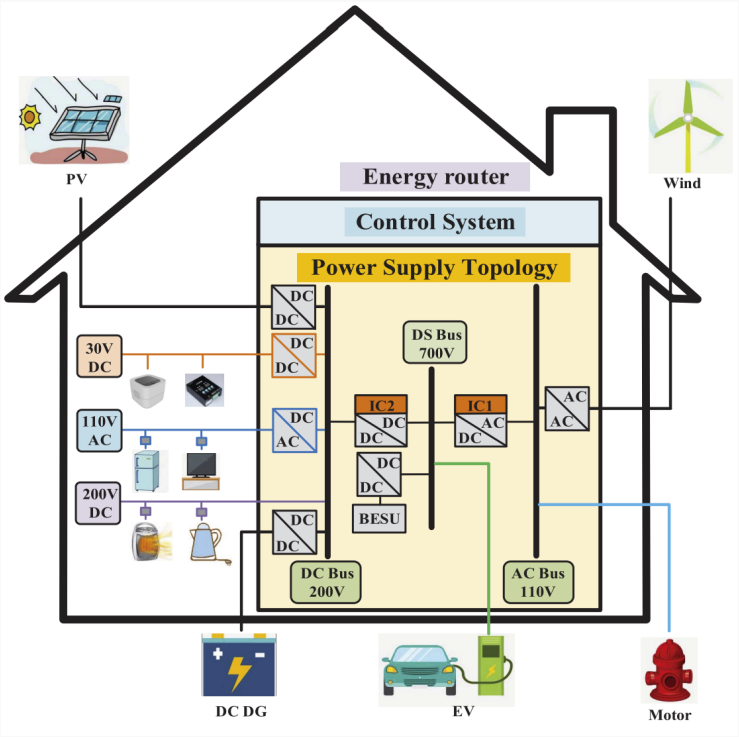
IEEE CONSUMER TECHNOLOGY SOCIETY (CTSOC)

ISSUE: DECEMBER2022

CTSOC-NCT NEWS ON CONSUMER TECHNOLOGY



A schematic diagram of the proposed energy router system for improving the renewable energy consumption and power supply flexibility.

TABLE OF CONTENTS

EDITOR'S NOTE

On behalf of the Editorial Board of IEEE CTSoc News on Consumer Technology (NCT) editor-in-chief Wen-Huang Cheng and editors, Luca Romeo, Jianlong Fu, Loh Yuen Peng and Chuan-Ju Wang, I am happy to introduce the December issue of the NCT in 2022.

This issue starts with a cover story which shows a schematic diagram of the proposed energy router system for improving the renewable energy consumption and power supply flexibility published in the CTSoc's journal, IEEE Transactions on Consumer Electronics. This paper presents a nine-port energy router topology to meet the requirement for hardware scheduling equipment in smart homes. The simulation and experimentation confirmed the viability of the energy router.

Next, the feature people provide an interview with Prof. Ir. Dr. Chee Seng Chan with the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia. The research of Prof. Chan aims to protect the ownership of deep learning or AI models and data. The interview shows the reason, current issues and importance of data and AI models protection and appeals to have some policies to govern AI so it creates some kinds of fear in certain communities.

Finally, this issue presents a featured article brought by Prof. Zhihui Lu of Fudan University, China, discussing on the task scheduling in a cloud-edge hybrid architecture which is a challenge in research and CE industry. The paper summarized the task scheduling algorithms with considering two major factors: restricted resources and service requirements. It explained the proposed dynamic collaborative scheduling method under the heterogeneous cloud-edge architecture, and their proposed data placement and retrieval mechanism under the cloud-edge hybrid architecture to cooperate with the execution of the scheduling method. These proposals reduce the response time of the entire system and improve the user's application experience. This article also summarized thire collaborative training method of models under the cloud-edge hybrid architecture.

We hope you can enjoy your reading!

Yafei Hou
Editor of NCT

ARTICLE TITLE
The Energy Management of Multiport Energy Router in Smart Home

AUTHOR(S)
Rui Wang, Shaoxu Jiang, Dazhong Ma, Qiuye Sun, Huaguang Zhang, and Peng Wang

JOURNAL TITLE
IEEE TRASACTIONS ON CONSUMER ELECTRONICS

JOURNAL VOLUME AND ISSUE
Volume: 68, Issue: 4

DATE OF THE ARTICLE
August 2022

PAGE NUMBERS FOR THE ARTICLE
344 - 353

Although the smart home has attracted a lot of attention recently, many academics place a greater emphasis on energy dispatch technology than on energy optimization strategy. This research presents a nine-port energy router topology to meet the requirement for hardware scheduling equipment in smart homes. In order to boost the consumption of renewable energy, the energy router achieves energy complementarity between wind, solar, and storage. It also comes with five voltage level output interfaces to expand the flexibility of the power supply system. Additionally, the entire topology utilizes an AC/DC hybrid structure made up of built-in power electronic converters, allowing for the minimization of power supply hardware. Then, with a focus on the energy management of the energy router, a hierarchical management control strategy is provided to realize the energy flow balance through multimode management. This strategy includes decentralized module control and power dispatch control. Finally, simulation and experimentation are used to confirm the viability of the energy router.

