Impact of Operator Splitting Algorithms in Huawei Cloud's Live Streaming Business

Xiaoming Yuan
Department of Mathematics
Faculty of Science

April 25, 2023

Summary of the Impact

- A series of algorithms were applied to allocate access traffic for live streaming business in cloud computing and validated in practice. They were embedded into a bandwidth allocation system named GSCO, which aims to minimize the total bandwidth cost while providing high-quality services.
- The GSCO has been adopted by Huawei Cloud for its B2B live streaming services since 2020 and has helped Huawei Cloud save about 30% of the total bandwidth cost, with a total of more than 49.6 million US dollars over the past two years.
- Besides, the metrics of quality of experiences, such as stall time, are optimized.

Underpinning Research

Background

- Live streaming delivers video in real time and supports applications such as sports broadcasting and interactive entertainment. The main components of live streaming based on cloud computing include Internet-enabled devices, live streaming platforms, and cloud service providers (CSPs).
- The recent explosive adoption of live platforms, especially against the backdrop of the COVID-19 pandemic, has elevated the prominence of live streaming. As of June 2021, China's online live streaming users reached 638 million. Besides, it is estimated that the market size of the global live streaming industry will climb from 59.14 billion US dollars in 2021 to more than 330 billion US dollars by 2030.
- The boom of the cloud computing and live streaming market emanates opportunities and challenges for CSPs. One fundamental problem is saving CSP's bandwidth cost, which is charged by Internet Service Providers (ISPs) using the 95th percentile billing scheme. Note that the bandwidth cost comprises a significant proportion of the total operational cost.

Underpinning Research

Background

- Operator splitting algorithms, including the classic alternating direction method of multipliers (ADMM), are particularly efficient for solving linearly constrained separable convex optimization problems.
- Because of the relatively low per-iteration computational cost and the ability to exploit sparsity in the problem data, operator splitting algorithms are particularly suitable for large-scale optimization.
- We have been dedicated to researching operator splitting algorithms for decades, and some of his results have become classic and state-of-the-art in the area. We proved the first O(1/n) convergence rate of ADMM in [Ref.1], gave the first counter example to show the possible failure of convergence when ADMM is extended to multi-block separable convex optimization problems in [Ref. 2], proposed a series of ADMM-like algorithms for multiple-block separable convex programming problems in [Ref 4], proposed the generalized primal-dual algorithm in [Ref 5] for saddle point problems, and reshaped the classic augmented Lagrangian method in a balanced way in [Ref 6].

Underpinning Research

Objectives

Our main objective is to design efficient algorithms for solving various bandwidth allocation problems and develop a comprehensive system for the live streaming business in cloud computing that minimizes the bandwidth cost while maintaining service quality. In particular, we were required to:

1. design an overall framework for modeling the bandwidth allocation systems,
2. formulate optimization models precisely,
3. analyze specific mathematical features and favorable structures of the proposed models,
4. develop core algorithms that can handle large-scale models and meet tight time constraints,
5. design a user-friendly and implementable system for real bandwidth allocation problems,
6. validate and improve the performance of the system for real live streaming business.

Underpinning Research

Roles: Played as the leader of the project, particularly in the problem analysis, architecture design, and algorithm design

- 2020 Q1: We showed our analysis of the total cost for the live streaming services and demonstrated the feasibility of our first roadmap of the GSCO system by testing some previous datasets.
- 2020 Q2: We implemented a traffic forecaster which is based on various machine learning techniques, and a network topology planner to create feasible communication links between the edge regions and the edge nodes to guarantee the quality of service.
- 2021 Q1: We deployed an offline solver to produce monthly bandwidth allocation by a two-phase strategy. Algorithms such as [Ref. 4-6] have been heavily applied for this module.
- 2021 Q3: We repaired some vulnerabilities in engineering so that the deployed modules interact better.
- 2022 Q1: We deployed an online solver to allocate traffic in real time. We applied several heuristic strategies that can be executed within milliseconds.
- After 2022 Q1: The GSCO system is substantially complete, and we mainly conduct vulnerability repair and general updates.

Underpinning Research

Contextual information about this area of research

- Traffic allocation problems in the live streaming business of cloud computing are different from other traffic engineering problems in traditional scenarios such as logistics management and transportation planning mainly because of the complexity of the 95th percentile billing scheme and many peculiar requirements, such as replicability of data packages and fast response time constraints (i.e., in milliseconds).
- In addition, there is little literature talking about cost-effective traffic allocation problems of cloud computing application scenarios. Microsoft proposed a mixed integer linear programming (MILP) model for inter-domain traffic allocation problems.
- Their method relies on a commercial MILP solver, which runs 15 hours to find a feasible solution. Besides, only traffic allocation between data centers and three ISPs was discussed in Microsoft CASCARA which is a relatively minor instance compared with Huawei Cloud's live streaming business.

