informs
https://pubsonline.informs.org/journal/ijaa

INFORMS JOURNAL ON APPLIED ANALYTICS
Vol. 54, No. 1, January–February 2024, pp. 37–53
ISSN 2644-0865 (print), ISSN 2644-0873 (online)

THE FRANZ EDELMAN AWARD
Achievement in Operations Research

Huawei Cloud Adopts Operations Research for Live Streaming Services to Save Network Bandwidth Cost: The *GSCO* System

Xiaoming Yuan,[a,*] Pengxiang Zhao,[a] Hanyu Hu,[a] Jintao You,[b] Changpeng Yang,[b] Wen Peng,[b] Yonghong Kang,[b] Kwong Meng Teo[b]

[a] Department of Mathematics, The University of Hong Kong, Hong Kong SAR 999077, China; [b] Huawei Cloud, Huawei Technologies Co., Ltd., Shenzhen 518129, China

*Corresponding author

Contact: xmyuan@hku.hk, https://orcid.org/0000-0002-6900-6983 (XY); pengxiangzhao@connect.hku.hk, https://orcid.org/0009-0004-9843-7282 (PZ); hhy1224@connect.hku.hk, https://orcid.org/0009-0008-3784-0982 (HH); youjintao5@huawei.com, https://orcid.org/0000-0002-0699-9295 (JY); yangchangpeng@huawei.com (CY); pengwen1@huawei.com (WP); kangyonghong@huawei.com (YK); teo.kwong.meng@huawei.com (KMT)

Accepted: October 11, 2023

https://doi.org/10.1287/inte.2023.0079

Abstract. The rapid evolution of cloud computing technologies has instigated a paradigm shift across various traditional industries, with the live streaming sector standing as a compelling exemplification of this transformation. Huawei Cloud, which has become an influential player in the business-to-business live streaming arena, with its services spanning over 60 countries since 2020, is at the forefront of this shift. Amid the flourishing live streaming market, Huawei Cloud faces the dual challenge of satisfying the escalating demand, while managing the mounting operational costs, predominantly associated with the network bandwidth. To offer premium services while minimizing the bandwidth cost, we developed a dynamic traffic allocation system called *GSCO*. This system was engineered using an array of operations research methodologies such as continuous optimization, integer programming, graph theory, scheduling, and network-flow problem solving, along with state-of-the-art machine learning algorithms. The *GSCO* system has been proven highly effective in cost optimization, reducing network bandwidth expenses by about 30% and leading to savings exceeding $49.6 million from Q1 2020 to Q3 2022. In addition, it has significantly bolstered Huawei Cloud's market share, amplifying peak bandwidth from an initial 1.5 terabits per second (Tbps) to a substantial 16 Tbps.

Keywords: cloud computing • live streaming • traffic allocation • mixed-integer programming • linear programming • Edelman Award

Introduction

Cloud computing is an Internet-based computing paradigm that enables faster innovation and economies of scale, whereby flexible hardware and software resources, such as servers, storage, databases, networking, and analytics, are provided to customers on demand. Cloud computing prevailed in the early 2000s because information processing could be done more efficiently on a large shared pool of computing resources, and most IT companies have embraced it (Marinescu 2022). According to Precedence Research (2022), the global cloud computing market is expected to skyrocket to approximately $1,614.10 billion by 2030, exhibiting a robust compound annual growth rate of 17.43% from 2022. As one of the rapidly burgeoning regions in this market, China's market size grew to $48.73 billion in 2020 and the compound growth from 2020 to 2025 is predicted to be 26% (Liu 2022). The accelerated advancement of cloud computing technologies has brought about a profound transformation across traditional sectors. A notable example that substantiates this transformation is the live streaming industry.

Live streaming delivers video in real time and supports applications such as sports broadcasting and interactive entertainment. Embracing the versatility of the Internet and media cloud, the live streaming industry has evolved from traditional linear television broadcasting systems toward a cloud computing-based model (Kleinerman 2022). Essential elements of this paradigm include Internet-enabled devices like smartphones, live platforms (e.g., *YouTube Live*, *Instagram*, and *Twich*) that enable real-time communication and engagement among users, and cloud service providers (CSPs) supplying essential services like data transmission. The recent explosive adoption of live platforms in online meetings and teaching, particularly during the COVID-19 pandemic, has reinforced the importance of live streaming. As of June 2021, live streaming users in China totaled 638 million, marking an impressive 47% annual increase and comprising 63% of all netizens (iiMedia Research 2022). In addition, projections suggest that the global live streaming industry is set to swell from $59.14 billion in 2021 to over $330 billion by 2030 (Grand View Research 2022).

The boom of the cloud computing and live streaming market provides both opportunities and challenges for CSPs. On the one hand, the robust demand from live streaming platform companies to migrate digital assets, services, databases, and applications into the cloud has created an expansive and lucrative business for CSPs. On the other hand, the competitive nature of the market has imposed a considerable cost burden on CSPs, which are keen to build or rent more and superior cloud infrastructure to confront the escalating demand and heightened expectations for service quality. Indeed, network infrastructure is the foundation of cloud computing. CSP expenditures on network bandwidth, which are charged by Internet Service Providers (ISPs), comprise a significant proportion of the operational cost of their live streaming services. In response to the burgeoning business requirements and the crucial necessity for bandwidth cost control, many CSPs have embarked on crafting effective traffic allocation systems to streamline the management of live streaming services.

Huawei Cloud

Huawei Cloud stands as a distinctive brand under Huawei's umbrella, catering to the realm of cloud services. It draws on Huawei's expertise accumulated over three decades in the field of information and communications technologies, products, and solutions and provides customers with reliable, secure, and sustainable cloud services. In 2021, Huawei Cloud maintained rapid growth with continuous innovation for inclusive technologies and constantly improved its cloud service capabilities and market share, enabling digital and intelligent upgrades across various industries.

According to a Gartner Report (Telecom Review 2021), Huawei Cloud ascended to the second position in the global Infrastructure as a Service (IaaS) market in China and the fifth place in the world. At present, Huawei Cloud has launched over 220 cloud services along with 210 technical solutions (e.g., efficient port scheduling, effective flight scheduling, and reliable financial data management) and has attracted over 30,000 partners worldwide and three million customers from a broad range of industries, including media entertainment, manufacturing, healthcare, finance, and logistics.

In China, Huawei Cloud serves 80% of China's top 50 Internet customers, 12 joint-stock commercial banks, and the top five insurance institutions. In addition, its services extend to over 30 smart airports, 30 urban rail networks, 29 provincial highways, 65% of provincial health insurance information platforms, over 30 automobile manufacturers, more than 20 major building materials and mining enterprises, and 15 top household appliance enterprises. In addition, Huawei Cloud built more than 40 industrial Internet innovation centers, helping 17,000 manufacturing enterprises in their digital transformation.

Live Streaming at Huawei Cloud

As a leading CSP in China, Huawei Cloud has been providing business-to-business (B2B) live streaming services since 2020. Its commitment extends to the provision of a complete suite of live streaming services, including management, transcoding, content delivery, live recording, and security. To support the global footprint, Huawei Cloud maintains a live streaming network with around 2,800 edge nodes in over 60 countries, boasting a total bandwidth capacity of up to 100 terabits per second (Tbps). This network supports over 10,000 domains with more than 15 million simultaneous online end users globally. Huawei Cloud has helped several major live platforms swiftly launch their live streaming services without building hardware stacks.

In the context of end-user experience, Huawei Cloud's technologies deliver high quality of service (QoS) marked by a stall frequency (i.e., the rate at which live streaming halts) of less than 2.5% and end-to-end latency (i.e., the delay between streamers and viewers) under three seconds. For example, Huawei Cloud successfully hosted live streaming services for a big sports event in China. Throughout the event, Huawei Cloud facilitated the streaming of over 60 games, with peak bandwidth utilization reaching 45 Tbps. Despite the high service pressure due to the immense bandwidth demand and unpredictable traffic bursts associated with this major global event, Huawei Cloud's live streaming services remained stable and reliable, outperforming various competitors and earning high praise from a broad range of end users.

Saving the Bandwidth Cost

Huawei Cloud's live streaming business includes two main types of infrastructure, that is, the edge nodes and the transmission network, which are used to transmit data from the source nodes to end users. Currently, only a few proportion of these infrastructures are self-built by Huawei Cloud, while the substantial majority are leased from three major ISPs in China. The success of the live streaming business requires not only a high-quality service supply but also effective cost control, particularly in the face of intense competition for limited network resources from other CSPs. In its early stage, Huawei Cloud was not in a highly competitive position and held a modest market share. Upon analyzing the workflow of live streaming services from source nodes to end users, we discerned that the bandwidth cost could account for more than 70% of total operational expenses. Furthermore, we recognized that the existing manual traffic allocation system, heavily reliant on expert knowledge, was neither efficient nor sustainable in light of the rapid business expansion and increasing complexity of the network. Motivated by the goals of reducing bandwidth costs and enhancing competitiveness, Huawei Cloud embarked on the development of an intelligent, automated traffic allocation system intended to eventually supplant the manual system.

However, traffic allocation problems in the context of cloud computing's live streaming business differ from those encountered in traditional traffic engineering scenarios such as logistics management (Lambert et al. 1998, Stock and Lambert 2001, Harrison et al. 2019) and transportation planning (Magnanti and Wong 1984, Steadie-Seifi et al. 2014). This divergence primarily arises due to the intricate nature of the 95th percentile billing scheme that charges network usage over a specific period based on its 95th percentile and several unique constraints, including data package replicability and tight response time requirements (typically in the order of milliseconds). Moreover, existing literature offers limited discussion on cost-effective traffic allocation issues within cloud computing application scenarios. For example, Singh et al. (2021) propose a mixed-integer linear programming (MILP) model to address the interdomain traffic allocation problem. However, their method relies on a commercial MILP solver, which requires 15 hours to generate a feasible solution. In addition, the scope of their study is limited to traffic allocation between data centers and three ISPs, a significantly smaller scale compared with the complexity of Huawei Cloud's live streaming business. Jalaparti et al. (2016) suggest using the average of the top 10% of bandwidth usages as an approximation of the 95th percentile utilization. Singh et al. (2021), however, point out that this approximation may not be valid for all network connections, particularly when traffic patterns fluctuate.