INTERVIEW WITH PROF. IR. DR. CHEE SENG CHAN

Prof. Chan is a Full Professor with the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia. He is leading a research team that specializes in computer vision and machine learning where his team has published more than 100 papers in top peer-review conferences and journals. From 2020-2022, he was seconded to the Ministry of Science, Technology and Innovation (MOSTI) of Malaysia as the Undersecretary for Division of Data Strategic and Foresight, as well as the Lead of PICC Unit under COVID19 Immunisation Task Force (CITF). He was the recipient of Top Research Scientists Malaysia (TRSM) in 2022, Young Scientists Network Academy of Sciences Malaysia (YSN-ASM) in 2015 and Hitachi Research Fellowship in 2013. Besides that, he is also a senior member of IEEE, Professional Engineer (BEM) and Chartered Engineer (IET). Prof. Chan is also the founding Chair for IEEE Computational Intelligence Society, Malaysia chapter, and is currently an Associate Editor of Pattern Recognition (Elsevier).

What is the current focus of you and your team's research?

The current focus of my research team is more on the security, in particular we are very interested to look into how to protect the ownership of deep learning models. This is because as a researcher, we know that it is not easy to train a successful and commercially viable Deep Learning model. For instance, we spend a lot of time, money, and resources to do so, but at the moment, there is no ownership protection at all on this very valuable deep learning models. That is why recently we have been working with WeBank (a private online bank founded by multiple Chinese companies including Tencent Holdings Ltd.) on looking into the possibility to create a technology to protect the ownership of deep models.

Our approach to this problem is not application centric at the moment because we are targeting types of Deep Learning models in general. We started off with protecting the deep learning Convolutional Neural Network (CNN) and after that we move to Generative Adversarial Network (GAN). The main reason is that in common CNNs and GANs, the inputs and outputs are totally different. In CNNs, you usually input an image and get a classification result while for GANs you would input a latent noise and the output would be

an image. Some CNN and GAN models would eventually have images as input and output as well, but we mainly look at the common difference between their inputs and outputs for our work. We also further extended our work to RNN recently, with the same intuition whereby a common RNN input is a string of text and the output is also a string of text.

So the way we apply our techniques to CNN, we may not be able to apply to GAN in a simple manner due to the nature of the input, and seems that it cannot be applied to RNN as well. So that is why at the moment we are not looking at application centric but rather at the nature of the deep learning model itself. As of now, we have covered most of the basic architectures that is available in the literature including, models where the input is an image and the output a regression, models where the input is an image and the output is an image, the input is a noise latent vector and the output is an image, and lastly the input is a text the output is also a text. As a summary, we have covered most architectures that is popular in this domain.

From your experiences, how has the research landscape changed from when you first started till now?

I think with the emergence of Deep Learning we can see that there are exciting models that can be put into the commercial market compared to last time. This is because of the power of the deep learning algorithms and also the power of the digital data that is available now compared to last time. Now almost every device that we use is digital-based, no longer analog. Because of this, we are getting more and more data now. So [in terms of visual data] from the "classical" 2D image, now most people have gone to work on video, and also probably beyond 3D. These had been the changes [I've seen throughout the years].

With many research as well as industrial advancements claiming to incorporate or innovate with AI, what is your perspective on this trend?

Now the industry has been very excited due to the financial gain that they can get. I would say in overall, the industry nowadays is very excited about the emergence of Deep Learning but not a lot of industries eventually can sustain for a long time because of the lack of effort to build their own deep model. You need a lot of financial resources in terms of hiring the correct people and then have enough infrastructure to build a deep model. So at the moment, almost all the current start-up companies have no issues for the first two years of business because they can rely on open-source codes to have their customer base. However, once they move to year 3 and above, which I call the "sustainable time", they might have a hard time because of competition where they may be no different to their rivals [that continue using open-source solutions]. At this time, the company would have also grown and the financial requirement to sustain the company is higher now compared to when they just started. So the challenge for most of the companies eventually start in year 3. That is why recently, we also can see that a lot of technology companies have retrenchment because it is no longer viable to invest a lot of resources. We can see that the benefits or the return on investment (ROI) is quite marginally small from one to another. That is why companies now would need to look at how to really sustain in the current competitive market.

(Quite a lot start on trendy type of approach and go for open source because they thought that is what they are going to use)

Yes, that is why for those start-up companies, you can see that those very successful start-ups eventually do have their own algorithms. Those that are able to sustain beyond year 3 and above, we always see that they not just have their own technology, but they are also pursuing research in their area. An example of successful start-up is YOLO (You-Only-Look-Once deep learning object detection). YOLO is from a start-up company and

now we know there are up to YOLOv7 [for their object detector], so the researchers are still improving their "product" and they are still active in research including publishing research papers, so on and so forth.

