telecommunications
Intended status: Standards Track S.Liu
Expires: 11 January 2023 China Mobile Research Istitute
H. Zheng
Huawei Technologies
25 July 2023

Architecture of Computing Power Optical Network
draft-sun-alto-arch-computing-optical-network-03

Abstract

This document describes the architecture of computing power optical network.

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 6 September 2022.

Table of Contents

1. Introduction

With the rapid popularization and application of cloud computing, artificial intelligence and other technologies, the total amount of data has increased explosively, and the demand for data storage, computing and transmission has increased significantly. This puts forward higher requirements for flexible network scheduling and quality of service. More importantly, the upgrading of industrial intelligence will bring about the diversity of devices, such as the application of Internet of things (IOT) sensors, cameras and other devices will produce diverse data. The processing of these heterogeneous data needs ubiquitous computing power to support.

The computing power network is the link that efficiently connects ubiquitous computing power resources and massive user data. With the advantages of ultra-large capacity, ultra-long distance, low latency, and flexible scheduling, optical networks provide a wide coverage, flexible and efficient super-capacity guarantee for computing resources.

The architecture of Computing Power Optical Network supports network-aware applications, networks, computing power and user needs, coordinates the scheduling of computing power resources and network resources, and provides the best user experience. This architecture combines the computing power network with the optical network to realize the collaborative linkage between edge computing and cloud computing.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