As a trailblazer in cloud technologies, Huawei Cloud decided to design innovative and effective traffic allocation algorithms to mitigate the bandwidth cost incurred in its live streaming business. Recognizing the potential hurdles, the company underscored the paramount importance of operations research (OR) techniques and sophisticated analytics in making optimal traffic allocation decisions. In a collaborative endeavor, research scientists from The University of Hong Kong and Huawei Cloud Algorithm Innovation Laboratory developed a cost-effective traffic allocation system named *GSCO*. Leveraging system engineering and OR principles, we dissected the usage of bandwidth resources across various stages of live streaming services, ultimately decomposing the entire process into several distinct modules. This structured approach facilitated bandwidth cost saving in live streaming by enabling the modular consideration of more manageable, solvable subproblems. Subsequently, our focus was directed toward each individual module, formulating efficient algorithms to solve the underlying mathematical models. We also scrutinized the interactive effects among different modules and tried to find practical and optimal strategies to reduce bandwidth cost from a holistic perspective. With the application of various OR techniques, including but not limited to continuous optimization, integer programming, graph theory, scheduling, and machine learning, the *GSCO* system has helped Huawei Cloud reduce the bandwidth cost by approximately 30%, leading to a total savings of more than $49.6 million from Q1 2020 to Q3 2022. Simultaneously, the *GSCO* system has facilitated an expansion of Huawei Cloud's peak bandwidth from 1.5 Tbps to 16 Tbps, all the while ensuring the maintenance of high QoS. We also detailed the portability of the *GSCO* system; for further insights into related cloud computing application scenarios, we direct interested readers to our previous work (Yang et al. 2022).

Problem Description and Challenges

An essential step in designing a traffic allocation system involves defining the optimization problem and constructing the associated mathematical models. In this section, we provide a high-level overview of the cost-effective traffic allocation problem encountered in Huawei Cloud's live streaming services. We first detail the hierarchical infrastructure underpinning Huawei Cloud's B2B live streaming business. Then, we establish the mathematical formulation for this intricate problem.

Hierarchical Infrastructure

The principal components of cloud computing-based live streaming include Internet-enabled devices, live platforms, and CSPs. For popular live platforms, there are millions of online audiences watching various live shows simultaneously. The audiences use different Internet-enabled devices and connect edge nodes that consist of servers or computing clusters managed by CSPs through the Internet to access cloud services. Subsequently, live streaming content is transmitted from live streamers to the audience (or vice versa) through specific networks and transmission protocols.

It is important to note that the edge nodes mentioned above serve as the interfaces between end users and the network. Indeed, the bandwidth cost accrued on edge nodes comprises the major bandwidth cost of the whole live streaming network (we cannot disclose the exact fraction due to confidentiality reasons). This is reasonable because of the massive egress and ingress traffic between millions of end users and thousands of edge nodes, while the same content only needs to be streamed once inside the transmission network. Therefore, minimizing the bandwidth cost accrued at edge nodes is an overriding step toward reducing the total bandwidth cost, which is the objective of our design of the cost-effective traffic allocation system. Specifically, the desired traffic allocation system acts as an intermediary, required to forward all the access requests from the end users to edge nodes in real time for connection establishment, while meeting four crucial requirements:

- Responding to every access request;
- Ensuring the bandwidth capacity of each edge node is not exceeded;
- Building connections exclusively over the transmission network of the same ISP;
- Accurately fulfilling all specific QoS requirements.

The first requirement is clear; serving customers well is Huawei Cloud's business philosophy, and thus the traffic allocation system must handle every access request and provide service for every end user. In regard to the second requirement, each network connection for live streaming between end users and edge nodes consumes bandwidth resources, and each edge node has the maximum bandwidth threshold contracted with ISPs. Hence, to avert extra costs and network congestion, the total traffic assigned to each edge node should not surpass this bandwidth capacity. The third requirement fundamentally reflects the practical constraint that inter-ISP connections are not allowed. The fourth requirement has two aspects. First, it is critical to ensure that services provided to customers meet the QoS standards. Furthermore, Huawei Cloud enters into various service level agreements (SLAs) with live platforms. These SLAs specify different QoS requirements and service prices. Therefore, the traffic allocation system should be able to build connections accurately according to these customized QoS requirements.

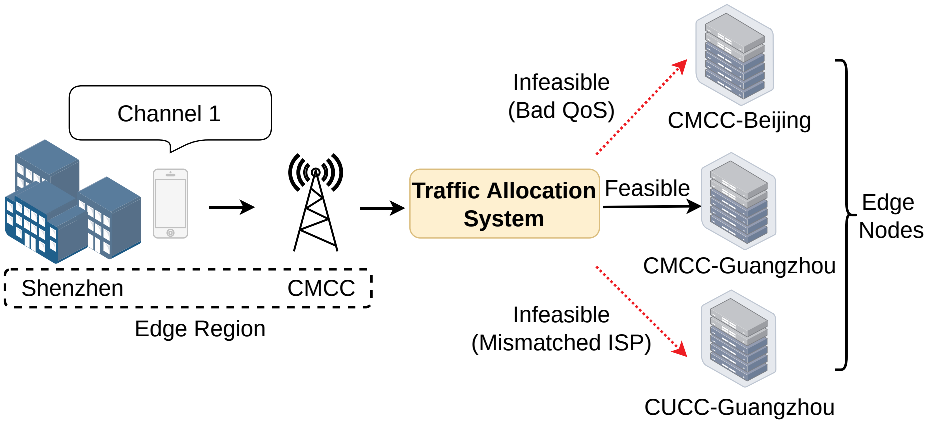
To meet the four fundamental requirements and manage traffic effectively while fully utilizing bandwidth resources to provide tiered QoS, we have established a hierarchical infrastructure for live streaming services. Access requests are analyzed in three principal dimensions: (1) live platforms, (2) channel groups differentiated by popularity, and (3) edge regions. First, as we mention above, different SLAs dictate various QoS requirements and service prices in the live streaming business. Therefore, it is necessary to recognize the live platform where an access request originates to provide QoS-differentiated services. Second, different live shows have various levels of popularity because the audiences' preferences and the streamers' reputations are diverse. Consequently, we classify all live shows into several channel groups based on their popularity (i.e., Tier 1 channel, Tier 2 channel, and so on). In general, channels with higher popularity require superior network infrastructure. This popularity-based classification contributes to the efficient management of varying resources. Third, each access request is labeled with ISP and location information, such as "CMCC-Beijing," after recognizing the live platform and channel group, where CMCC is the abbreviation of *China Mobile Communication Company Limited*. ISP information is necessary for traffic allocation due to the prohibition of inter-ISP connections. Location information assists in maintaining specific QoS levels, because shorter communication distances generally imply lower latency and more stable connections. Similarly, edge nodes are also identified by ISP and location information. Then, the target of cost-effective traffic allocation is to establish a matching between edge regions and edge nodes through QoS-qualified network connections; this matching should minimize the total bandwidth cost accrued at all edge nodes under the 95th percentile billing scheme, ensuring the bandwidth capacity of each edge node is not exceeded, and all access requests are addressed.

Figure 1 depicts an example of an access request routing on Huawei Cloud's hierarchical infrastructure. In this scenario, there is a client's access request from a Tier 1 channel in Shenzhen using CMCC. Location information "Shenzhen" and ISP information "CMCC" together constitute an edge region referred to as "CMCC-Shenzhen." To reach an audience that could reside in any part of China, this request must be forwarded to an edge node that utilizes the matching ISP and can guarantee acceptable QoS levels, such as the edge node labeled "CMCC-Guangzhou." The figure also illustrates why edge nodes like "CMCC-Beijing" and "CUCC-Guangzhou" would be unsuitable. The former fails to meet the necessary QoS requirements, while the latter does not match the original ISP. In addition, if multiple feasible edge nodes exist, the traffic allocation system should select a connection plan that minimizes the total bandwidth cost.

Problem Formulation

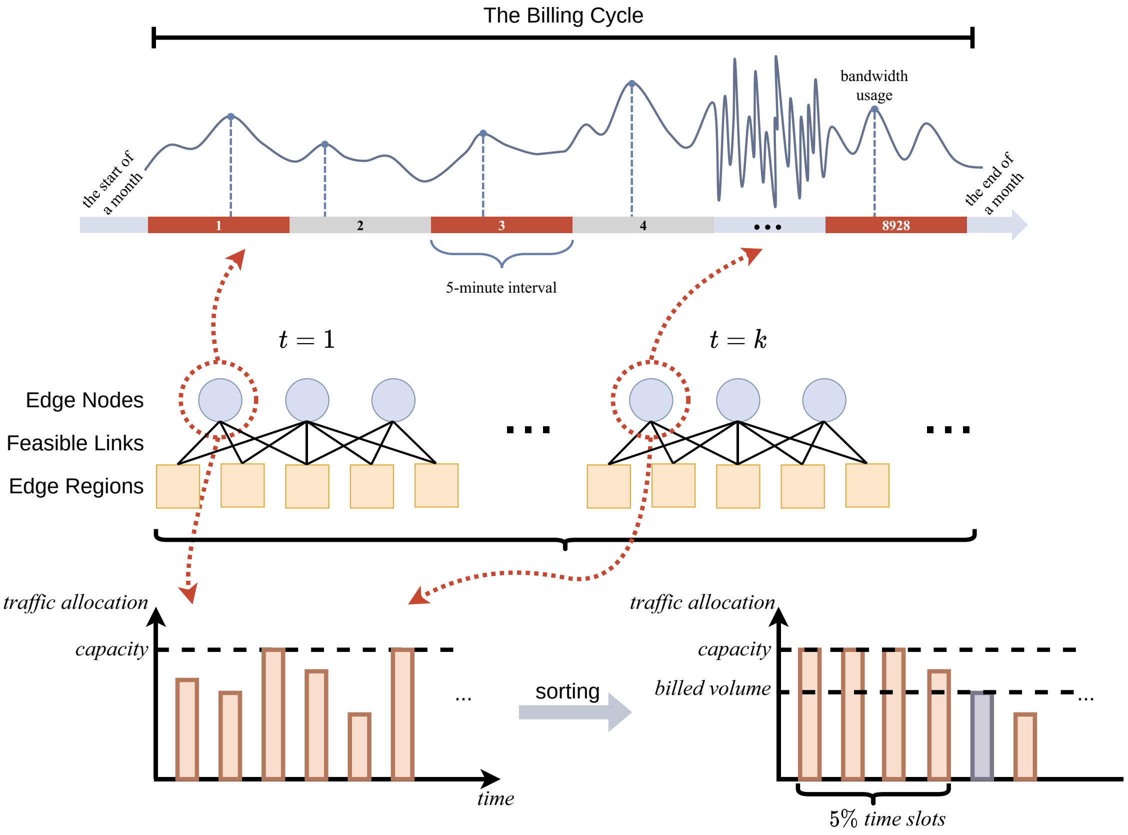
According to Huawei Cloud's hierarchical infrastructure of live streaming services and the 95th percentile billing scheme, we can frame the cost-effective traffic allocation problem as a generalized assignment problem, where the network topology is a bipartite graph with a time dimension. Figure 2 presents an abstract depiction of this network topology and offers an example of the 95th percentile billing scheme. Although we have mentioned this billing scheme multiple times, we give its explicit definition in Huawei Cloud's live streaming services here: the peak bandwidth usage on each edge node is measured every five minutes, and the 95th percentile of these measurements forms the billed bandwidth usage over a billing cycle (e.g., per day, per week, or per month). This arrangement indicates that a 5% window of time slots, corresponding to the top 5% of measurements, is free of charge.