I think publishing now is not like the conventional where it is only centric on academia, but companies now, in order to survive long enough, they need to start to look into some research element so that they can be differentiated even though very slightly from their rivals. This is so that they can be unique in their own domain to attract unique customers. We all know that deep learning solutions is not one-size-fits-all as it has always been customized to a certain application and if you have your own technology in a particular domain, then you will always be able to attract your type of customers that will eventually become your loyal customer in the future. So that is why, to have your own technology, and then move on to improve your technology is very important.

So the main message here is that you would definitely need to do (at least some) research. A lot of people always say that research papers are only for academia but it is no longer true because when you are able to publish your papers in very reputable conferences, from a company's perspective, it can give investors and clients more confidence in the company. This is because publications eventually tell the world that you may already have a new technology within the company, just that due to some kind of constrain such as hardware, the solution is not yet efficient or feasible financially to be deployed. The publication is one of the proof that the company already have a future solution waiting. I think this is very important but has been overlooked by a lot of start-up companies, particularly in Malaysia.

Nowadays, many tools and resources are open and accessible to anyone and everyone to devise their own AI projects and solutions, would this be a concern especially from the ethical side of things?

I think this is very important when it comes to ethics, and a lot of people have shown that it is possible to use technology [for ethics]. For example, in some of our work that we have shown, it is possible to use technology to protect deep models [a part of ownership ethics]. Some researchers in the domain of explainable AI have also shown that it is possible to use certain solution to make the model "more ethical" by knowing what is happening behind the scenes. But unfortunately, policy-wise is not ready. So that is the hiccup here because ethics are not merely just technological, you need lawmakers to come up with certain guidelines or policies.

At the moment, this is a work in progress because we know that it is not easy to come up with a policy that would be agreed by every parties and worse still the real understanding of any AI model is still in quite an infancy stage. In most cases, we still do not really know why an AI model behave in such a way. However, I do see efforts on improving ethics not only in terms of how we should use a model, but also on how should we use the data. For example, a particular set of data can be used to train a model for "good", but at the same time it can be used to train a model for "bad" purposes as well. How do we govern that? In my opinion, unfortunately, we are not there yet although there are some laws in Europe and also China for such protection in their early stages. However, these are mostly to tell that one would need consensus to use data, including the consent of the individual whose data is captured by even a CCTV, otherwise an alternative route away from the CCTV has to be provided by the developers.

In Europe, there is the General Data Protection Regulation (GDPR) that provides such protection to data. But what is still lacking is the governance of the AI's behavior. What if the AI acts

worse than expected, who would be the one responsible for the model's action? The company, the inventor, or eventually the user? At the moment we do not have a clear answer to this, so that is why there is still a long way to go when it comes to ethics.

(And because from the scientist perspective, understanding the AI also is still a work in progress so it is hard for the policy makers or even the public to even understand further.)

Exactly. As we all know, underlay of all these AI models are all algorithms achieved by some mathematics. But we also know that the real world environment keeps on changing from time-to-time and human behavior change as they age too. The data we use to train the model are typically historical data. There is no guarantee people that behave one way in the past will have the same behavior in the future. There are still a lot of such uncertainties in the real world environment where we try to apply algorithms on, so it may not be able to adapt well to this.

Most consumers would have used some form of AI technology and are focused on the convenience they provide, but do you think they are sufficiently aware of the potential negative implications that such tech would bring to their privacy?

I think everyone, even myself are very excited with all these technologies because no doubt that the advancement of technology has really improved our lives. It gives us more luxury to spend time with family and loved ones, and so on and so forth. But eventually none of us look into the [potential] negatives of it because we enjoy the benefits more than the so-called negative. However, we as an inventor or researcher in this area, we must be aware of every single possibility of negative impact that might happen so as to safe guard everyone.

We must know that, we have been enjoying the benefits of technology like AI because there are currently no rules and regulations. For example, the situation is just like driving on the highway. If there is no speed limit, users can drive as fast as possible and enjoy the shorter travel time from A to B. However, if an accident happens due to a lack of safety consideration, the consequences are very serious. Worse still is how to decide who to be held responsible in such a situation. From the consumer or user's point of view, they are only using the technology and service. A driver is only driving as fast as possible as there is no speed limit in place with the presumption that safety measures have been put in place by the service providers, such as road barriers, road quality, and then from the car manufacturers, a car that is well equipped with safety systems to handle collisions. This is the hypothetical situation that we have with AI right now.

