Industrial Internet of Things
Internet-Draft
Intended status: Informational
Expires: November 20, 2021

C. Tang
Chongqing University
S. Ruan
B. Huang
H. Wen
X. Feng
ChongQing University

May 19, 2021

Research on multipriority scheduling technology for real-time interconnection between industrial field data and cloud information
draft-tang-iiot-industrial-scheduling-04

Abstract

This document describes the multipriority scheduling technology for the interconnection between industrial field and cloud data in the application of 5G communication. The technology includes spectrum resource scheduling based on 5G slice in the process of accessing industrial data and task collaborative scheduling based on edge computing.

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 November 20, 2021.

1.

The rapid development of 5G mobile communication is driven by different application scenarios and diversified service deployment. Industrial Internet based on 5G technology has also accelerated research and deployment. In maximizing the role of 5G technology in the industrial system, the priority is to realize the interconnection between the industrial site and the cloud. On the one hand, the spectrum resources for 5G access are limited; on the other hand, the constraints of industrial equipment computing resources prompt factories to unload a portion of their industrial applications to computing systems with sufficient computing resources, such as microservers, cloud servers, or data centers. Therefore, the multipriority scheduling between industrial factory data and cloud computing is an important issue to be solved.

In the industrial environment, industrial factory data mainly refer to the real-time data generated by industrial production equipment and target products under the operation mode of the Internet of Things. These data include the those reflecting the operation state of equipment and products, such as the operation and operation conditions, working conditions, and environmental parameters. These data can be uploaded to the cloud for data processing and analysis through 5G base stations. The data can then be reused by factories for intelligent design, intelligent production, networked collaborative manufacturing, intelligent service, and personalized customization.

In the future smart factories, the demand for industrial field applications, including Internet of Things data acquisition, intelligent robots, industrial augmented reality (AR), and other services, is expected to increase. As applications have different demands for network service quality，multipriority scheduling technology should be studied on the basis of service quality.

From the industrial factory to the cloud computing center, the priority scheduling problem can be decomposed into two parts. The first part includes the allocation of spectrum resources corresponding to the schedule for different priority services in the process of industrial data access through 5G communication technology. For this purpose, we propose an uplink scheduling scheme

for industrial field data based on 5G slice. The other part includes the allocation and scheduling of computing resources for tasks in edge computing nodes and cloud computing nodes. For this purpose, we propose a collaborative scheduling scheme for big data tasks in industrial fields based on edge computing.

1.1.

1.1.1.

The Internet of Things is the Internet of everything. The Internet of Things data contain all types of essential information, such as sound, light, heat, electricity, mechanics, chemistry, biology, and location. Its goal is to combine all types of information from sensing devices with the Internet to realize the interconnection of people, machines, and objects at any time and place. Therefore, massive machine communication brings great demand for network coverage. The information collection of industrial control systems includes the sensor data collection designed in these industrial systems. These systems mainly collect the physical events and data generated in industrial production and manufacturing factories, including various physical quantities, identification, positioning, and other data.

1.1.2.

Machine vision has become increasingly popular in manufacturing enterprises, such as automobile factories, because of its effectiveness in detecting product defects. This type of application requires a large network bandwidth. As intelligent robots require complete corresponding intelligent operations, the fast response of a highly reliable 5G network is also a prerequisite.

1.1.3.

5G and AR are projected to become important applications of the Industrial Internet. The combination of 5G and AR can be applied to multiple scenes in industrial factories, including man-machine collaboration, monitoring of production processes, pre-job training for new employees, product quality detection, and remote assistance and guidance. For example, when industrial equipment is damaged and needs to be repaired, remote technicians can control the robot remotely through AR to complete the maintenance process. In such a case, the industrial network needs to provide a reliable network bandwidth and address low latency communication requirements.

1.1.4.

The industrial field has other business needs, including security monitoring. Different businesses have different requirements for network services.

2.

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 ERROR: Undefined target: RFC2119.

3.

In the access stage from the industrial field to 5G New Radio (NR), the base station side needs to perform multipriority uplink scheduling for different service data and allocate spectrum resources because of the service demand and communication quality of different data on site.