Underpinning Research

Innovativeness

- The GSCO system is a novel intelligent bandwidth allocation system, which is highly automated, user-friendly, and transportable to a wide range of media services.
- The developed state-of-the-art optimization algorithms were implemented and were able to solve corresponding large-scale optimization problems within ten seconds.
- Benefiting from our emergency response mechanism in the online solver, it can adapt to traffic uncertainties in real business to meet demand bursts and guarantee user experience.
- We innovatively tackled the complexity of the 95th percentile billing objective function and the sheer problem scale by various Operations Research techniques such as relaxations of the mathematical model and neighborhood search algorithms.

Underpinning Research

Significance

- Cost Savings: The GSCO system's intelligent bandwidth allocation and optimization techniques have led to substantial cost savings, with more than $49.6 million saved in network bandwidth costs.
- Enhanced Performance: The system's ability to increase peak bandwidth capacity from 1.5 Tbps to 16 Tbps demonstrates its effectiveness in improving overall performance and scalability.
- Adaptability: The research findings emphasize the GSCO system's emergency response mechanism, which allows it to adapt to traffic uncertainties and meet demand surges, ensuring a consistently high-quality user experience.
- Application in Diverse Industries: The system's transportability and successful implementation in companies like China Mobile show its potential for adoption in various industries, including telecommunications and media services.
- Labor Efficiency: The GSCO system's automation, visualization, and ease of use have led to a significant reduction in labor efforts, with 98.6% labor-saving reported.

Knowledge to be Exchanged

- Specific knowledge exchanged includes advanced optimization algorithms, state-of-the-art machine learning techniques, and innovative approaches to address complex bandwidth allocation challenges. These insights stem from rigorous research, development, and real-world applications, offering valuable information to industry professionals and academics alike.
- The knowledge generated in the GSCO project is curated through thorough analysis, evaluation, and testing of various methodologies and techniques. It is then disseminated through the competition of the Franz Edelman Award at the INFORMS Business Analytic Conference 2023.

Knowledge to be Exchanged

The GSCO system shares valuable insights and ideas, particularly in the areas of operator splitting algorithms and interdisciplinarity, which are being applied in practice across various industries:

- Operator Splitting Algorithms: The GSCO system leverages innovative operator splitting algorithms to efficiently solve complex optimization problems in bandwidth allocation. These algorithms break down large-scale problems into smaller, more manageable subproblems, allowing for a more efficient and accurate solution. In practice, this approach enables organizations to make data-driven decisions, optimize network resources, and reduce costs without sacrificing performance.
- Interdisciplinarity: The GSCO system embraces interdisciplinarity by combining expertise from various fields, such as Operations Research, machine learning, graph theory, scheduling, network flow problems, and continuous optimization. This interdisciplinary approach fosters innovation and ensures that the system remains flexible and adaptable to the evolving needs of different industries. In practice, the GSCO system's interdisciplinarity enables it to address diverse network optimization challenges and cater to a wide range of applications, such as telecommunications, media services, and cloud computing.

Engagement

Engagement Process

- Identifying the Problem: The motivation behind the GSCO system began with the recognition of a critical challenge in providing high-quality live streaming services while minimizing bandwidth costs for businesses and end-users.
- Initial Analysis: The research team conducted a thorough analysis of the total cost for live streaming services, identifying opportunities for optimization and cost reduction, which served as the foundation for the GSCO system's development.
- Feasibility Study: The team tested previous datasets to demonstrate the feasibility of the GSCO system's first roadmap, building confidence in the project's potential and motivating further research and development.
- Leveraging Advanced Techniques: The GSCO system was designed with a strong emphasis on integrating Operations Research techniques, machine learning solvers, and various operator splitting algorithms developed by us to create an innovative and efficient solution to the bandwidth allocation problem.
- Scalability and Transportability: The motivation to create a versatile solution that could be applied across different industries and markets drove the development of the GSCO system's scalability and transportability features.

Engagement

External Partners: Huawei Cloud

Huawei Cloud provides customers with reliable, secure, and sustainable cloud services. According to Gartner's Market Share: IT service, Worldwide 2020, Huawei Cloud rose to No. 2 in the global IaaS market in China and No. 5 worldwide. To date, Huawei Cloud has launched more than 220 cloud services with 210 technical solutions and has attracted over 30,000 partners and 3 million customers from a broad range of industries, including media entertainment, manufacturing, health care, finance, and logistics.