We simply do not have any regulations around the word to dictate when A can or cannot be used. We also do not have limits on how far we can use the AI in a given population or conditions. That is why we can still enjoy a "free ride". The pro would be that the technology can improve a lot but the "side effect" would be, unbearable consequences if problems arise, especially from an ethical perspective. So I think what we need now is a check and balance.

There has been a saying from a movie, "Your scientists were so preoccupied with whether they could, they didn't stop to think if they should." Do you find that current day AI or computer scientists to have such a problem?

As a scientist and researcher, we would always be excited about technology, there is no doubt about that. But as I mentioned, it is because we have been enjoying the "free ride" [due to no regulations], we may eventually not think if we should do something or not because as a researcher, we go into it with a genuine attempt and a purpose. However, somehow, the work can be reversed and used against humans by someone else. So in such case, as I said, policies would be very important. There is no standardization so far and we have to be careful on what we choose to use or improve.

Based on your observations, has there been sufficient ethical awareness among computer scientists and the public?

I would say some of the researchers might be aware of this but definitely not the public. Probably 80% of the public is not aware because they are on the consumption side of things, enjoying the benefits and there is also not much information provided. There is a lack of awareness for example when they pass through an area with CCTV monitoring, and their data has been used as part of the training for an algorithm that no one had informed them about.

For some companies too where they are using face recognition systems and similar touchless mechanisms due to the pandemic, there has been no information revealed to consumers that their data has been captured by the devices and how long would it be kept. Matured industries like finance, the institutions have clear guidelines that details what would be the data used for, and how long it would remain with their servers, for instance 3 days, 5 days, and then it will be permanently deleted. There does not seem to be these kind of information related to AI.

Another simple example is social media. Often it does not seem that permission has been given to e-commerce platforms to trace our search history. But apparently, most of us has an experience where our search terms for a certain product in a search engine will somehow automatically appear as an advertisement in the social media platform as well as the e-commerce platform that we commonly use. To me, this is some kind of tracking that was done without my explicit permission [which is a breach of ethics]. Can I make a police report on this? It may not have any actions because based on current laws that I am aware of, there is nothing to be done. So as much as we want to use technology as much as possible, at the same time we also want to be protected. Unfortunately, the protection part is not there yet.

Has there been notable efforts to improve ethical accountability in computer science research?

I think there is a lot of efforts that is ongoing, such as the data protection act [of different countries]. We would start with data first, such as the European GDPR and a new California Privacy Rights Act (CPRA), and slowly move on from there. Unfortunately, policies are not something that can be done within a day. It usually takes years to be endorsed and passed by governments. I would say that it will be a long journey ahead because it seems to me, every country would have their own view on how data [and eventually AI] can and should be used.

Unity on this matter that can cut across the world will not happen in a short time or in the near future because we understand that there will always be disparity between developed, developing, and under-developed countries when it comes to technology. Also bear in mind that related policies will also need to involve cultural aspects as come matters are acceptable to certain countries and regions but not the others. So if we were to find a one-size-fits-all policy, it will be challenging. However, having some form of restrictions can be done soon. The community from academia and lead technologists are trying to help governments in various parts of the world to see to it. But as I said, it still needs time.

What should be the way forward since technology such as AI is ubiquitous to most consumers?

At the moment, technologies have shown and proved that they are very useful but now is the time for the policy makers to really sit down to come out with a check-and-balance. When this technology can be used and when it should be used? What are the terms and conditions to be set? If we do not do that soon, the more advance that we go, the harder it is for us to track all these changes. Especially nowadays where the world is on a full digital transformation. We really need to know now, and govern now or else it will be out of control very soon. So it is not about slowing down the technology but to catch up with the policy, that should be the way forward.

Any parting words for those who are following the development and trends of technology and AI?

For me I think, we definitely need technology to improve our lives. But having new technology is always an unpleasant thing for humans because humans always have a nature for reluctance of change. We can see through the industrial revolutions from 1st to 4th now. When a technology or a new invention has been put in place, always people have fear. For example, losing work opportunities and so on. If you look at the 2nd industrial revolution, from riding on a horse to the used of a car, people had reluctance on that. Up until now we have cars everywhere and then ride-hailing service like Uber, people also reluctant to that. In the coming future then might be worse with autonomous cars. So I think we need to look at the pros and cons on this. No doubt that when a technology an invention is been introduced, there will be a change of the job market but I would say that is how life is.

Regardless, all these technologies that have been put in place are always human-centric. It is the knowledge of humans that are represented by a set of algorithms, or mathematical equations. So a lot of technologies are always looking at humans. So I would say that as long as we humans are able to understand the works, and then having the policies to govern that, I think humans and technology can always work hand-in-hand to continue to improve human life. Just that unfortunately, at the moment we do not have a policy to govern AI so it creates some kind of fear in certain communities. Once this is resolve, I think everyone will know their position in the world, including technology. This is because now technology is everywhere, it is borderless without governance, yet humans are governed by certain policies which causes an imbalance. So we need to balance it so that humans would be able to enjoy more of this with less fear.