In this work, interslice scheduling and intraslice user scheduling are used for uplink resource allocation. In terms of satisfying the service quality required by different 5G slice scenarios and different industrial field data, it can effectively improve the fairness and throughput of scheduling.

3.1.

First, when UE needs to send uplink data, it puts the required data into the cache and then submits its buffer state report to the base station through the physical uplink control channel. At the same time, the scheduling request is sent to inform the base station gNB (5G base station) that it needs to send data.

Second, the uplink scheduler of gNB receives an uplink scheduling request from UE, and gNB allocates resources to UE on the basis of UE's cache status report and the uplink channel condition of UE. The uplink channel status of UE is obtained by the sounding reference signal that UE periodically sends to gNB. The allocation results are sent to UE via the physical downlink control channel by using the uplink grant.

Third, UE uses the resources allocated by the base station to send data to the base station through the physical uplink shared channel.

The uplink scheduler of gNB receives the cache status report and upline channel status of UE and completes the dynamic scheduling of

time-frequency resources according to the built-in scheduling algorithm. There are three common scheduling algorithms: Round-Robin(RR) algorithm, Max C / Ι algorithm, Proportional Fairness(PF) algorithm.

The uplink scheduler of gNB receives the cache status report and the upline channel status of UE and completes the dynamic scheduling of time-frequency resources according to the built-in scheduling algorithm. The three most common scheduling algorithms are the round robin (RR) algorithm, max C/Ι algorithm, and proportional fairness(PF) algorithm.

The RR algorithm allocates resources for different users of request scheduling in a circular way. This algorithm only considers the fairness among users and loses the system throughput. The max C/I algorithm always provides resources for the best users of the channel. It can maximize the system throughput, but it cannot guarantee the fairness between cell users. The PF algorithm considers the ratio of instantaneous rate and long-term average rate when selecting users. It adjusts different users by using weight values to achieve the purpose of consideration of the overall throughput of the system and fairness of users. However, it does not consider quality of service (QoS) information.

The explosive growth of data rate and capacity demand, as well as the large-scale, high reliability, low delay, and other differentiated demands, have brought about the development of 5G. Therefore, in the face of different industrial scenarios and different QoS requirements of businesses, a highly reasonable multipriority scheduling scheme needs to be designed. Under the condition of limited wireless resources, a reasonable scheme allocates wireless resources for different 5G slices to meet the service requirements of high-priority services in intelligent manufacturing plants and improve the resource utilization rate and fairness among users as much as possible.

This section presents a multipriority resource scheduling method for industrial field data to improve resource utilization as much as possible and thereby meet the service quality required by different industrial field services under different 5G slice scenarios.

3.2.

The general flow of the uplink scheduling scheme based on 5G slice is as follows:

Step 1: During a scheduling cycle, determine if the task cache queue is empty. If it is null, then wait for the next scheduling cycle; otherwise, proceed to the next step.

Step 2: Use the interslice scheduling algorithm to allocate resources to the three network slices according to requirements.

Step 3: For the resources obtained from each slice, perform resource scheduling for each user in the slice. Upon completion, wait for the next scheduling cycle.

3.3.

For Step 2 of the uplink scheduling scheme based on 5G slice, interslice resource scheduling needs to meet the following:

1. The resources obtained from different 5G slices are isolated and independent in the frequency domain, and they can be adjusted flexibly. The congestion of one 5G network slice does not affect the other 5G network slices.

2. By allocating spectrum resources with good channel conditions to a high-priority slice, the throughput of the system and the service guarantee for high-priority services can be improved.

Assume that the number of resource blocks (RBs) that the scheduler can configure is "Q" and that at the time t of scheduling, the total number of users requesting resources is N. In this work, the priority of each slice in each RB is defined in formula (1):

where P(i,j) represents the priority of the i-th slice at the j-th RB in one scheduling cycle. The greater P(i,j) is, the higher the scheduling priority of the i-th user in the j-th RB will be. In improving the system throughput, the priority is based on the rate and calculation of all users in the j-th RB in the i-th slice. R(i,j) represents the rate sum of all users in the j-th RB in the i-th slice. According to the calculation, we can obtain the priority matrix.