Figure 1. (Color online) The Graphic Shows an Example of an Access Request Routing on Huawei Cloud's Hierarchical Infrastructure



Note. CMCC and CUCC are two major ISPs in China.

Figure 2. (Color online) The Graphic Illustrates the Abstracted Network Topology and Shows an Example of the 95th Percentile Billing Scheme



Notes. The middle part is the network topology, where we take the circled edge node as an example. The top part represents the bandwidth usage of this edge node during a billing cycle. The bottom left part depicts the sampled bandwidth usage of this edge node, where we sample the peak bandwidth usage in each five-minute interval. The bottom right part is the sorted samples where the 95th percentile is the billed volume of this edge node.

To analyze quantitatively how to allocate bandwidth between edge regions and edge nodes for minimizing the 95th percentile bandwidth cost while ensuring QoS, we developed a mathematical programming model, as we describe in the appendix.

Challenges

Addressing the cost-effective traffic allocation problem and implementing a traffic allocation system pose significant challenges from both engineering and optimization perspectives. In this section, we delineate the principal obstacles we encountered in these domains.

Engineering Challenges. Designing the overall framework of a traffic allocation system is a complex task that entails systems engineering involving diverse technological areas and myriad factors. This begins with an exhaustive examination of the live video delivery process, from the source to the end users, and then modularizing this process into smaller, more manageable tasks. To ensure the successful implementation of these modules, we first defined the inputs and outputs of each module. We then collaborated closely with frontline engineers to confirm the availability of the necessary input data. Gradually, we designed and tested different modules, and iteratively refined them based on feedback from the engineers. To ensure a logically coherent system, it was essential to carefully consider the interactive effects among different modules and optimize their interactions in a systematic and holistic manner. This comprehensive approach enabled us to develop a streamlined and efficient system that meets our performance objectives.

To craft a sound traffic allocation plan that optimally reduces the total bandwidth cost while maintaining satisfactory QoS, an accurate forecast of future demand is critical. However, achieving this precision is challenging due to the difficulty of predicting unexpected events. For instance, a sudden surge in access requests may be triggered by the emergence of breaking news.

Addressing real-time dynamic access requests while providing satisfactory QoS necessitates a traffic allocation system capable of producing decisions within milliseconds. As such, high-speed algorithms and high-performance programming are imperative.

The network infrastructure also imposes additional restrictions on the traffic allocation strategies. As illustrated in the subsection *Hierarchical Infrastructure*, we can only serve end users in an edge region with edge nodes that provide satisfactory QoS and are serviced by the same ISP. Further, there may be specific restrictions imposed on the usage of the network infrastructure due to management concerns. Therefore, a cost-effective traffic allocation system that is viable for real-world business implementation must respect all these additional restrictions.

Optimization Challenges. Achieving optimal online traffic allocation strategies necessitates the discovery of effective offline solutions to the proposed mathematical programming model. However, the underlying model is analytically difficult due to the following two challenges.

First, the objective function of the model is particularly intricate. ISPs charge Huawei Cloud using the 95th percentile billing scheme. As we mention above, this implies that only the 95th percentile of bandwidth utilization over a billing cycle is billed, which results in a nonconvex and nonsmooth cost function. Indeed, minimizing the 95th percentile billing cost has been proven to be NP-hard. Studies on optimizing the 95th percentile billing cost for live streaming business are lacking as we mention in the *Saving the Bandwidth Cost* section. Therefore, we needed to work from scratch to design algorithms that could adapt to such challenges.

Second, the mathematical model in practice is vast in scale. There are more than 4,800 edge regions and 2,800 edge nodes with heterogeneous capacities and computing abilities, resulting in millions of decision variables in the corresponding models. The scale of this model could require more than 300 GB of computer memory for storage. Therefore, we could not even load this model into a computer without a sufficient hardware configuration, let alone find an effective solution using off-the-shelf solvers. Hence, we needed to develop decomposition methods to efficiently solve the original problem. For example, we applied operator splitting algorithms to reduce the large original problem into several smaller and more manageable subproblems, iteratively solve them, and aggregate the results to find a good solution to the original problem.

Technical Solution

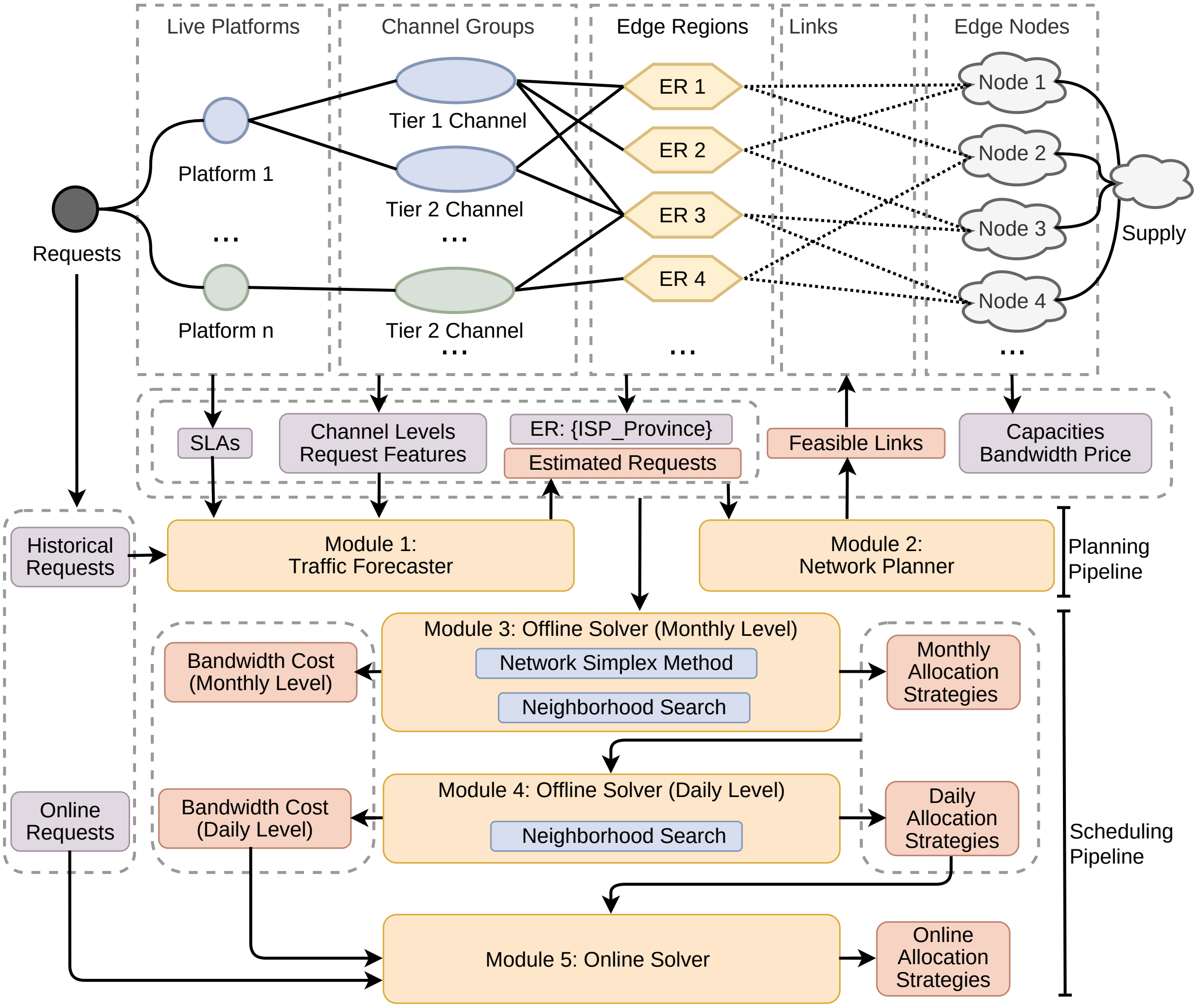
After realizing the significance of reducing the bandwidth cost and maintaining satisfactory QoS, as well as the inherent challenges, we started to develop a cost-effective and robust traffic allocation system called the *GSCO* in early 2020. In this section, we delve into the underlying logic of the *GSCO* system, discussing the design and function of its constituent modules and the OR techniques employed in its creation.

Overview of the *GSCO* System

Figure 3 depicts the architecture of the *GSCO* system, which consists of five main modules classified into a planning pipeline and a scheduling pipeline. The planning pipeline executes to produce the necessary input data, subsequently facilitating the operation of the scheduling pipeline.