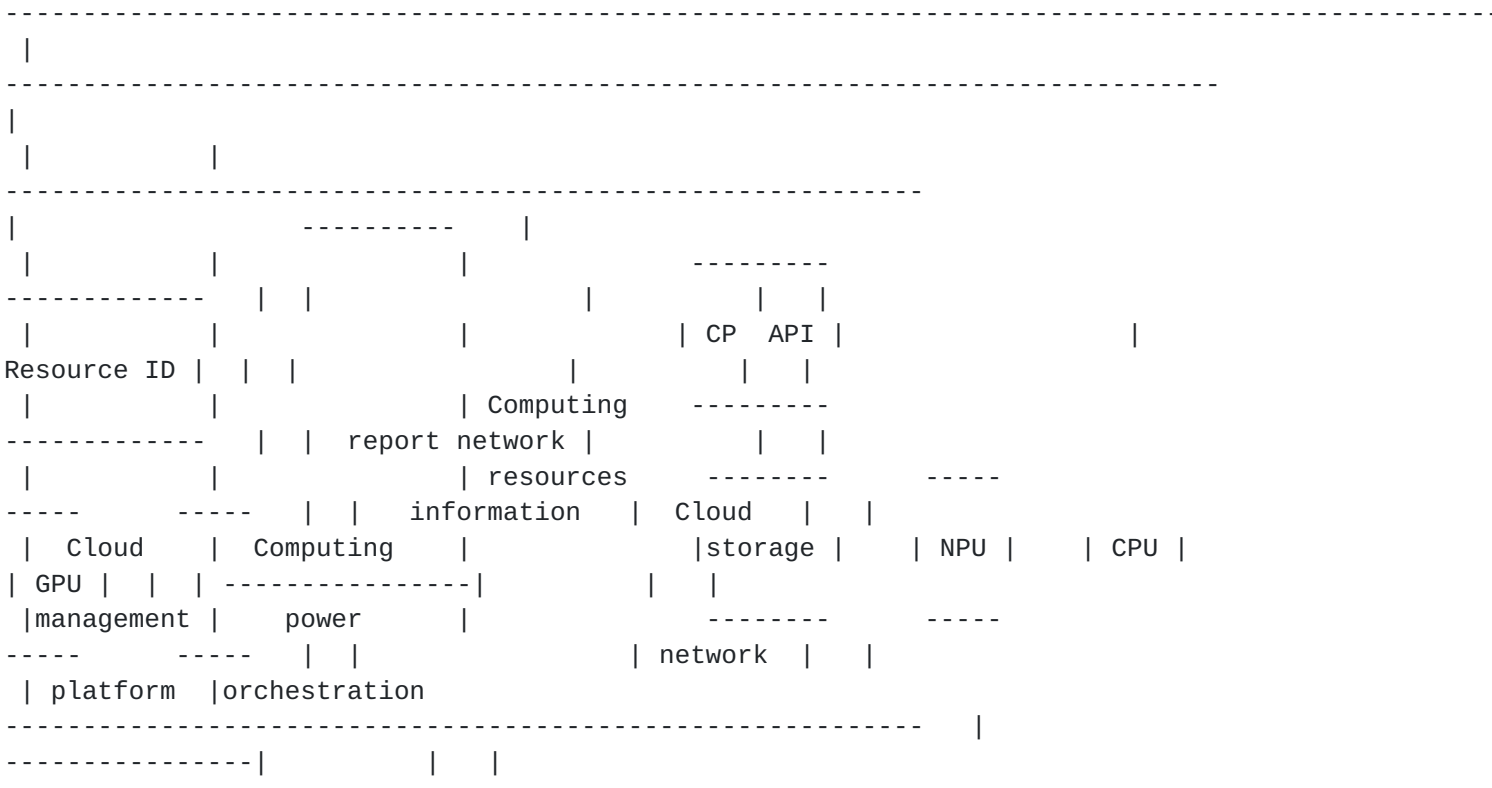
2. Scenarios

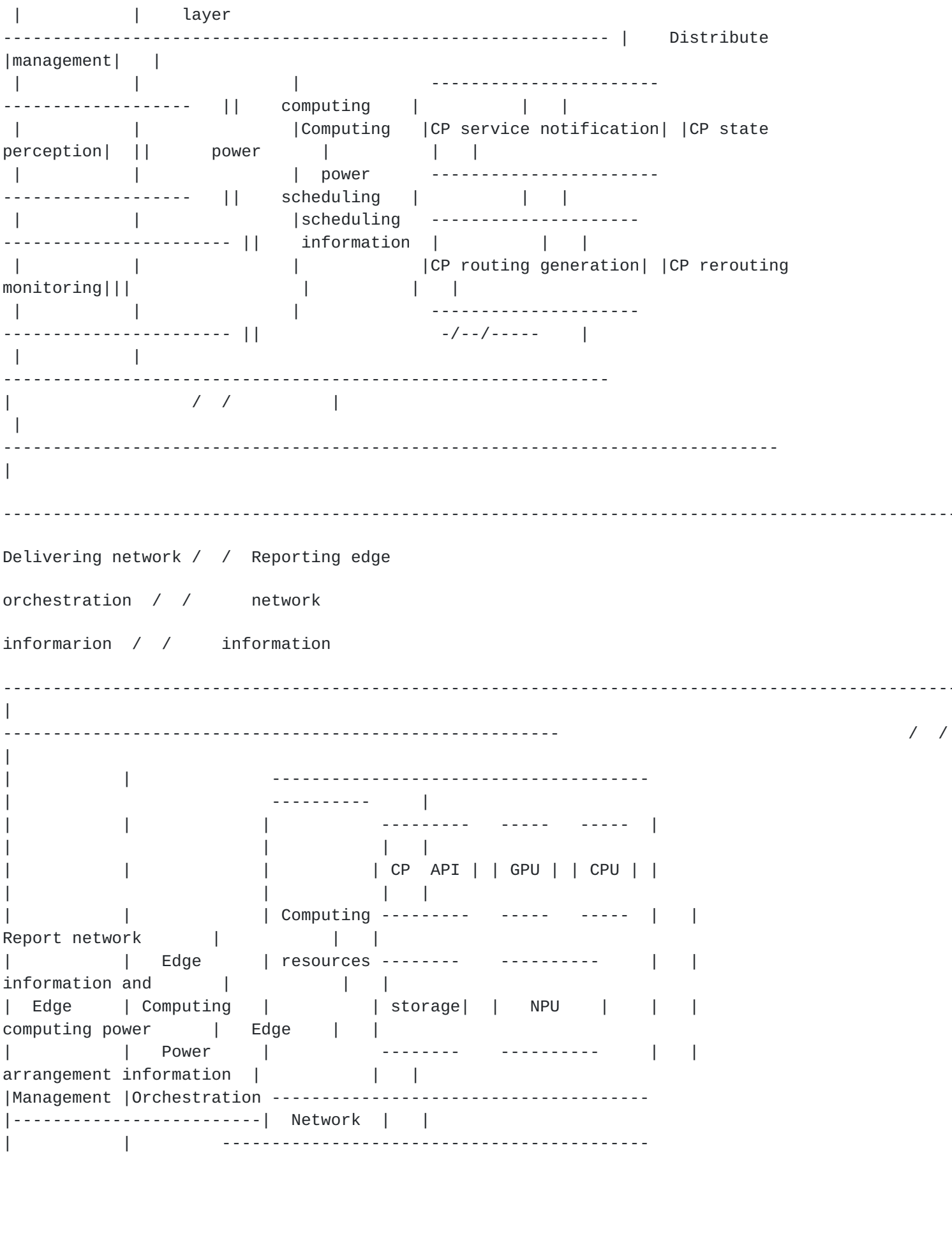
With the prevalence of cloud services, enterprise services and other services, the architecture of computing power optical network has become the choice to solve supported services. The following scenarios provide some typical applications.