Impacts Achieved

Beneficiaries

- Telecommunications: Telecom operators, such as China Mobile, have benefitted from the GSCO system's ability to optimize network resources, reduce costs, and improve overall performance. The system allows these operators to manage their bandwidth allocation more effectively, catering to the growing demand for high-quality digital services.
- Media Services: Companies providing live streaming and video-on-demand services can leverage the GSCO system to ensure a seamless user experience by efficiently managing network resources. By minimizing bandwidth costs and guaranteeing service quality, these organizations can stay competitive in a rapidly evolving market.
- Cloud Computing Providers: Cloud service providers can utilize the GSCO system to optimize their network infrastructure, resulting in cost savings and enhanced performance. This allows them to offer more competitive and reliable services to their clients.
- E-commerce and Online Platforms: Businesses operating in the e-commerce sector or offering online services can benefit from the GSCO system's ability to manage network resources effectively. This ensures a smooth user experience for customers, contributing to increased customer satisfaction and retention.
- Government and Public Sector: The GSCO system's applications extend to government agencies and public sector organizations that rely on robust and efficient network infrastructures to deliver essential services and maintain effective communication channels.

Impacts Achieved

Nature and extent of the impact

- Cost Savings: Beneficiaries of the GSCO system have experienced significant cost savings, as the system has helped save about 30% of network bandwidth cost, amounting to more than $49.6 million. These savings have made a direct positive impact on the organizations' bottom lines, allowing them to invest in other areas of growth and development [see submitted reference from Mr. Jia Li].
- Improved Performance: By optimizing network resources, the GSCO system has enabled organizations to achieve a higher level of performance. For instance, Huawei Cloud increased its peak bandwidth from 1.5 Tbps to 16 Tbps over two years, enabling the delivery of better and faster services to their customers.
- Enhanced User Experience: By guaranteeing service quality and minimizing bandwidth costs, the GSCO system has had a direct impact on the end-users' experience. This has resulted in increased customer satisfaction and retention, further contributing to the growth of the organizations using the system.
- Labor Efficiency: The GSCO system's automation and ease of use have led to significant labor-saving benefits, allowing organizations to reallocate their workforce to other essential tasks or areas of growth.

Ongoing Impact: The impact of the GSCO system remains present today, as the system is continuously updated and refined to address the evolving needs of the industries it serves. By staying ahead of the technological curve, the GSCO system continues to provide lasting benefits to its users and beneficiaries.

Impacts Achieved

Evidence

[1] Finalists Selected for the World's Leading Operations Research and Analytics Award: 2023 INFORMS Franz Edelman Award Competition Elevates Research that is Saving Lives, Saving Money and Solving Problems

[https://www.informs.org/News-Room/INFORMS-Releases/Awards-Releases/Finalists-Selected-for-the-World-s-Leading-Operations-Research-and-Analytics-Award-2023-INFORMS-Franz-Edelman-Award-Competition-Elevates-Research-that-is-Saving-Lives-Saving-Money-and-Solving-Problems]

[2] A News Article on People's Daily

[https://wap.peopleapp.com/article/6984879/6842824]

[3] A News Article on Prnewswire

[https://www.prnewswire.com/in/news-releases/huawei-cloud-becomes-a-franz-edelman-award-finalist-301724231.html]

[4] A News Article on Huawei Central

[https://www.huaweicentral.com/huawei-cloud-enters-franz-edelman-award-finals/]

[5] A News Article on TMT News

[http://www.tmtnews.tech/archives/32550]

[6] A News Article on Sina.com

[https://sina.com.hk/news/article/20230118/0/2/21/%E8%8F%AF%E7%82%BA%E9%9B%B2%E5%85%A5%E5%9C%8D%E9%81%8B%E7%B1%8C%E8%88%87%E7%AE%A1%E7%90%86%E5%AD%B8%E6%9C%80%E9%AB%98%E7%8D%8E%E9%A0%85Franz-Edelman-Award%E5%85%A8%E7%90%83%E7%B8%BD%E6%B1%BA%E8%B3%BD-14954545.html]

[7] A testimony letter from Huawei Cloud

References

[1] He, B., & Yuan, X. (2012). On the O(1/n) convergence rate of the Douglas–Rachford alternating direction method. SIAM Journal on Numerical Analysis, 50(2), 700-709.

[https://epubs.siam.org/doi/abs/10.1137/110836936]

[2] Chen, C., He, B., Ye, Y., & Yuan, X. (2016). The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Mathematical Programming, 155, 57-79.

[3] Yang, C., You, J., Yuan, X., & Zhao, P. (2022). Network Bandwidth Allocation Problem for Cloud Computing. arXiv preprint arXiv:2203.06725.

[https://arxiv.org/abs/2203.06725]

[4] He, B., & Yuan, X. (2018). A class of ADMM-based algorithms for three-block separable convex programming. Computational Optimization and Applications, 70, 791-826.

[https://link.springer.com/article/10.1007/s10589-018-9994-1]

[5] He, B., Ma, F., Xu, S., & Yuan, X. (2022). A generalized primal-dual algorithm with improved convergence condition for saddle point problems. SIAM Journal on Imaging Sciences, 15(3), 1157-1183.

[https://epubs.siam.org/doi/abs/10.1137/21M1453463]

[6] He, B., & Yuan, X. (2021). Balanced augmented Lagrangian method for convex programming. arXiv preprint arXiv:2108.08554.

[https://arxiv.org/abs/2108.08554]