In the planning pipeline, the *Traffic Forecaster* extensively exploits machine learning techniques to estimate future requests. In addition, the *Network Planner*

Figure 3. (Color online) The Flowchart Provides an Overview of the GSCO System for Efficient Bandwidth Allocation

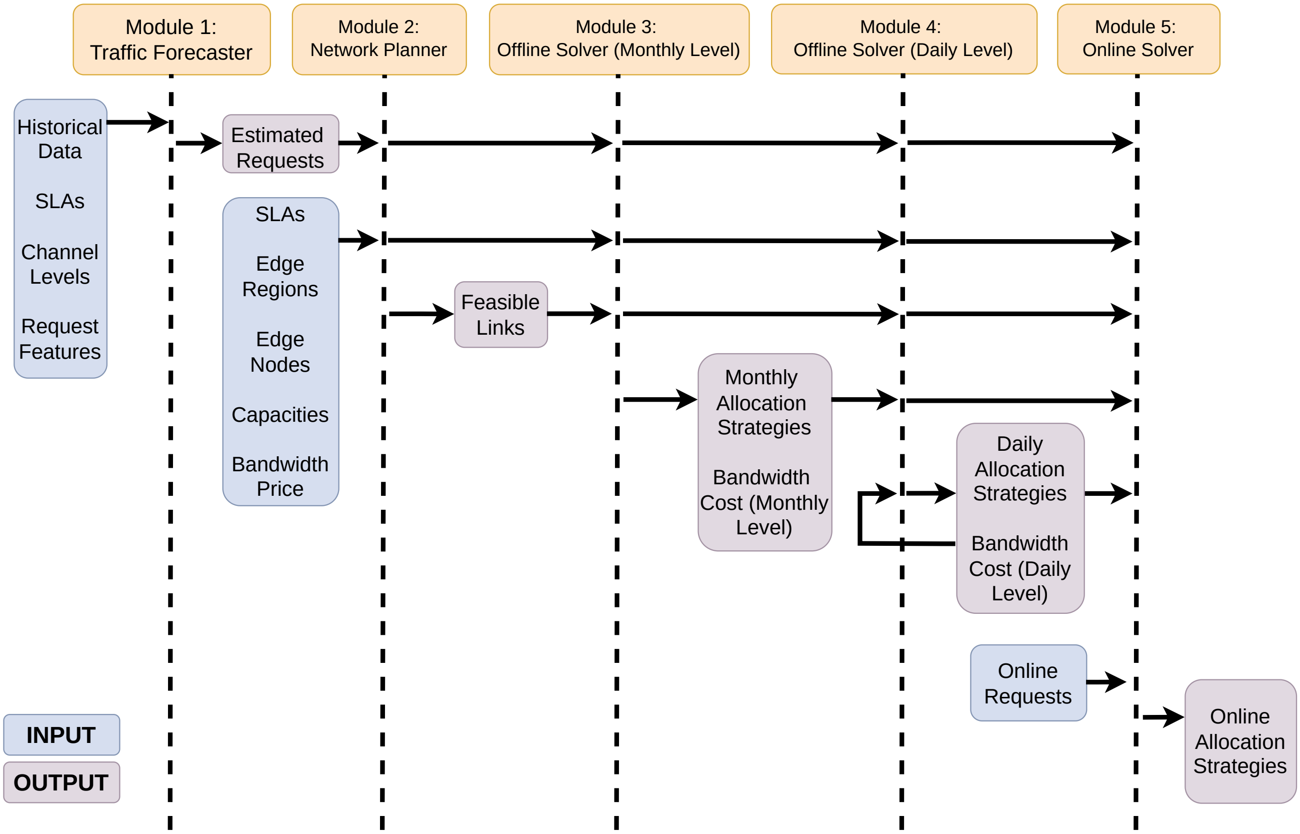


generates QoS-qualified and ISP-compliant connections between edge regions and edge nodes.

Because the online traffic allocation strategies must be generated within milliseconds, we leverage scheduling theory and design a scheduling pipeline from coarse to fine to allocate access requests. First, the *Offline Solver (Monthly Level)* creates monthly allocation strategies and appraises the bandwidth cost at the monthly level. At this phase, the static offline traffic allocation problem at a single time point is considered. Then, we implement the neighborhood search algorithm (NSA) by Mladenović and Hansen (1997) to tune results and obtain a higher multiplex rate, which measures the utilization of the billed bandwidth of edge nodes, and better QoS without increasing the total bandwidth cost. Several scalable heuristic algorithms are embedded into the NSA. Second, the *Offline Solver (Daily Level)* produces daily allocation strategies and evaluates the bandwidth cost at the daily level. This module applies the NSA to adjust the input allocation strategies. Third, during the *Online Solver* phase, we utilize real-time optimization strategies to allocate network traffic. This allocation is based on previously calculated ratios derived from the traffic volumes assigned to feasible connections between edge regions and edge nodes. In this module, cost predictions from the offline solvers are leveraged, complemented with dynamic mechanisms that are integrated to manage uncertainty and ensure optimal performance.

The above five core modules are interlinked and mutually reinforcing. The I/O flow of the *GSCO* system is summarized in Figure 4. The *Traffic Forecaster* takes historical data collected by the data management system, SLAs contracted with live platforms, channel levels predetermined by operations experts, and request features analyzed by data scientists as input, and generates the estimated requests, which serve as reference information for the other modules. The information on the network topology (including edge regions and edge nodes), and problem parameters (including bandwidth

Figure 4. (Color online) The Graphic Illustrates the I/O Flow of the GSCO System



capacities of edge nodes and bandwidth prices), comprise a part of the input to Modules 2–5. In addition, the *Network Planner* incorporates SLAs to create feasible network connections between edge regions and edge nodes that meet tiered QoS requirements. Moreover, these connections comply with ISP-compliant regulations. The three traffic allocation solvers, namely Modules 3–5, utilize the generated feasible links as one of their input parameters. This ensures that traffic allocation strategies are formulated exclusively based on these links. Then, the *Offline Solver (Monthly Level)* outputs monthly allocation strategies and the corresponding monthly bandwidth cost based on monthly data from estimated requests. Similarly, the *Offline Solver (Daily Level)* outputs daily allocation strategies and the corresponding daily bandwidth cost based on daily data from updated forecasts and monthly allocation strategies. The *Online Solver* utilizes the results derived from the two offline solvers to handle online requests in real time. Specifically, the monthly level results are employed when processing access requests on the first day of a month, whereas the daily level results are leveraged for all other days. The previous allocation strategies created by these two offline solvers are leveraged by the *Offline Solver (Daily Level)* based on the same principle.

GSCO System Modules

The five modules we show in Figure 3 are the principal components of the *GSCO* system. We describe the technical details and OR methods adopted in each of them in this section.

Module 1: Traffic Forecaster. The forecaster system, which is used to predict traffic distribution between edge regions and edge nodes, primarily relies on state-of-the-art machine learning methods, including BHT-ARIMA (Shi et al. 2020). This method is based on multiway delay-embedding transform tensorization (Yokota et al. 2018) and uses low-rank Tucker decomposition to implement tensor ARIMA (Box and Jenkins 1968). Throughout the implementation process, we gained several valuable insights.

Foremost, there is uncertain volatility caused by various factors. A salient example is the sudden surge in access requests triggered by breaking news. Additionally, there exists manual management of traffic allocation for tackling unexpected malfunctions. The information regarding such manual interventions is not communicated to the forecasting module, thereby reducing the prediction accuracy. Finally, the collection of adequate historical data to train our machine learning methods presents a significant challenge, particularly because privacy considerations limit access to detailed client information. These challenges constitute serious impediments that may negatively impact prediction accuracy.

Although our initial focus was on refining the prediction accuracy of the *Traffic Forecaster* to better align predicted and actual demand, the aforementioned factors necessitated a shift in our approach. We came to appreciate the inherent difficulties in achieving perfect demand alignment using prediction models alone. With this realization, we pivoted toward developing a system capable of adjusting to and accommodating changes in demand, even in the face of suboptimal predictions. Consequently, we designed the *GSCO* system with this adaptability feature. It utilizes the most recent data available in offline solvers to update traffic allocation strategies, dynamically responding to demand changes in real-time operations through the *Online Solver*. This agile handling of demand uncertainties and the ability to adapt to real-time changes underscore a significant advantage of the *GSCO* system in optimizing traffic allocation strategies. By acknowledging the limitations of prediction models and pivoting toward adaptive strategies, we created a system that is robust and well-equipped to meet the demands of modern network infrastructures.

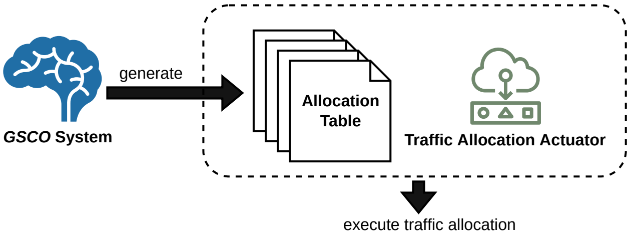
Module 2: Network Planner. The primary task of the *Network Planner* is to identify feasible network connections between edge regions and edge nodes that comply with ISP-compliant regulations and meet the QoS requirements as stipulated in SLAs. First, the *Network Planner* is vigilant to maintain ISP compliance, restricting inter-ISP connections. For instance, it would not sanction the establishment of a connection between edge regions "CMCC-Shenzhen" and "CUCC-Guangzhou" due to mismatching ISPs. Second, the *Network Planner* assesses each possible link between an edge region and an edge node to verify that it can fulfill the pertinent SLAs. Referring back to the example in Figure 1, should a link between the edge region "CMCC-Shenzhen" and the edge node "CMCC-Guangzhou" meet the QoS requirements for the Tier 1 channel group, this connection would be considered a viable option for accommodating corresponding access requests. Conversely, a link that fails to meet these requirements, such as the connection between "CMCC-Shenzhen" and "CMCC-Beijing," would be deemed infeasible for traffic allocation. The measure of various metrics, including stall frequency and end-to-end latency, are implemented in this module to continuously monitor the QoS.

Determining ISP-compliant connections is relatively straightforward. Initially, we perform a comprehensive matching process and only need to make minor adjustments when the information pertaining to the edge regions and edge nodes is updated. However, maintaining QoS-compliant connections is a more dynamic process. To this end, the *Network Planner* automatically reassesses the feasibility of network connections every two minutes based on current QoS metrics. This ensures that the SLAs of our clients are stringently upheld.

Module 3: Offline Solver (Monthly Level). The monthly level offline solver is run at the end of the current month to devise an initial allocation strategy and provide an estimate of the total cost for the following month. In this module, a static offline traffic allocation problem based on our clients' expected demand for the next month at a time point is considered. It is formulated as a minimum-cost network flow problem (MCNFP) by leveraging a linear approximation of the total cost (see the appendix for further details). The MCNFP can be efficiently solved by the combination of several algorithms, such as the generalized primal-dual algorithm (He et al. 2022), the balanced augmented Lagrangian method (He and Yuan 2021), the extended alternating direction method of multipliers (He and Yuan 2018), and the network simplex method (Cunningham 1976). We implemented these advanced optimization algorithms in Python on a PC with an Intel® Core^{TM} i7-8700 CPU and 32GB RAM; the MCNFP can be solved within 10 seconds for real datasets.