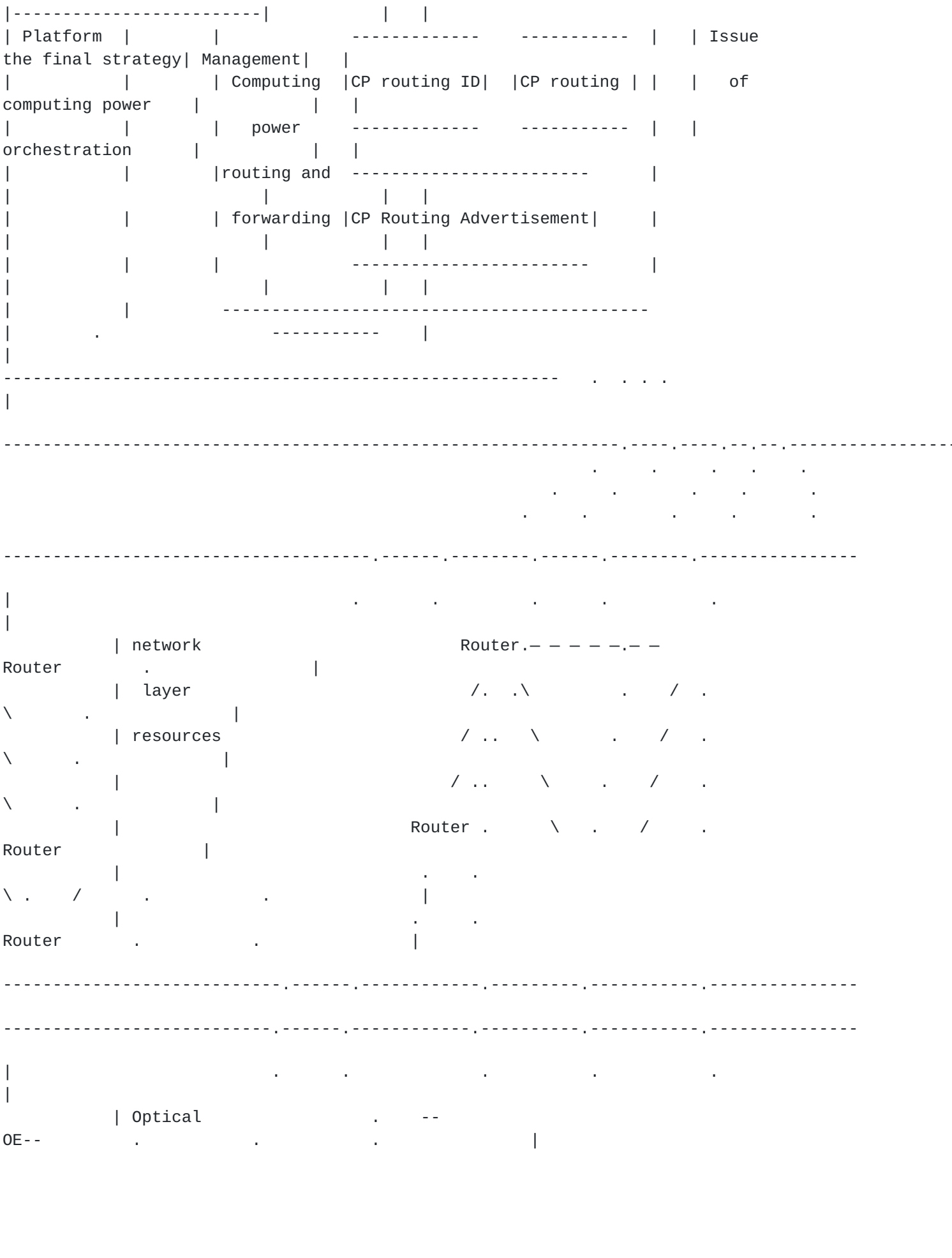
2.1. Network Resource Acquirement

The edge network management layer receives information from the client, obtains complete user information, and provides it to the cloud management platform for network resource synchronization. The cloud management platform obtains regular information about applications and networks.

3. The architecture of Computing Power Optical Network




layer
Distribute
management
computing
Computing
power
power
scheduling
scheduling
information
CP service notification
CP state perception
CP routing generation
CP rerouting monitoring
Delivering network orchestration informarion
Reporting edge network information
information
CP API
GPU
CPU
Computing resources
storage
NPU
Edge Computing power
Edge Power arrangement information
Report network information and computing power
Management
Orchestration
Network


Platform
Management
Computing power orchestration
routing and forwarding
CP routing ID
CP routing
Issue the final strategy of computing power
CP Routing Advertisement
network layer resources
Router
Optical
OE--

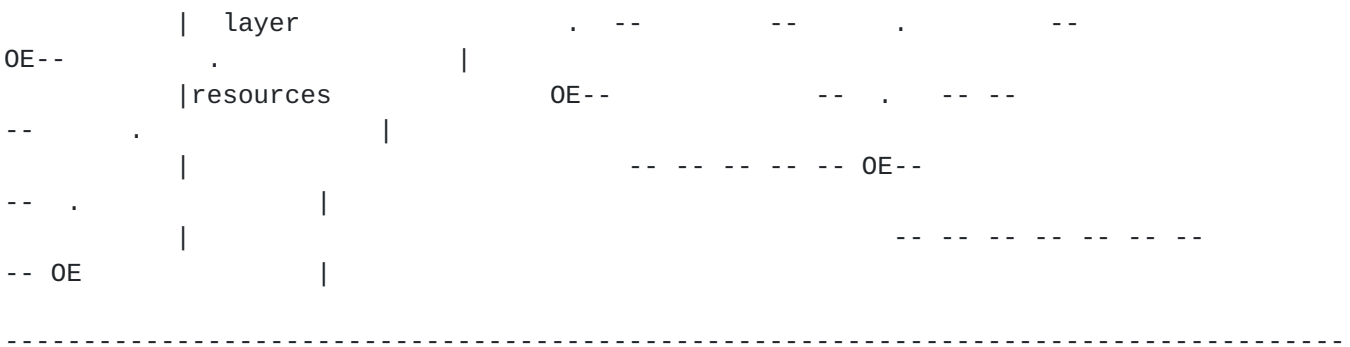


Fig.1 The architecture of computing power optical network.