COLLABORATIVE SCHEDULING OF WORKLOAD TASKS IN A CLOUD-EDGE HYBRID ARCHITECTURE

Prof. Lu Zhihui
lzh@fudan.edu.cn
Zhihui Lu is a Professor at School of Computer Science, Fudan University. He received Ph. D from Fudan University in 2004, and he is a member of the IEEE and China computer federation's service computing specialized committee. His research interests are cloud computing and service computing technology, big data architecture, mobile edge computing, and IoT distributed system.

Xin Du
xdu20@fudan.edu.cn
Xin Du is a Ph.D student at School of Computer Science, Fudan University. His research interests are edge computing and distributed computing.

Abstract:

Task scheduling in a cloud-edge hybrid architecture is a challenge in research and industry. The task scheduling algorithm needs to consider restricted resources and service requirements. The restricted resources include hardware resources, network topology and latency, storage resources and data placement. Service requirements include service latency, data transmission time, etc. In the last few years, we not only design a dynamic collaborative scheduling method under the heterogeneous cloud-edge architecture but also propose a data placement and retrieval mechanism under the cloud-edge hybrid architecture in order to cooperate with the execution of the scheduling method. It reduces the response time of the entire system and improves the user's application experience. Besides, a collaborative training method of models under the cloud-edge hybrid architecture is proposed and designed. According to the types of tasks it handles, appropriate model deployment is selected and personalized model training is performed.

Why we need collaborative scheduling for workload tasks?

Recently, in academia and industry, the collaborative scheduling of workload tasks in the cloud-edge hybrid architecture is a key area of research at home and abroad. The latest research results have carried out joint research on task service deployment and task scheduling, but they are for applications with a large amount of data and a large amount of transmission. More importantly, the current joint scheduling mostly stays in the design of the model, the resource utilization method is

relatively rough, and the data placement, retrieval and model coordination need to be considered when the scheduling method is implemented.

In the heterogeneous cloud-edge hybrid architecture, the same task type may be deployed on different edge servers, and different edge servers may deploy different type of services. There are various task scheduling schemes, and different scheduling schemes have a significant impact on the completion time of tasks. When a user task request arrives at the edge layer, the system should be able to select a suitable instance on the edge server to respond to the task calculation according to the task attribute and the resource attribute at the location of the edge node (including the parameter configuration and data cache of the edge node), so as to ensure the real-time and validity of the calculation.

There are three problems involved here: 1. The edge server itself has limited resources and cannot deploy all types of tasks. The system should be able to perform edge-edge coordination according to the changing characteristics of user requests, support refined resource joint scheduling of hardware, and select appropriate resources to adjust service deployment to utilize the limited edge resources to meet the needs of user services. 2. The execution of tasks requires data support, and the unloading, placement and retrieval of data need to be reflected in the entire architecture, which can more comprehensively support the work tasks. 3. In the process of executing tasks interactively with the edge cloud, users need to consider the data security of the execution model and the wishes of the actual stakeholders, which is also one of the limitations of task coordination and scheduling. In the dynamic collaborative training of the model, the training results of each model in the node and the training data are used to jointly train the model, and the cloud computing power and edge node computing power have also been effectively used, realizing the dynamic collaborative scheduling of the edge cloud model, thus greatly improving the training efficiency and resource utilization of the model under the heterogeneous edge-cloud architecture.

How we make collaborative scheduling for workload tasks?