Subsequently, the NSA is deployed to refine the initial allocation strategies. The aim is to increase the multiplex rate, bolster robustness, and enhance QoS without increasing the total bandwidth cost. The NSA integrates some removal and repair heuristics to improve the current solution. The concept is to destroy a part of the solution (i.e., remove it from the current solution) and repair it afterward. The objective of the removal heuristics is to create opportunities for the repair heuristics to optimize the solution. In practice, we apply two pairs of removal and repair heuristics. The first one is to improve the multiplex rate. This heuristic tunes the allocation strategy by capitalizing on excess capacity at edge nodes serving channel groups with higher popularity to provide services for those with lower popularity. As we note above, each edge node has 5% free time slots during a billing cycle. If these capacities are only partially utilized or remain entirely unused during these slots, it would represent a marked inefficiency in bandwidth resource utilization. Thus, it is advantageous for us to consolidate access requests onto edge nodes possessing redundant capacities during such slots. This strategy takes advantage of the fact that these edge nodes are not fully utilized by the access requests complying with their respective SLAs. Essentially, this redistribution does not violate our QoS commitments; rather, it ensures they are upheld more diligently. We only allocate this excess capacity to requests that have less rigorous QoS requirements, ensuring they are adequately served by nodes

Figure 5. (Color online) The Graphic Illustrates the Relationship Among the *GSCO* System, the Allocation Table, and the Traffic Allocation Actuator



Note. The actuator executes the traffic allocation strategies generated by the *GSCO* system, as represented by assignment probabilities in the allocation table.

capable of meeting even higher QoS standards. This approach, therefore, maintains and even enhances our overall compliance with QoS requirements. The second one is designed to improve system robustness. This heuristic refines the allocation strategy with the intention of increasing the number of edge nodes allocated to a single edge region, thereby considerably enhancing the resilience of the system. Consequently, if an unexpected event occurs, such as an edge node failure, the impact will be minimal. In addition, we can promptly redirect access requests, initially linked to the affected node, to other functioning edge nodes. Such a configuration allows for a swift recovery and ensures uninterrupted service, thus preserving the robustness and reliability of the system.

Module 4: Offline Solver (Daily Level). The *Offline Solver (Daily Level)* operates daily, with the exception of the last day of the month when the *Offline Solver (Monthly Level)* takes precedence, as we mention above, to refresh allocation strategies and update the estimated total bandwidth cost. During the operation, it employs the NSA, utilizing the same heuristic algorithms as detailed in *Module 3: Offline Solver (Monthly Level)*. The focus of this application is to refine the allocation strategies, which have been formulated either on a preceding day or by the *Offline Solver (Monthly Level)* on the initial day of a month. This refinement process takes into consideration the most recent predicted demand and the current network topology. In this way, the allocation strategies are constantly updated and optimized, reflecting the dynamic nature of the demand and the network conditions. The output of this module is an optimized set of allocation strategies and the estimated total bandwidth cost. These data are subsequently leveraged extensively within the *Online Solver* for real-time operations.

Module 5: Online Solver. The *Online Solver* is tasked with generating real-time allocation strategies while adhering to strict time constraints. It creates an allocation table that embodies these strategies in the form of probabilities, which represent the likelihood that access requests from each edge region will be routed to each respective edge node by the traffic allocation actuator. Figure 5 describes the relationship among the *GSCO* system, the allocation table, and the traffic allocation actuator. The *GSCO* system is analogous to that of a brain, generating traffic allocation strategies for the actuator to execute the traffic allocation process. The allocation table, meanwhile, provides the tangible output of these strategies. Table 1 shows an example of the allocation table. In this instance, the access requests from the edge region "CMCC-Shenzhen" will be assigned to the edge node "CMCC-Shenzhen" with a probability of 0.8 and to the edge node "CMCC-Guangzhou" with a probability of 0.2. In practice, we uniformly draw a random number from the interval $[0, 1]$ when an access request arrives. If it is less than 0.8, this access request will be assigned to "CMCC-Shenzhen." Otherwise, it will be assigned to "CMCC-Guangzhou." It should be noted that the given example is a simplified representation for illustrative purposes. In a real-world scenario, we employ the multinomial distribution to determine the assignment of an edge node to an edge region. This particular probability distribution proves useful when there are more than two potential edge nodes to be assigned to an edge region. Each node is assigned a probability, as defined in the allocation table, the summation of which equals 1.0. Upon the arrival of an access request, a random variable adhering to the defined multinomial distribution is generated and the outcome corresponds to a specific edge node, which is then assigned the incoming access request.

The allocation table expedites the traffic allocation process, enabling it to be completed in milliseconds. This efficiency is because the *GSCO* system only needs to execute a search algorithm to obtain the corresponding probabilities from the allocation table, and then a random number generation method to finalize the allocation decision. This allocation table is generated based

Table 1. The Table Provides an Example of the Allocation Table that Contains the Probabilities of Assigning Traffic from Each Edge Region to Each Edge Node

	CMCC-Shenzhen	CMCC-Guangzhou	CMCC-Beijing
CMCC-Shenzhen	0.8	0.2	0.0
…	…	…	…

on the outputs of the *Offline Solver (Monthly level)* and the *Offline Solver (Daily level)*. The probabilities used in the allocation table are derived by calculating the ratios of the traffic volume assigned to each edge node based on the offline allocation strategies, and further adjusted to accommodate dynamic conditions. This process ensures that the online strategy is responsive to real-time changes in the network environment, thereby maintaining our commitment to QoS.

Comparison of the Scheduling Modules

Three scheduling modules in the *GSCO* system (i.e., Modules 3–5) produce allocation strategies in different time horizons. Collectively, these three modules consist of a comprehensive workflow for allocating bandwidth as delineated in the preceding sections of this paper, and their relevance varies across distinct scenarios.

In scenarios where there are no extraordinary events, Modules 3 and 5 are the dominant modules because of the roughly similar trends of traffic requests in the short term. However, in the case of a super event, Modules 4 and 5 are the dominant modules because allocation strategies are expected to be adjusted daily, and we need to handle a sudden surge of access requests. In emergency situations, Module 5 becomes the principal module. In these instances, allocation strategies necessitate immediate and timely adjustments to manage the abrupt and unpredictable changes in network traffic.

Managerial Challenges

We encountered two major managerial challenges throughout the development of the *GSCO* system.

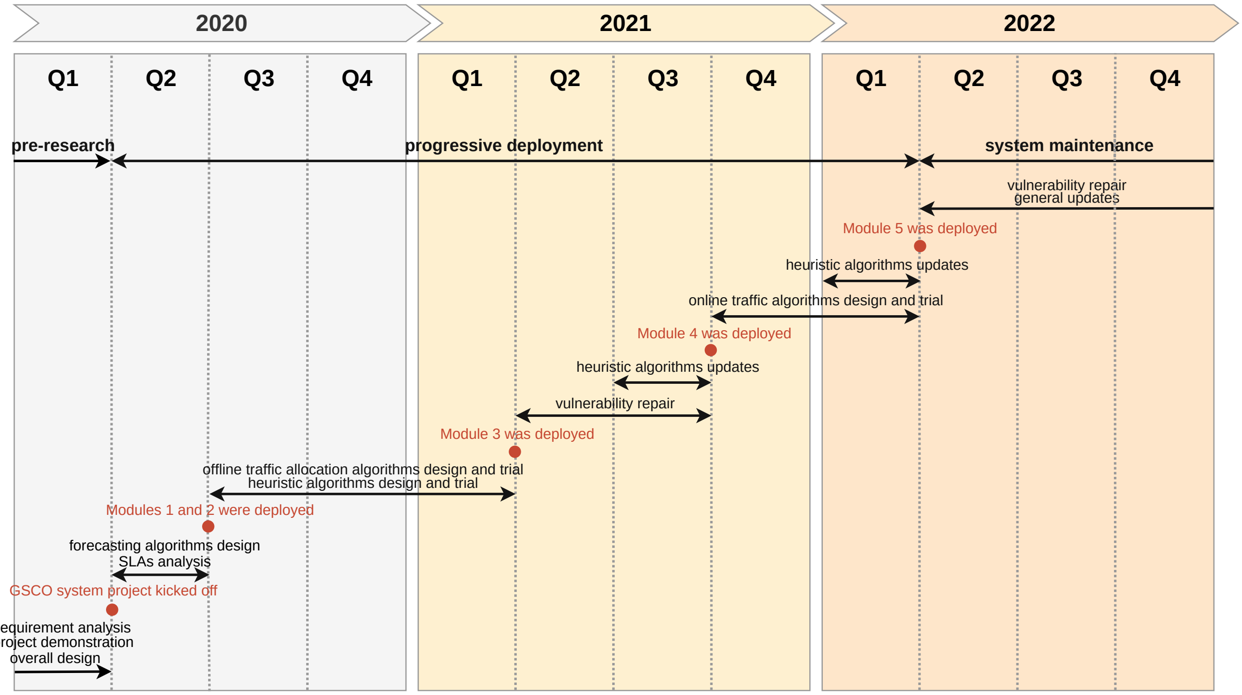
First, advocating the value of OR to senior managers lacking a deep understanding of OR was challenging. This difficulty slowed the initial development phase of the *GSCO* system. To address this, we initiated a series of informative lectures, showcasing successful examples of how OR can effectively reduce bandwidth costs while maintaining high QoS. These educational presentations captivated the interest of senior managers and frontline engineers, leading to a broader understanding and eventual acceptance of our proposed system.

Second, we faced the challenge of persuading the operations and maintenance teams of live streaming services to adopt the *GSCO* system. Their primary concern was the stability of the system. To alleviate these concerns, we initially deployed the system to a small selection of edge nodes, thereby demonstrating its robustness and stability over several months. Then, we showed the superior performance of the system in terms of bandwidth cost efficiency and QoS to convince the operational staff of its benefits. As a result, the *GSCO* system gradually gained acceptance and is now utilized for live streaming services in Huawei Cloud.

Milestones

Figure 6 shows the timeline of the development of the *GSCO* system. We highlight the following milestones.

Figure 6. (Color online) The Development of the *GSCO* System Spans More Than Two Years


2020
2021
2022
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
pre-research
progressive deployment
system maintenance
vulnerability repair
general updates
Module 5 was deployed
heuristic algorithms updates
online traffic algorithms design and trial
Module 4 was deployed
heuristic algorithms updates
vulnerability repair
Module 3 was deployed
offline traffic allocation algorithms design and trial
heuristic algorithms design and trial
Modules 1 and 2 were deployed
forecasting algorithms design
SLAs analysis
GSCO system project kicked off
requirement analysis
project demonstration
overall design

- Before Q1 2020: Amidst increasing business expansion and network complexity, Huawei Cloud recognized the inefficiency of manual traffic allocation. This led to the decision to develop an automated system based on systems engineering and OR principles.
- Q1 2020: With a comprehensive cost analysis and demonstration of our initial *GSCO* system roadmap, we illustrated the superiority of an OR-based approach over the previous manual system. This was crucial in obtaining senior management approval, leading to the official commencement of the *GSCO* system project.
- Q2 2020: We established two integral components of the *GSCO* system: a traffic forecaster utilizing various machine learning techniques, and a network topological planner, ensuring QoS through the creation of viable connections between edge regions and nodes.
- Q1 2021: We implemented Module 3 of the *GSCO* system, an offline solver capable of generating monthly traffic allocation via a two-phase strategy, heavily leveraging advanced optimization algorithms and the NSA.
- Q3 2021: We enhanced the interaction between deployed modules by fixing several engineering vulnerabilities and updating the heuristic algorithms within the NSA. Additionally, we completed the development of the second offline solver, Module 4 of the *GSCO* system, which was designed to generate daily traffic allocation.
- Q1 2022: We launched Module 5 of the *GSCO* system, an online solver designed for real-time traffic allocation, utilizing various fast-execution strategies.
- After Q1 2022: The *GSCO* system was essentially complete. Our focus shifted primarily to repairing vulnerabilities and rolling out general updates. The system, according to recent statistics, continues to efficiently generate traffic allocation strategies, accommodating up to eight million access requests per minute as of Q4 2022.