3.1. Cloud management platform

In order to realize the perception of ubiquitous computing and services, cloud management platform includes computing power scheduling and network management. Computing power scheduling can be divided into computing power resources and computing power scheduling. Computing resources are the use of existing computing infrastructure to provide computing resources. The computing infrastructure is mainly composed of edge computing nodes and network devices, which are controlled by the computing network control layer through the north interface, and provide computing, storage and network facility resources for the serverless edge computing network reference architecture. The computing infrastructure includes a combination of various computing capabilities such as single core central processing unit (CPU), graphics processor (GPU), network processor (NPU). In order to meet the diverse computing needs of the edge computing field, this layer can provide functions such as algorithm library, computing application programming interface (API), computing network resource identification, etc.

The computing power scheduling layer is the core of the computing power aware network, which consists of computing power service notification, computing power state awareness, computing power route generation and computing power route monitoring. Based on the abstracted computing network resources, and considering the network status and computing resource status comprehensively, the computing power routes that can flexibly schedule services to different computing resource nodes on demand are generated, and the real-time monitoring of computing power routes is carried out. Computing power nodes, including terminals, edges and cloud data centers, need to collect and distribute information about their computing power resource status, such as CPU processing capacity, queue status, cache status, computing power node address, etc. through the control surface network mechanism, that is, create a global computing power routing table at the forwarding node of the whole network computing power network, in case of

application requests for optimal routing scheduling in the whole network computing
resource pool. The cloud network management layer reports the current information
of the network to the computing force arrangement layer and accepts the computing
force arrangement information issued by the computing force arrangement layer.

3.2. Edge management platform

Edge management platform includes edge computing force arrangement and edge network management. Edge computing scheduling can be divided into computing resources and computing routing forwarding. Computing resources include computing application programming interface (API), central processing unit (CPU), graphics processor (GPU), network processor (NPU), and storage composition. They are controlled by the computing network control layer through the north interface, providing computing, storage and other resources for the server free edge computing network reference architecture.

The computing power route forwarding layer is composed of computing power route identification, computing power route addressing, computing power route notification .
Through the distributed edge computing nodes, through the automatic deployment of services, optimal routing and cross layer optimization, the edge computing power aware network is built, which can truly call different computing resources on demand and in real time, improve the utilization efficiency of computing resources, and finally realize the optimization of user experience Optimization of computing resource utilization and network efficiency.

The management layer of the edge network reports the network information and the arrangement information of the edge computing power to the arrangement layer of the edge computing power, and accepts the final strategy of the arrangement of the distributed computing power to be implemented. And the edge network management layer reports the edge network status information to the cloud network management layer, and accepts the distributed network resource arrangement information.

When the edge management platform receives an application request from a user, it will forward the request to the cloud management platform after verifying the user.
The cloud network management layer reports the network information to the computing power scheduling layer, which receives the network information, informs the computing power service and perceives the computing power status through the computing

power
  scheduling layer, so as to generate the computing power route and monitor the route
  in real time. The cloud management platform sends the generated computing power
  arrangement information to cloud network management. Cloud network management will
  distribute the received computing power arrangement information to edge network management.
  Edge network management reports network information and computing power scheduling
  information. Edge computing power scheduling performs computing power routing forwarding
  operations, and issues the final computing power scheduling strategy through computing
  power routing addressing, computing power routing notification.
Computing resources protect CPU, GPU and other resources to meet the different requirements of
 general computing and proprietary computing. These computing resources have greater advantages
 in parallel computing efficiency and low latency computing performance.
Heterogeneous computing s
olutions are deployed based on business requirements to meet the performance requirements of computing
, I/O, and network intensive applications.

4. Manageability Considerations

TBD

5. Security Considerations

TBD

6. IANA Considerations

This document requires no IANA actions.

7. References

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

Acknowledgments

TBD

Authors' Addresses

Zhengjie Sun
Beijing University of Posts and Telecommunications

Email: sunzhengjie@bupt.edu.cn

Hui Yang
Beijing University of Posts and Telecommunications

Email: yanghui@bupt.edu.cn

Chao Li
Beijing University of Posts and Telecommunications

Email: lc96@bupt.edu.cn

Sheng Liu
China Mobile

Email: liushengwl@chinamobile.com

Haomian Zheng
Huawei Technologies Co

Email: zhenghaomian@huawei.com