For the service requirements of different businesses using industrial field data, the uRLLC slice needs low delay and high reliability, such as those exhibited by a real-time remote cooperative robot, which needs to prioritize the allocation of resources and reduce queuing delay. The eMMB slice has high data volume business requirements. Hence, the priority allocation of resources will improve the overall network throughput. However, its delay requirements are not as high as those of the uRLLC slice. The mMTC slice has the lowest scheduling priority, and in most cases, the amount of uplink data is not large, and the delay requirement is low. Therefore, given the order of the uRLLC slice, eMMB slice, and mMTC slice, RB resources are configured according to the priority matrix.

According to the resource scheduling requirements of slices, the final interslice resource scheduling scheme is as follows:

Step 1: Calculate the priority matrix according to formula (1).

Step 2: According to the priority order of slices, select the i-th slice for resource scheduling, and initialize i to 1.

Step 3: Select the i-th slice and the j-th RB with priority ranking, and initialize j to 1.

Step 4: Determine whether the currently scheduled RB is adjacent to the RB allocated by the slice. If yes, then perform step 5. Otherwise, set j = j + 1 and repeat step 3.

Step 5: Assign priority j-th RB to the i-th slice and remove the RB from the RB sequence.

Step 6: Determine whether the resource request for the i-th slice has obtained enough resources. If so, then perform step 7. Otherwise, set j = j + 1 and repeat step 3.

Step 7: Determine whether the RB sequence is empty or whether all slices have obtained sufficient resources. If so, then end slice resource scheduling. Otherwise, execute i = i + 1 and repeat step 2.

3.4.

After resource scheduling among slices, the slices obtain their respective continuous and isolated RB groups. It does not interfere with the intraslice user scheduling.

The user scheduling in the slice can be understood as a logical cell, and the scheduler conducts resource scheduling on the users belonging to this cell through the resources obtained by intraslice scheduling. Given the different QoS requirements of data in the 5G industrial field, the performance indicators that need to be comprehensively considered in the process of user scheduling in the slice include transmission rate, delay demand, packet loss rate, and the amount of data to be transmitted.

The priority of user i in the j-th RB group at time t is calculated by formula (2):

p(i) represents the maximum rate of packet loss of user i, i.e., p(i) in (0, 1). Therefore, -log(p(i)) indicates that the lower the maximum packet loss rate is, the higher the priority of user will be. Td(i) represents the maximum wait delay for user i. The smaller

Td(i) is, the higher the user priority will be. r(i, j) represents the instantaneous transfer rate of the i-th user in the j-th RB group during the t scheduling cycle. R(i) represents the average transmission rate before the i-th user. The higher the instantaneous transmission rate is, the better the channel quality condition is, and the higher the priority is. d(i) represents the amount of data that user i is waiting to send at time t. The higher the value of d(i)/D is, the higher the proportion of the pending business volume of user i in the total business volume of all requesting users at time tis, and the higher the priority is.

Therefore, the intraslice user scheduling scheme is as follows:

Step 1: Complete the interslice scheduling.

Step 2: For all RBs of a single slice, they are divided into RB groups of the same size according to the number of slice users.

Step 3: Calculate the priority of each user in the slice on each RB group according to formula 2.

Step 4: Assign the RB group with the highest priority to each user in turn according to user priority.

Step 5: Determine whether the RB sequence is empty. If so, then end the resource allocation. Otherwise, repeat step 4.

3.5.

The scheme proposed in this section has the following advantages:

1. It analyzes the different service requirements for industrial field data in a 5G environment. The interslice scheduling scheme is used to complete the resource allocation of three 5G slices to ensure the flexible scheduling and isolation of resources between slices. The scheme also ensures that the required resources for high-priority businesses, such as the uRLLC slice business, are allocated and improve the throughput of the system.

2. The intraslice scheduling comprehensively considers the data service transmission rate, delay demand, packet loss rate, data volume to be transmitted, and other performance indicators. It also gives priority to the scheduling of users with good channel conditions, high delay requirements, high-reliability requirements, and large data volume to be sent.

4.

With the rapid development of industrial Internet and mobile communication technology, applications such as face recognition, short video traffic, autonomous driving, drone operations, industrial detection, and others have high requirements for computing. However, relying only on the current centralized cloud computing architecture model does not meet the required computing power of businesses. Amid the continuous generation of big data in industrial production, entertainment, education, and other industries, cloud-centric computing architecture should rely on new distributed computing architecture, such as edge and fog computing, to alleviate computational stress.