In essence, the deployment of the *GSCO* system was an iterative process, manifested not only in the step-by-step implementation of core algorithms and modules but also in the gradual broadening of its operational sphere within the network. To clarify, we initially applied the developed algorithms within a restricted set of edge nodes and edge regions. Subsequently, we consistently rectified vulnerabilities, refined algorithms, and widened the operational range, until we successfully integrated a module into the entire live streaming service. This phased development strategy enabled us to address problems promptly and to ensure the production of high-quality software.

Financial Benefits

With various built-in OR techniques, the *GSCO* system has been implemented to solve the traffic allocation problems of Huawei Cloud in regard to its live streaming services. To quantify *GSCO*'s financial benefits, we define the unit cost for a given quarter to be the quotient of the total fee we pay to ISPs to purchase bandwidth resources and the total amount (in bandwidth) of live streaming service provided:

$$\text{the unit cost} = \frac{\text{the total fee paid to ISPs}}{\text{the total amount of live streaming service provided}}.$$

Note that the unit cost is not the procurement price from the ISPs for bandwidth resources. Instead, this equation measures the cost of providing one unit of live streaming service in bandwidth (e.g., 1 Gigabit per second) for our clients in each quarter. Therefore, if the price of network bandwidth charged by the ISPs remains unchanged, a decrease in this unit cost indicates that the *GSCO* system enables us to provide the same amount of service but with a lower bandwidth cost. It is worth mentioning that such assumptions on the static unit price of bandwidth resources are reasonable. The network bandwidth industry differs from the traditional retailing business, where the scale effect on the unit price is prominent. Network bandwidth is indeed a kind of scarce resource because currently, the construction of network infrastructure is usually slower than the increasing demand. In the network bandwidth market, many expanding cloud providers compete for limited bandwidth resources. Hence, even though our market share has increased with a smoother demand and larger traffic bandwidth, the unit bandwidth price charged by the ISPs usually does not decrease. In addition, in some cases, such as a super event, the unit bandwidth price even increases due to fierce competition. However, it is exactly the *GSCO*'s ability to reduce the bandwidth cost that motivates Huawei Cloud's clients to increase the scale of their business, thereby proportionally enhancing Huawei Cloud's market share.

We then set the unit cost as of Q1 2020 as the benchmark, prior to the implementation of the *GSCO* system. Specifically, we compute and normalize the unit cost of each quarter such that the unit cost in Q1 2020 is 1.0. Then, we can obtain the bandwidth cost-saving percentage. For example, the normalized unit cost is about 0.97 in Q2 2020, and the bandwidth cost-saving percentage is 3% ($1.0 - 0.97 = 3\%$). By the unit cost calculation above, such a percentage indicates the proportion of bandwidth cost savings in supplying the same amount of services compared with Q1 2020. Therefore, we determined the estimated savings on bandwidth costs by multiplying the actual total cost paid to ISPs by the corresponding bandwidth cost-saving percentage. For example, the actual total cost we paid to ISPs in Q2 2020 was $6.96 million, the corresponding bandwidth cost-saving percentage was 3%, and thus the estimated savings on the bandwidth cost in this quarter was about \$0.21 million (6.96 × 0.03 = 0.2088). That is, to maintain the same level of live streaming service, we would have needed to pay an additional \$0.21 million to the ISPs if we had not implemented the *GSCO* system. Finally, we obtained the estimated accumulated bandwidth cost saving of \$49.6 million by summing the bandwidth cost saving of each quarter from Q1 2020 to Q3 2022. In addition, we use the bandwidth cost-saving percentage of Q3 2022 to represent the latest cost-saving ability of the *GSCO* system, which is 30%.

Figure 7 summarizes the aforementioned metrics and financial data. In this figure, each circular marker represents the normalized unit cost for each respective quarter. The dotted bars represent quarterly actual total bandwidth costs paid by Huawei Cloud to ISPs. All numbers of the dotted bars are real costs provided by Huawei Cloud's finance department. The slashed bars represent quarterly estimated savings on bandwidth costs with the *GSCO* system. That is, the sum of a dotted bar and the corresponding slashed bar represents the estimated total bandwidth cost if we had not implemented the *GSCO* system during that quarter.

It is important to note that obtaining the baseline cost without the *GSCO* system through comprehensive A/B tests would have been a more appropriate approach. However, due to the nature of our cross-regional scheduling and the need to optimize for the overall system, it was difficult to set up two identical test environments. Even if we had managed to find two identical environments to conduct the experiment, using different scheduling algorithms would have resulted in different end-user experiences, which would not have met the client's requirements for consistency in user experiences. Therefore, while we acknowledge the limitations of not having a proper A/B test, we have taken steps to minimize any potential bias in our results.

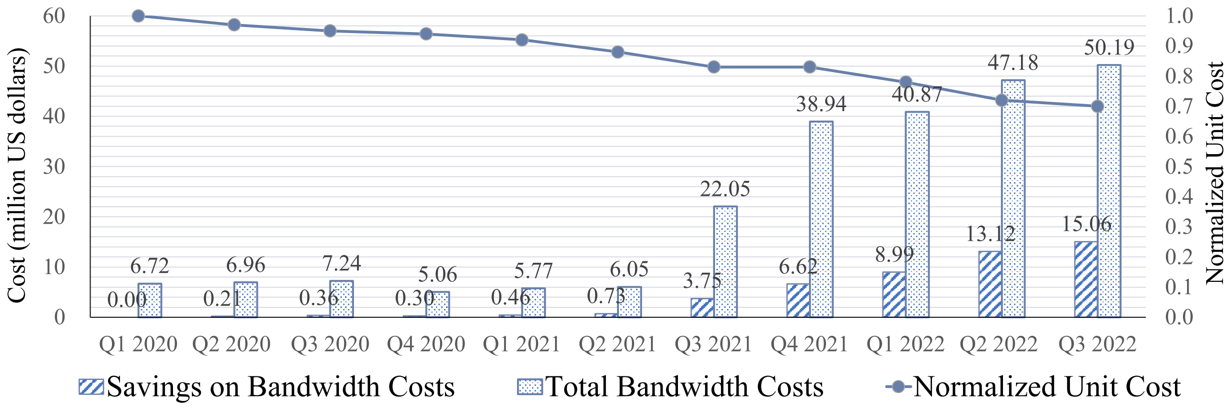
Impact

Beyond the financial benefits discussed above, the *GSCO* system has also considerably extended Huawei Cloud's market share within the live streaming sector, enhanced our employee experience, and established a precedent for successful utilization of OR methods in addressing issues prevalent in cloud computing. The ensuing discussion in this section will illustrate these additional benefits.

Expanding Market Share

With lower cost and better QoS (i.e., lower stall frequency and end-to-end latency), the *GSCO* system significantly increased Huawei Cloud's live streaming market share from 1.5 Tbps to 16 Tbps measured by the peak bandwidth. Each of Huawei Cloud's live streaming service clients chooses its cloud providers from many tenders, mainly by the providers' QoS and quotation. Take one of our major clients as an example. To strictly evaluate the service of its cloud providers, the client has a scoring system, which consists of the price and marketing strategies (50%), quality of service (30%), and SLA (20%). The client then determines its business scale in accordance with the scores of all bidding cloud providers. The *GSCO* system has helped Huawei Cloud provide live streaming services to the client at a lower price and with a higher QoS. Thus, it has been increasing its business scale with Huawei Cloud from 2020 to 2022. For example, Huawei Cloud's market share of this client's live streaming fluctuated between 2% and 4% in 2020, during which the *GSCO* system project was in its infancy. In 2021, Modules 1–3 of the *GSCO* system were deployed, and we realized a 17% reduction in the unit cost, leading to a significant increase in Huawei Cloud's market share of this client's live streaming services from 6% in Q1 2021 to 12% in Q4 2021. In 2022, we deployed the complete *GSCO* system, the unit cost decreased by an additional 15%, and Huawei Cloud's market share climbed gradually from 18% to 20%.

Figure 7. (Color online) The Graph Illustrates Financial Benefits in Terms of Quarterly Bandwidth Cost Savings



Improving Employee Experience

The *GSCO* system has revolutionized Huawei Cloud's traditional, experience-based model by automating the generation of traffic allocation results. This has markedly elevated operational efficiency and considerably lessened the need for labor. Before its implementation, we needed three experts to manually monitor and schedule traffic, each working eight hours a day. Now, we need only one expert working less than 20 minutes each day to update *GSCO*'s configurations. Therefore, the implementation of the *GSCO* system led to a decrease of 98.6% in expert labor.

We designed an online graphical user interface (GUI) to visualize all the data throughout the operation of the *GSCO* system. This interface conveniently presents a variety of information in graph and chart formats, including traffic request distribution, current traffic allocation strategies, traffic allocation results, the billed bandwidth, and QoS metrics. This GUI equips engineers and decision makers with the tools needed to constantly monitor the system's operational status and swiftly adjust allocation strategies during emergencies. For example, on the homepage of the GUI, the whole network topology of the live streaming services can be visualized, including the live video providers, edge nodes, data centers, and feasible communication links. On the page dedicated to displaying SLAs, managers have the flexibility to alter the classification standards of SLAs simply by adjusting the values of the relevant QoS metrics in their respective text boxes. Moreover, a dedicated page visualizes traffic request data, enabling managers to easily view or export historical request data within a specific time slot.

Prevailing OR Techniques

As we state above, the development of the GSCO system was enabled exclusively through the application of various OR techniques. This realization has been embraced across Huawei Cloud from C-level executives (i.e., senior management) to frontline engineers, affirming OR as a pivotal technology capable of enhancing cost efficiency and fostering productivity. The successful implementation of OR techniques in our projects has encouraged more departments within Huawei Cloud to incorporate them into their research and development efforts. A testament to this ripple effect is that another team within Huawei Cloud embarked on a project to tackle virtual machine scheduling issues using mixed-integer programming (MIP) models.