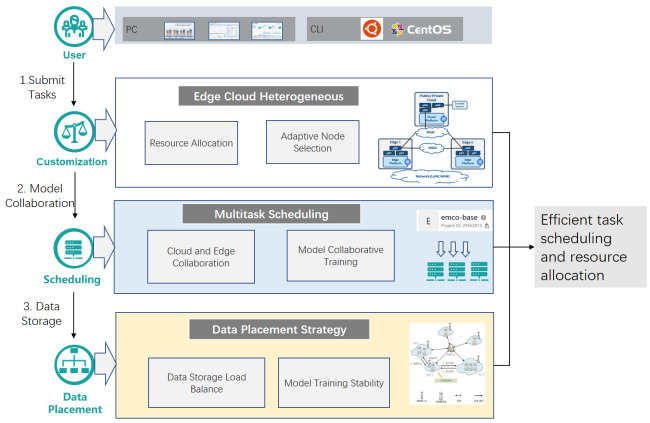


Figure 1: Technical route of our researches

In the last few years, we make some tries to solve the problem of low efficiency of the workload collaborative scheduling mechanism in the cloud-edge hybrid architecture. By proposing and implementing the collaborative scheduling method, combined with the innovation of the data placement and retrieval mechanism, it will be deployed in the cloud-edge hybrid architecture. The model in the system is collaboratively trained on multiple nodes, so that it can not only realize the refined joint scheduling of hardware resources in the system and improve the efficiency of resource utilization, but also meet their requirements for response time and data security from the user's point of view. We take into account the heterogeneous work task resource application characteristics of multi-edge nodes and the data storage and computing resource deployment of the cloud-edge hybrid system. Based on some open-source edge computing frameworks, a dynamic collaborative task scheduling algorithm is designed to fully integrate and utilize the system. It can reduce the average processing delay of tasks, improve user service quality, and meet the high real-time processing requirements of massive edge tasks. The data placement and retrieval mechanism under the cloud-edge hybrid architecture are proposed, and the data required for task scheduling is efficiently and safely used in an environment with high privacy protection requirements, and the computing power of heterogeneous edge nodes is effectively utilized to realize edge cloud. Model dynamic scheduling improves training efficiency and resource utilization, and improves system security

and privacy protection capabilities. To express the content of these studies more intuitively, the technical route of our research is shown as Figure 1.

A. Dynamic collaborative scheduling methods

We propose some dynamic collaborative scheduling methods under the heterogeneous cloud-edge architecture. On the basis of sensing the deployment status of services and data on edge nodes, select the corresponding edge nodes that can meet the task request delay submitted by users. The node with the smallest requirement and the smallest average completion time performs the cooperative processing of the task. By virtue of the fact that edge nodes are located at the edge of the network and are close to mobile terminal equipment, they can effectively reduce the delay of data transmission and better support the real-time nature of local services. At the same time, due to its strong computing power, the cloud system can command and dispatch the task deployment and computing resource allocation of edge nodes at the macro level. Our goal is to deploy the collaborative scheduling method in the KubeEdge, relying on its powerful container orchestration and scheduling capabilities to deploy tasks, support refined joint scheduling of resources at the hardware level, and improve computing resource scheduling performance.

We allocate the required computing resources for different user tasks, maximize the use of heterogeneous cloud-edge distributed characteristics, and processes tasks in parallel in the form of multi-task dynamic coordinated scheduling, reducing the response delay when edge computing processes jobs submitted by multiple users, optimize the processing efficiency of heavy tasks in heterogeneous cloud-edge networks. Based on the task processing model and resource management model, we design a task scheduling dynamic coordination mechanism and the task coordination scheduling mechanism. For the user's high concurrency and heavy workload requests, the cloud makes task scheduling decisions for incoming requests based on the resource capacity information of the edge server, the instance information of the services deployed on the edge server, and the instance running time.

First, according to the feature information of concurrent requests, the request-response sequence is calculated to form priority scheduling based on service deployment. In the case of meeting the time and resource requirements of the task itself, the edge nodes that can meet the delay requirements and can reduce the task completion time are selected from the edge nodes to perform task scheduling processing. When the edge node cannot respond to the task request, the service configuration recommendation algorithm is used to generate the relevant service configuration recommendation according to the spatiotemporal distribution characteristics of the task request, and the edge node is selected to deploy the relevant service. At the same time, a timing mechanism is designed. The edge nodes periodically send metadata to the central nodes for tasks and resources in the system, so as to realize task rescheduling, that is, to dynamically schedule tasks. Finally, a prototype verification system based on the scheduling component Dispatcher of the open-source project is designed to verify the effectiveness of the proposed cooperative scheduling algorithm.

The main function of the Dispatcher is divided into two parts: the first part is to schedule task requests generated by edge devices and find the optimal edge node for task processing; In the second part, when the required service is missing or cannot meet the user's QoS guarantee, select the appropriate edge node for service deployment. The performance reference indicators of the verification system include delay, deadline, QoS, energy consumption, economic indicators and other aspects. For the delay analysis, the service delay of the task can be obtained by weighting the processing delay of different nodes according to the distribution probability, and the delay optimization effect of the collaborative scheduling algorithm can be judged by analyzing the weighted average delay and the processing of time-sensitive tasks. In addition to minimizing the delay, the task deadline also indicates the urgency of the task. The delay sensitivity of different tasks is different. If some tasks are not completed before the deadline, there will be serious consequences, so they are defined as

hard deadline-constrained tasks. Otherwise, Soft deadline-constrained tasks. In addition to minimizing the delay, the task deadline also indicates the urgency of the task. The delay sensitivity of different tasks is different. If some tasks are not completed before the deadline, there will be serious consequences, so these tasks are defined as hard deadline-constrained tasks. Other tasks are defined as soft deadline-constrained tasks. If the completion time is greater than the deadline, there is a delay. The verification system describes the requirements of the task completion time according to the time-related efficiency function, and performs the final collaborative scheduling algorithm effect verification. The verification system also needs to consider QoS. According to the user's expectations for the application and the state of edge computing resources, it supports the refined joint scheduling of resources of the system hardware and improves the performance of computing resources.