Correspondingly, the emergence of big data poses a challenge to the improvement of the performance of end devices. According to the type of data and service quality requirements, higher requirements are put forward for computing speed and processing capacity. Increased endpoint computing power also provides a good boost to distributed computing architecture. For example, some tasks with particularly high latency requirements are suitable for distributed processing mechanisms with the help of high-performance end devices because by relying only on cloud processing, real-time performance cannot be met under heavy network load. Therefore, edge computing needs to be utilized to sink computing power and dynamically allocate computing resources on the basis of tasks and real-time performance. We indeed know the importance of high computing power and effectively distributed computing architecture. However, a number of challenges exist in the industrial field environment; these challenges include varied sensor data, the corresponding instruction requirements generated by devices, and service processing. To a certain extent, the computing power of field devices is insufficient as well, and data present complex characteristics. Relying solely on edge terminal equipment to implement business logic is difficult. In complex industrial sites, the response requirements of various services are inconsistent. Thus, the system's response capabilities, processing capabilities, and throughput capabilities are put to the test. Thus, end devices and servers should be combined to achieve good collaboration.

Therefore, in the industrial Internet big data scenario, the collaboration of tasks, the allocation of resources, and the efficient processing of data offer a huge research space and important practical value, thereby attracting the attention of many scholars. However, the traditional research point focuses on resource allocation and the utilization of edge nodes, the coordinated scheduling of the edge and the cloud, and the issue of task priority. However, these algorithms have drawbacks. For

example, they consider either the resource allocation problem between nodes or the task priority problem, and they do not finish the coordinated consideration. Moreover, the network link bandwidth exerts an impact on the system. Hence, an improved task scheduling algorithm should be proposed on the basis of task requirements, along with edge computing, network bandwidth, and real-time task requirements, to maximize resource utilization and user satisfaction.

4.1.

On the basis of the complex industrial site environment, task requirements, and the different amounts of calculation, we take on a new perspective as we comprehensively consider computing resources and user satisfaction through edge computing technology. The goal is to perform tasks between the terminal and edge server resource scheduling problem. In the case of meeting the minimum resource requirements, user satisfaction can also be guaranteed to meet the real-time task processing generated in a variety of industrial production processes.

According to the actual needs of industrial field task data, we considered and provided a collaborative scheduling algorithm for industrial field big data tasks based on edge computing. The steps are as follows: the terminal publishes the service, and the scheduler obtains the service information and calculates the number of tasks and delay requirements. According to the task delay requirements, a task's urgency and priority are evaluated. Subsequently, whether the task should be offloaded to the edge server is determined according to the current bandwidth resources. Meanwhile, the task cache status of the scheduler and the computing information of each edge server are obtained. On the basis of the system resource status, number of queued processes, current business task volume, delay requirements, etc., the reasonable task scheduling of the server and terminal is performed. The process is repeated until all tasks are allocated and executed.

The specific steps are as follows:

Step 1: The terminal publishes the service, and the scheduler obtains the service information, such as the number of calculation tasks N(i) and the delay requirements T(i) (i with (1,N)).

Step 2: The task priority LEVEL(i) (i with (1,N)) is evaluated according to the time delay requirements of the task. It exerts an impact on the subsequent task scheduling and further reflects user satisfaction.

Step 3: The current network link information, that is, the remaining bandwidth of the network R, is obtained. Assuming that the upload rate of the task is α, the task upload needs to meet. If the current bandwidth is not enough, then the task needs to be unloaded, and the waiting delay is recorded.

Step 4: The task scheduling threshold of the scheduler is obtained, along with the calculation information of each edge server, the number of tasks queued, and the queue's waiting delay.

Step 5: The reasonable task scheduling of the server and terminal is performed according to the state of system resources, current business task volume, and delay requirements.

4.2.

The following is the processing flow of the scheduling strategy.