Portability

Currently, the *GSCO* system has primarily been deployed for traffic allocation in Huawei Cloud's live streaming services. It has also been successfully adopted by other companies as a solution for their live streaming business. We have also begun applying the *GSCO* system to other cloud media services, particularly the content delivery network (CDN) service and real-time communication network (RTC) service. This extension is driven by the fact that these services have major operational expenses tied to the bandwidth cost and the traffic allocation problems they face can be addressed using similar mathematical models. For instance, under the common 95th percentile billing scheme, minimizing bandwidth costs in these services can be expressed as distinct MIP models, but the objective functions bear a significant resemblance to each other. The QoS constraints may vary, but these variations are often slight and related to specific parameters. Therefore, we can apply the same methodologies and philosophies that informed the design of the *GSCO* system to CDN and RTC services, also enabling significant bandwidth cost savings for these services.

Summary

This paper discusses the development and implementation of the *GSCO* system, an innovative solution designed to optimally reduce the network bandwidth cost while maintaining the QoS level in Huawei Cloud's live streaming business. By incorporating various OR methodologies and machine learning methods, the *GSCO* system dynamically allocates network traffic, resulting in a significant decrease in operational expenses. In the face of multiple technical and practical hurdles, the system has been exceptionally successful, reducing the network bandwidth cost by 30% and saving more than $49.6 million from Q1 2020 to Q3 2022. In addition, it has notably expanded Huawei Cloud's market share, with its peak bandwidth growing from an initial 1.5 Tbps to a considerable 16 Tbps. Huawei Cloud has begun to extend the *GSCO* system to other cloud media services, such as CDN and RTC services, which not only underscores the portability of the *GSCO* system but also promises additional cost reductions and improvements in efficiency.

Acknowledgments

The authors thank Zohar Strinka and Xinhui Zhang for constructive comments and feedback when verifying their submission for the 2023 Franz Edelman Award competition. The authors are grateful to Aaron Burciaga, Zohar Strinka, and Xinhui Zhang for voluntarily coaching their team when preparing for the final presentation.

Appendix. Model Formulation and Details of Optimization Methods

Model Parameters

- Let $I = \{1, 2, \ldots, m\}$ be the index set of all nodes abstracted from edge regions, and $|I| = m$, where $|\cdot|$ denotes the cardinality of a set. We use the letter i to indicate the index of the edge regions.

- Let $J=\{m+1,m+2,\dots,m+n\}$ be the index set of all nodes abstracted from edge nodes, and $|J|=n$. We use the letter j to indicate the index of the edge nodes.
- Let $V=I\cup J$ be the index set of all nodes in the overlay network of the cloud edge, and $|V|=m+n$.
- Let $E=\{(i,j)|i\in I, j\in J\}$ be the set of edges abstracted from feasible network connections that satisfy QoS and ISP requirements, and $|E|\le mn$.
- Let $T=\{1,2,\dots p\}$ be a series of time points during a billing cycle, and $|T|=p$. We use the letter t to indicate the index of the time points.
- Let $[t,t+1], t\in T$ be a five-minute time slot in the billing cycle T.
- Let $G=(V,E,T)$ be the topology of the overlay network of the cloud edge.
- Let $c_j>0$ be the bandwidth capacity of the edge node $j\in J$.
- Let $d_i^{(t)}\ge 0$ be the traffic demand of the edge region $i\in I$ in the time slot $[t,t+1], t\in T$.
- Let $u_j>0$ be the unit price of the bandwidth at the edge node $j\in J$.

Decision Variables

To find the most cost-effective traffic allocation through QoS-satisfied connections between edge regions and edge nodes in our problem, we introduce decision variables for the traffic across each feasible connection throughout the billing cycle. The definition is as follows:

- Let $x_{ij}^{(t)}\ge 0$ be the traffic assigned to the feasible connection $(i,j)\in E$ in $[t,t+1], t\in T$.

The total number of variables equals $|E|\times|T|$.

Objective Function

We aim to minimize the total bandwidth cost accrued at the cloud edge with the 95th percentile billing scheme while fulfilling the QoS requirements. Considering that the QoS requirements are implicitly considered, with traffic demands always catered to via QoS-compliant connections, our focus in the objective function is exclusively on minimizing the total bandwidth cost. Then, we have:

$$\min\sum_{j\in J}\left(u_j\cdot Q_{95}\left(\left\{\sum_{(i,j)\in E}x_{ij}^{(t)}\right\}_{t=1}^{p}\right)\right), \tag{A.1}$$

where $Q_{95}(\cdot)$ is the operation to determine the 95th percentile of a set of numbers.

Constraints

Our traffic allocation system has the following constraints:

$$\sum_{(i,j)\in E}x_{ij}^{(t)}=d_i^{(t)},\qquad \forall i\in I, t\in T; \tag{A.2a}$$

$$\sum_{(i,j)\in E}x_{ij}^{(t)}\le c_j,\qquad \forall j\in J, t\in T; \tag{A.2b}$$

$$x_{ij}^{(t)}\ge 0,\qquad \forall (i,j)\in E, t\in T. \tag{A.2c}$$

We interpret these constraints as follows:

- Constraint (A.2a) mandates that every access request must be acknowledged and served.
- Constraint (A.2b) states that the bandwidth capacity of each edge node should not be exceeded.
- Constraint (A.2c) specifies the domain of decision variables.

Additionally, both Constraints (A.2a) and (A.2b) stipulate that all demands must be met through viable network connections E.

The deterministic model we present offers a vital framework for confronting the dynamic traffic allocation problem. Although it simplifies the dynamic intricacies inherent to network systems, it is essential in investigating the characteristics of the problem and designing effective offline algorithms. Although this model forms the core of our problem-solving approach, it does not operate in isolation. Instead, it is supplemented by auxiliary mechanisms meticulously designed to manage the problem's dynamic nature, details of which are further expounded in the body of the paper.

Optimization Methods

We present the application of the generalized primal-dual algorithm (He et al. 2022) and the balanced augmented Lagrangian method (He and Yuan 2021) for solving the MCNFP, which is a critical step in the proposed offline solver modules.

MCNFP

To solve the offline cost-effective traffic allocation problem with the 95th percentile billing scheme, which we demonstrate in Equations (A.1) and (A.2a)–(A.2c), we adopt a linear relaxation of the complex objective function. Specifically, we select a representative time point $t^*\in T$ and conduct the traffic allocation, which we formulate as the MCNFP:

$$\min\sum_{j\in J}\left(u_j\cdot\sum_{(i,j)\in E}x_{ij}^{(t^*)}\right) \tag{A.3}$$

$$\text{subject to}\sum_{(i,j)\in E}x_{ij}^{(t^*)}=d_i^{(t^*)},\qquad\forall i\in I; \tag{A.4}$$

$$\sum_{(i,j)\in E}x_{ij}^{(t^*)}\le c_j,\qquad \forall j\in J; \tag{A.5}$$

$$x_{ij}^{(t^*)}\ge 0,\qquad\forall (i,j)\in E. \tag{A.6}$$

Define the vector $\mathbf{u}$:

$$\mathbf{u}=[\underbrace{\underbrace{u_1,u_2,\dots,u_n}_{n},\underbrace{u_1,u_2,\dots,u_n}_{n},\dots,\underbrace{u_1,u_2,\dots,u_n}_{n}}_{m}]^{\top}_{mn\times 1} \tag{A.7}$$

the vector $\mathbf{x}$:

$$\mathbf{x}=\Big[x_{11}^{(t^*)},x_{12}^{(t^*)},\dots,x_{1n}^{(t^*)},x_{21}^{(t^*)},x_{22}^{(t^*)},\dots,x_{2n}^{(t^*)},\dots,x_{m1}^{(t^*)}, x_{m2}^{(t^*)},\dots,x_{mn}^{(t^*)}\Big]^{\top}_{mn\times 1}, \tag{A.8}$$

the vector $\mathbf{d}$:

$$\mathbf{d}=[d_1,d_2,\dots,d_m]^{\top}_{m\times 1}, \tag{A.9}$$

and the vector $\mathbf{c}$:

$$\mathbf{c}=[c_1,c_2,\dots,c_n]^{\top}_{n\times 1}. \tag{A.10}$$

The objective function (A.3) is equivalent to $\mathbf{u}^\top \mathbf{x}$. In addition, define an indicator function $\sigma(\cdot)$ as follows:

$$\sigma(i,j) = \begin{cases} 1, & \text{if} \quad (i,j) \in E; \\ 0, & \text{otherwise.} \end{cases} \tag{A.11}$$

Then, define a $m \times mn$ matrix $\mathbf{\Psi}$ as

$$\mathbf{\Psi} = \begin{bmatrix} \sigma(1,1) & \sigma(1,2) & \cdots & \sigma(1,n) & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \sigma(2,1) & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & \sigma(m,1) & \cdots & \sigma(m,n) \end{bmatrix}_{m \times mn}, \tag{A.12}$$

and an $n \times mn$ matrix $\mathbf{\Phi}$ as

$$\mathbf{\Phi} = \begin{bmatrix} \sigma(1,1) & 0 & \cdots & 0 & \cdots & \sigma(m,1) & 0 & \cdots & 0 \\ 0 & \sigma(1,2) & \cdots & 0 & \cdots & 0 & \sigma(m,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \cdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma(1,n) & \cdots & 0 & 0 & \cdots & \sigma(m,n) \end{bmatrix}_{n \times mn}. \tag{A.13}$$

Then, the compact formulation of the MCNFP is written as

$$\min \quad \mathbf{u}^\top \ \mathbf{x} \tag{A.14}$$

$$\text{subject to} \quad \begin{bmatrix} \mathbf{\Psi} \\ -\mathbf{\Phi} \end{bmatrix} \mathbf{x} \geq \begin{bmatrix} \mathbf{d} \\ -\mathbf{c} \end{bmatrix}, \tag{A.15}$$

$$\mathbf{x} \geq \mathbf{0}, \tag{A.16}$$

where $\mathbf{0}$ is a $mn \times 1$ all-zeros vector, and $\geq$ denotes elementwise inequalities. For simplicity and clarity, we denote $\begin{bmatrix} \mathbf{\Psi} \\ -\mathbf{\Phi} \end{bmatrix}$ by $\mathbf{A}$ and $\begin{bmatrix} \mathbf{d} \\ -\mathbf{c} \end{bmatrix}$ by $\mathbf{b}$ in the following.