B. A data placement and retrieval mechanism

We propose a data placement and retrieval mechanism under the cloud-edge hybrid architecture, which greatly reduces the response time of the entire system and improves the user's application experience. By considering the resource heterogeneity and dynamic variability of edge nodes, combined with the respective advantages of cloud data centers and edge nodes, a more reasonable data placement and retrieval mechanism is proposed, which fully considers the scalability of edge nodes. The speed of data placement and retrieval is improved, and the error rate of data retrieval is reduced.

The actual execution of the task is inseparable from the placement and retrieval of the corresponding data, which is often realized through heterogeneous edge computing systems. In our study it consists of a controller and several edge nodes. The controller manages the edge nodes in the system network and is generally not used to store data or indexes. An edge node consists of a network access point and its edge server, and the edge server deploys the network access point. In order to meet the differentiated demands of different regions of the edge layer, the distribution of resources among edge nodes is often uneven. Edge nodes can play the roles of storage server and index server at the same time. A storage server refers to a server that stores data items, and an index server refers to a server that stores indexes. Given an index key, it is easy to find an index server and find the corresponding index value in the DIT record of the index server that stores the data item.

C. A collaborative training method

We design a collaborative training method under the cloud-edge hybrid architecture. In heterogeneous edge nodes, according to the types of tasks they process, select appropriate model deployments and conduct personalized model training to achieve a trade-off between training speed and security. Refine the consensus on the training results of each model to collaboratively train intelligent models deployed on different edge nodes to improve model training efficiency and accuracy. When dealing with complex intelligent learning tasks, according to tasks of different difficulty, the models that have been deployed in the edge cloud are reasonably scheduled, the complex model in the cloud handles the more difficult tasks, and the lightweight model on the edge node handles the simple tasks, reducing the task processing delay and improving throughput.

Model collaborative training in traditional federated learning requires edge nodes to deploy the same model, which cannot be applied to the needs of heterogeneous edge nodes to deploy heterogeneous models for different tasks in the context of the Internet of Things. At the same time, for different types of tasks, complex models deployed in the cloud will increase communication overhead and task processing delays. Lightweight models deployed entirely by edge nodes are generally less accurate and efficient when dealing with complex tasks. Our research proposes and designs a model dynamic co-scheduling method in a heterogeneous edge-cloud environment. In the heterogeneous edge nodes, according to the type of processing tasks, select the appropriate model deployment, and carry out personalized model training. At the same time, the consensus on the training results of each model is

refined to collaboratively train heterogeneous models deployed on different edge nodes to improve model training efficiency and accuracy. When dealing with complex tasks, according to tasks of different difficulty, the models deployed on the cloud edge are reasonably scheduled. The complex model on the cloud handles the more difficult tasks, and the lightweight model on the edge node handles the simple tasks, reducing the task processing delay and improving the throughput.

The method proposed in our research makes full use of the training results of each model in the heterogeneous edge nodes and the collaborative training model, and the computing power of the cloud and the computing power of the edge nodes are also effectively used, realizing the dynamic collaborative scheduling of the edge-cloud model. It greatly improves the training efficiency and resource utilization of the model under the heterogeneous edge-cloud architecture.

What we have done for collaborative scheduling?

In our research, the paper ORHRC [1] (IEEE ICWS'20) placed the performance prediction module in the cloud layer based on the cloud-fog hybrid architecture, and placed the configuration selection module in the nodes of the edge layer. ORHRC processes the workload in the fog nodes and sends the characteristics of the workload to the cloud layer for modeling. The paper ARVMEC [2] (JPDC'20) uses an ensemble learning algorithm based on XGBoost to make accurate predictions on the workload performance of various VM types according to the objectives of different user needs. The paper [3] focuses on the application of edge service distribution strategy and proposes a novel edge service distribution strategy based on intelligent prediction, which reduces the bandwidth consumption of edge service providers and minimizes the cost of edge service providers. Based on the edge-cloud computing paradigm, Reference [4] not only constructs a data placement model that includes shared datasets within the individual and among multiple workflows across various geographical regions, but also proposes a data placement strategy (DYM-RL-DPS) based on algorithms of two stages. In [5], we design a microservice-based service deployment strategy to reduce the average latency of IoT devices in a mixed environment. Aiming at the heterogeneity of edge server capacity, dynamic geographic information of IoT devices, changes in device preferences for applications, and complex application structures, we first propose a method based on heterogeneity and dynamic characteristics in an edge-cloud hybrid environment. Reference [6] proposes an adaptive mechanism (ADST) for dynamic collaborative service deployment and task scheduling under a heterogeneous edge-cloud architecture. In the process of task scheduling and service deployment, in order to meet the request delay requirements, ADST uses a greedy way to make decisions, and selects edge nodes and instances to reduce the average completion time. [7] (IEEE ICWS'20) considers the existence of multiple scientific workflows and multiple cloud data centers when building the data placement model, and constructs a new data placement model. [8] (IEEE JIOT'21) proposed a blockchain empowered secure and incentivized federated learning (BESIFL) paradigm. [9] (IEEE ICWS'22) designs a novel blockchain-based intelligent edge cooperation system to make CEC effective, among which incentive and trust mechanisms and performance optimization are crucial for latency-sensitive service provision. [10](JDPC'22) takes into account the heterogeneity and scalability of edge nodes in real scenes when formulating an efficient indexing mechanism, and adopts a method that maps all edge nodes to points in a two-dimensional coordinate system according to the network distance. [11] (IEEE Access'22) propose an Online Pre-filtering Task Offloading System (OPTOS) that is able to mitigate the impact of vehicle mobility on task offloading performance in the edge network.