We consider the task delay requirements, the remaining capacity of the terminal, the remaining computing capacity of the edge server, and the total delay of the task assigned to the terminal as the input of the machine learning algorithm. The results of the former calculation are then fed to the fully connected network layer, and the output layer is maximized through the fully connected layer. The softmax layer estimates the probability of assigning to the terminal or the edge server. Therefore, the internal parameters of the network are learnable parameters so that it can perform adaptive adjustment to provide a basis for subsequent optimization on the basis of the customized loss, system resource conditions, network load, and optimization goals. The number of iterations can be determined accordingly. The updated parameters are used to allocate real resources to the tasks in the current scheduler.

4.3.

Relative to the existing task resource collaborative scheduling algorithm in the industrial field, the algorithm proposed in this work has the following advantages:

It can realize the reasonable scheduling of industrial field tasks in terminal equipment and edge servers, fully consider the system's computing resources, improve the system's ability to process tasks, and minimize the resource consumption of the system. It can also avoid calculation delay caused by an unbalanced resource allocation. The algorithm considers the execution status of the system, the task calculation amount, and the delay requirements for optimal scheduling when performing task scheduling. It also comprehensively considers

the construction of new optimization goals to achieve system resource utilization efficiency and user satisfaction.

5.

5.1.

For edge computing equipment, security problems are caused by indirect or self-inflicted causes during operation (e.g., energy supply, cooling and dust removal, and equipment loss). Although threats to operations are not as devastating as the damage caused by natural disasters, the lack of a good response will still lead to disastrous consequences, resulting in the performance degradation of edge computing, service interruption, and data loss. Particularly in the Industrial Internet scene, factories conduct sophisticated equipment maintenance and overhaul, but dealing with the operation and maintenance of IT equipment timely is difficult.

5.2.

Relative to cloud computing data centers, edge nodes have limited capabilities and are highly vulnerable to hackers. The damage of a single edge node is not extensive, and the network can quickly find alternative nodes nearby. However, if hackers use the compromised edge nodes as "broilers" to attack other servers, then they could affect the entire network. Most existing security protection technologies have complex computational protection processes, which are not suitable for edge computing scenarios. Therefore, an important network security requirement is to design lightweight security technology suitable for edge computing architecture in the Industrial Internet scene.

5.3.

In edge computing, users outsource data to edge nodes and transfer the control of data to them. The process introduces the same security threats as cloud computing. First, ensuring the confidentiality and integrity of data is difficult because the outsourced data may be lost or modified incorrectly. Second, unauthorized parties may misuse the uploaded data to seek other benefits. Relative to the cloud, edge computing avoids the long-distance transmission of multiple routes and greatly reduces the outsourcing risk. Therefore, the security problem of data belonging to edge computing is increasingly prominent. For example, in such a complex and changeable environment, the safe and rapid migration of data after the collapse of an edge node should be realized.

5.4.

Application security, as the name implies, guarantees the security of application processes and results. In the era of marginal big data processing, applications can be guaranteed to get short response times and high reliability by moving application services from cloud computing centers to network edge nodes. Meanwhile, network transmission bandwidth and intelligent terminal power consumption can be greatly reduced. However, edge computing suffers from common application security problems in information systems, such as the denial of service attack, unauthorized access, software vulnerability, abuse of authority, and identity impersonation. Moreover, it has other application security requirements because of its characteristics. In the scenario in which multiple security domains and access networks coexist at the edge, managing user identity and realizing authorized access to resources become important in ensuring application security.

6.

This memo includes no request to IANA.

7.

We thank all the contributors and reviewers and are deeply grateful for the valuable comments offered by the chairpersons to improve this draft.

8. References

Authors' Addresses

Chaowei Tang
ChongQing University
No.174 Shazheng Street, Shapingba District
Chongqing 400044
China

Email: cwtang@cqu.edu.cn

Ruan Shuai
ChongQing University
No.174 Shazheng Street, Shapingba District
ChongQing
China

Phone: +86 189-6826-0296
Email: rs@cqu.edu.cn

Huang Baojin
ChongQing University
No.174 Shazheng Street, Shapingba District
ChongQing
China

Email: baojing-huang@foxmail.com

Wen Haotian
ChongQing University
No.174 Shazheng Street, Shapingba District
ChongQing
China

Email: wenhaotianrye@foxmail.com

Feng Xinxin
ChongQing University
No.174 Shazheng Street, Shapingba District
ChongQing
China

Email: xxfeng@cqu.edu.cn