The Generalized Primal-Dual Algorithm for the MCNFP

Following He et al. (2022), we first write down the Lagrangian function of the MCNFP (A.14)–(A.16):

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{u}^\top \mathbf{x} - \boldsymbol{\lambda}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}), \tag{A.17}$$

where $\boldsymbol{\lambda} \in \mathbb{R}_+^{m+n}$ is the Lagrangian multiplier. Then, the iterative scheme of the generalized primal-dual algorithm reads as

$$\begin{cases} \mathbf{x}^{k+1} = \arg\min \left\{ \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^k) + \frac{r}{2} \|\mathbf{x} - \mathbf{x}^k\|^2 | \mathbf{x} \in \mathbb{R}_+^{mn} \right\}, & \text{(A.18)} \\ \overline{\mathbf{x}}^{k+1} = 2\mathbf{x}^{k+1} - \mathbf{x}^k, & \text{(A.19)} \\ \boldsymbol{\lambda}^{k+1} = \arg\max \left\{ \mathcal{L}(\overline{\mathbf{x}}^{k+1}, \boldsymbol{\lambda}) - \frac{s}{2} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^k\|^2 | \boldsymbol{\lambda} \in \mathbb{R}_+^{m+n} \right\}, & \text{(A.20)} \end{cases}$$

where $r > 0$ and $s > 0$ are parameters, and we can determine the less restrictive condition of $r \cdot s$ by exploring the structure of $\mathbf{A}$ to improve the efficiency of the algorithm. For example, He et al. (2022) leverage a heuristic to set $r \cdot s$ as the average of the trace (instead of the trace itself) of the matrix $\mathbf{A}^\top \mathbf{A}$ for solving the classic assignment problem. In the context of live streaming, the structure of $\mathbf{A}$ is related to the feasible connections between edge regions and edge nodes. In addition, both subproblems (A.18) and (A.20) have closed-form analytical solutions and thus are easy to solve.

The Balanced Augmented Lagrangian Method for the MCNFP

Following He and Yuan (2021), the iterative scheme of the balanced augmented Lagrangian method reads as

$$\begin{cases} \mathbf{x}^{k+1} = \arg\min \left\{ \mathbf{u}^\top \mathbf{x} + \frac{r}{2} \|\mathbf{x} - \mathbf{q}_0^k\|^2 | \mathbf{x} \in \mathbb{R}_+^{mn} \right\}, & \text{(A.21)} \\ \boldsymbol{\lambda}^{k+1} = \arg\min \left\{ \frac{1}{2} (\boldsymbol{\lambda} - \boldsymbol{\lambda}^k)^\top \mathbf{H}_0 (\boldsymbol{\lambda} - \boldsymbol{\lambda}^k) \right. & \\ \qquad\qquad \left. + (\mathbf{s}_0^k)^\top \boldsymbol{\lambda} | \boldsymbol{\lambda} \in \mathbb{R}_+^{m+n} \right\}, & \text{(A.22)} \end{cases}$$

where $\mathbf{q}_0^k = \mathbf{x}^k + (1/r)\mathbf{A}^\top \boldsymbol{\lambda}^k$, $\mathbf{s}_0^k = \mathbf{A}(2\mathbf{x}^{k+1} - \mathbf{x}^k) - \mathbf{b}$, $\mathbf{H}_0 = (1/r) \mathbf{A}\mathbf{A}^\top + \delta \mathbf{I}$, and $r > 0$ and $\delta > 0$ are parameters. Note that both subproblems (A.21) and (A.22) have analytical solutions and thus are easy to solve.

References

Box GE, Jenkins GM (1968) Some recent advances in forecasting and control. *J. R. Statist. Soc. Ser. C. Appl. Statist.* 17(2):91–109.

Cunningham WH (1976) A network simplex method. *Math. Programming* 11(1):105–116.

Grand View Research (2022) Video streaming market worth $330.51 billion by 2030. Accessed August 30, 2022, https://www.grandviewresearch.com/press-release/global-video-streaming-market.

Harrison A, Skipworth H, van Hoek RI, Aitken J (2019) *Logistics Management and Strategy: Competing Through the Supply Chain* (Pearson, London).

He B, Yuan X (2018) A class of ADMM-based algorithms for three-block separable convex programming. *Comput. Optim. Appl.* 70(3): 791–826.

He B, Yuan X (2021) Balanced augmented Lagrangian method for convex programming. Preprint, submitted August 19, https://arxiv.org/abs/2108.08554.

He B, Ma F, Xu S, Yuan X (2022) A generalized primal-dual algorithm with improved convergence condition for saddle point problems. *SIAM J. Imaging Sci.* 15(3):1157–1183.

iiMedia Research (2022) iiMedia Report—Development Status and Market Research Analysis Report of China's Live Streaming Industry in 2022. Accessed August 30, 2022, https://www.iimedia.cn/c400/84858.html.

Jalaparti V, Bliznets I, Kandula S, Lucier B, Menache I (2016) Dynamic pricing and traffic engineering for timely inter-datacenter transfers. *Proc. 2016 ACM SIGCOMM Conf.* (ACM, New York), 73–86.

Kleinerman K (2022) Cloud computing in the media and entertainment industry. Accessed December 13, 2022, https://www.ridge.co/blog/cloud-computing-in-the-media-and-entertainment-industry/#cloud-media-processing.

Lambert D, Stock JR, Ellram LM (1998) *Fundamentals of Logistics Management* (McGraw-Hill/Irwin, New York).

Liu R (2022) China's overall cloud computing market forecast, 2021–2025. Accessed December 13, 2022, https://www.idc.com/getdoc.jsp?containerId=CHE47428121.

Magnanti TL, Wong RT (1984) Network design and transportation planning: Models and algorithms. *Transportation Sci.* 18(1):1–55.

Marinescu DC (2022) *Cloud Computing: Theory and Practice* (Morgan Kaufmann, Cambridge, MA).

Mladenović N, Hansen P (1997) Variable neighborhood search. *Comput. Oper. Res.* 24(11):1097–1100.

Precedence Research (2022) Cloud computing market size to hit US $1,614.1 billion by 2030. Accessed August 30, 2022, https://www.globenewswire.com/en/news-release/2022/05/13/2443081/0/en/cloud-computing-market-size-to-hit-us-1-614-1-billion-by-2030.html.

Shi Q, Yin J, Cai J, Cichocki A, Yokota T, Chen L, Yuan M, Zeng J (2020) Block Hankel tensor ARIMA for multiple short time series forecasting. *Proc. Conf. AAAI Artif. Intell.* 34(04):5758–5766.

Singh R, Agarwal S, Calder M, Bahl P (2021) Cost-effective cloud edge traffic engineering with Cascara. *Proc. 18th USENIX Sympos. Networked Systems Design and Implementation (NSDI 21)* (USENIX, Berkeley, CA), 201–216.

SteadieSeifi M, Dellaert NP, Nuijten W, Van Woensel T, Raoufi R (2014) Multimodal freight transportation planning: A literature review. *Eur. J. Oper. Res.* 233(1):1–15.

Stock JR, Lambert DM (2001) *Strategic Logistics Management*, vol. 4 (McGraw-Hill, Irwin Boston).

Telecom Review (2021) Huawei Cloud ranked world's 5th largest IaaS provider. Accessed August 30, 2022, https://www.telecomreview.com/articles/cloud-and-enterprise-business/5196-huawei-cloud-ranks-as-world-s-5th-largest-iaas-provider.

Yang C, You J, Yuan X, Zhao P (2022) Network bandwidth allocation problem for cloud computing. Accessed August 30, 2022, https://arxiv.org/pdf/2203.06725v1.pdf.

Yokota T, Erem B, Guler S, Warfield SK, Hontani H (2018) Missing slice recovery for tensors using a low-rank model in embedded space. Accessed August 30, 2022, https://arxiv.org/pdf/1804.01736.pdf. *Proc. IEEE Conf. Comput. Vision Pattern Recognition* (IEEE, Piscataway, NJ), 8251–8259.

Xiaoming Yuan is professor and head of the Department of Mathematics at The University of Hong Kong, and the Chief Scientist of the Algorithm Innovation Laboratory of Huawei Cloud. His research interests include optimization, cloud computing, AI, and optimal control. He has been included in the Clarivate Analytics List of Highly Cited Researchers multiple times. He led the team and developed a series of core algorithms used in the *GSCO* system. His research on operator splitting algorithms is a solid foundation for the success of the *GSCO* system.

Pengxiang Zhao is currently pursuing his PhD degree at The University of Hong Kong, under the mentorship of Xiaoming Yuan. He specializes in optimization, machine learning, and cloud computing, with a keen focus on the intersection of these domains. He has garnered significant expertise in virtual machine scheduling and the design of bandwidth allocation algorithms.

Hanyu Hu is a PhD candidate at The University of Hong Kong, supervised by Xiaoming Yuan. His research interests primarily lie in the fields of optimization, machine learning, and cloud computing. He has acquired extensive expertise in designing bandwidth allocation algorithms, optimizing network resources and enhancing the performance of cloud-based systems.

Jintao You received his PhD degree from the Department of Industrial Engineering at Tsinghua University. He previously served as a research fellow at the Department of Industrial Engineering at the National University of Singapore. He collaborated with Xiaoming Yuan on algorithm design for bandwidth allocation problems in live streaming while at the Algorithm Innovation Laboratory of Huawei Cloud. He is currently a research scientist at the Shenzhen Research Institute of Big Data.

Changpeng Yang holds a joint PhD degree from Nanyang Technological University and the University of California, Berkeley. He was a visiting scholar at Stanford University. He is now the director of the Media Innovation Laboratory and architect of the *GSCO* system at Huawei Cloud. He was the head of the resource scheduling algorithm team at SF Express, and undertook SF network planning, intra-city express network, national trunk/branch line optimization, and other projects.

Wen Peng obtained his master's degree from the University of South California. He is an engineer in the Algorithm Innovation Laboratory at Huawei Cloud and an algorithm expert of the *GSCO* project. He has rich experience in the implementation of the Group's capital planning, tax planning, and cloud resource scheduling. He is responsible for the scheduling algorithms, system design, and development of the *GSCO* project.

Yonghong Kang was an expert in the media service field. He has rich technological experience in big data, cloud resource scheduling, and cost control in cloud computing.

Kwong Meng Teo obtained his PhD degree from the Massachusetts Institute of Technology. He was an assistant professor at the Department of Industry Engineering, National University of Singapore, and is now a senior scientist at Huawei Cloud. He is an expert in solving engineering problems with operations research techniques.