What is the next?

In the future, we will combine with some existing open-source frameworks and design fine-grained scheduling of resources and computing models. The data placement and indexing will be fully considered in the design, which lays the foundation for the collaborative training of the model in the architecture.

The corresponding collaborative scheduling algorithm, data placement strategy and specific retrieval mechanism will be proposed around the dynamic scheduling model and data placement model in the heterogeneous cloud-edge hybrid architecture. In addition, from the perspective of landing, the methods will support the refined resource joint scheduling of hardware and improve the performance of computing resources.

Finally, we will realize the prototype system based on the heterogeneous cloud-edge hybrid architecture, the dynamic scheduling method of resources and data, and the collaborative training strategy of the model, and ensure the security and accuracy of the model training.

References

[1] A. Xiao, Z. Lu, X. Du, J. Wu and P. C. K. Hung, "ORHRC: Optimized Recommendations of Heterogeneous Resource Configurations in Cloud-Fog Orchestrated Computing Environments," 2020 IEEE International Conference on Web Services (ICWS), 2020, pp. 404-412, doi: 10.1109/ICWS49710.2020.00059.

[2] Yajing Xu, Junnan Li, ZhiHui Lu, Jie Wu, Patrick C. K. Hung, Abdulhameed Alelaiwi: ARVMEC: Adaptive Recommendation of Virtual Machines for IoT in Edge–Cloud Environment. J. Parallel Distributed Comput. 141 (2020) 23–34.

[3] Y. Wang, X. Wan, X. Du, X. Chen and Z. Lu, "A Resource Allocation Strategy for Edge Services Based on Intelligent Prediction," 2021 IEEE 6th International Conference on Smart Cloud (SmartCloud), 2021, pp. 78-83, doi: 10.1109/SmartCloud52277.2021.00021.

[4] Xin Du, Songtao Tang, Zhihui Lu, Keke Gai, Jie Wu, and Patrick C. K. Hung. 2022. Scientific Workflows in IoT Environments: A Data Placement Strategy Based on Heterogeneous Edge-Cloud Computing. ACM Trans. Manage. Inf. Syst. 13, 4, Article 42 (December 2022), 26 pages. https://doi.org/10.1145/3531327.

[5] Chen L, Xu Y, Lu Z, et al. IoT Microservice Deployment in Edge-cloud Hybrid Environment Using Reinforcement Learning[J]. IEEE Internet of Things Journal, 2020.

[6] Y. Xu, L. Chen, Z. Lu, X. Du, J. Wu and P. C. K. Hung, "An Adaptive Mechanism for Dynamically Collaborative Computing Power and Task Scheduling in Edge Environment," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2021.3119181.

[7] Du X, Tang S, Lu Z, et al. A Novel Data Placement Strategy for Data-Sharing Scientific Workflows in Heterogeneous Edge-Cloud Computing Environments[C]//2020 IEEE International Conference on Web Services (ICWS). IEEE, 2020: 498-507.

[8] Yajing Xu, Zhihui Lu, Keke Gai, Qiang Duan, Junxiong Lin, Jie Wu, and Kim-Kwang Raymond Choo, BESIFL: Blockchain Empowered Secure and Incentive Federated Learning Paradigm in IoT, IEEE Internet of Things Journal, 2021.12 Early Access.

[9] X. Du, X. Chen, Z. Lu, Q. Duan, Y. Wang and J. Wu, "BIECS: A Blockchain-based Intelligent Edge Cooperation System for Latency-Sensitive Services," 2022 IEEE International Conference on Web Services (ICWS), 2022, pp. 367-372, doi: 10.1109/ICWS55610.2022.00061.

[10]Tang S, Du X, Lu Z, et al. Coordinate-based Efficient Indexing Mechanism for Intelligent IoT systems in Heterogeneous Edge Computing[J]. Journal of Parallel and Distributed Computing, 2022.

[11] J. He, Y. Wang, X. Du, Z. Lu, Q. Duan and J. Wu, "OPTOS: A Strategy of Online Pre-Filtering Task Offloading System in Vehicular Ad Hoc Networks," in IEEE Access, vol. 10, pp. 4112-4124, 2022, doi: 10.1109/ACCESS.2022.3141456.