International Journal of Artificial Intelligence Trends (IJAIT)
Vol. 4, Issue VIII; No. 58, August, 2025, pp. 605-618

REVIEW OF EXISTING CONGESTION CONTROL TECHNIQUES ACROSS WIRELESS AND CLOUD COMPUTING ENVIRONMENTS

[1*]Ezema Christopher I. [1]Mgbachi C.A.C., [2]Onuigbo Chika M.
[1*,1,2]Department of Electrical Electronic Engineering, Enugu State University of Science and Technology, Agbani, Enugu, Nigeria

Article Info
Received: 23/8/ 2025
Revised: 2/9/2025
Accepted 9/9/2025
Corresponding Authors Email: ikeezema2000@gmail.com
Corresponding Author's Tel:
[1*]+ 2348063631088

ABSTRACT
The growing appreciation and usage of cloud computing has heightened the need for a sound network-management practice, particularly under a dynamic and resource intensive workload. The current systematic literature review (SLR), in its turn, methodically explores the available approaches to congestion-control in wireless, and cloud computing network contexts, with a dual goal of identifying the gaps that exist in the current congestion-control approaches and developing a more dynamic, intelligent-based approach to the same congestion-control, which suits the needs of cloud-based systems. To this extent, this review compiles the results of available researches of the recent past across the subject of AI-driven protocols, decision-based algorithms, and classical congestion-control operations. Interestingly, despite the efficacy demonstrated by the Distributed Congestion Control Protocol (DCCP) and the Markov Decision Processes, these tasks commonly fall short when it comes to addressing cloud-specific environments typified by momentary surges, resource contention, and network congestion due to high concurrent request rates. The systematic analysis thus points out a small extent of scalability, flexibility, and context-awareness of the existing solutions in real-time cloud deployments of congestion control. Collectively, these results highlight the absolute need to develop a powerful and smart congestion-control architecture that has the adaptive capacity to dynamically react to chaotic, variable conditions in the cloud network. The proposed direction contributes to improving quality of service (QoS), enhancing network efficiency, and advancing the design of next-generation congestion control mechanisms in cloud computing environments.

Keywords: Cloud Computing; Congestion Control; Distributed Congestion Control Protocol (DCCP); Network Bottlenecks; Artificial Intelligence

1. INTRODUCTION

Cloud computing is a key enabler of data networks, providing the computational power, storage, and scalability necessary to support the high demands of communication technology. Cloud networks are expected to deliver ultra-fast speeds, low latency, and massive connectivity, which require robust infrastructure (Imam-Fulaniet al., 2023). Cloud computing meets these needs by offering centralized resources that can be dynamically allocated to handle the varying loads and applications typical in communication environments. This integration allows for the efficient management of resources, ensuring that the network can scale up or down based on demand, thereby improving the overall performance and reliability of cloud services (Gaili and Yigit, 2021).

According to Haile et al. (2021), one of the significant advantages of cloud computing is its ability to support network slicing. Network slicing enables the creation of multiple virtual networks within a single physical network, each tailored to specific use cases or applications. Cloud computing provides the necessary platform to host these slices, allowing for the flexible and efficient management of resources (Guo and Yuan, 2021). This capability is particularly important, where different applications such as autonomous vehicles, smart cities, and IoT deviceshave varying requirements in terms of bandwidth, latency, and reliability. By applying cloud computing, network operators can optimize each slice according to the specific needs of the application it supports (Ilievet al., 2019).

Furthermore, cloud computing plays a crucial role in edge computing. Edge computing brings data processing closer to the user or IoT devices, reducing latency and enhancing real-time processing capabilities (Mahawish and Hassan, 2022). By integrating cloud computing with edge nodes, networks can offload data processing tasks from centralized

data centers to the edge of the network. This not only improves the efficiency of data processing but also reduces the amount of data that needs to be transmitted back to the core network, thereby saving bandwidth and reducing latency. This is essential for applications like augmented reality, gaming, and real-time analytics, where immediate responses are critical (Khan et al., 2022).

Cloud computing networks face significant congestion challenges due to inefficient resource allocation, leading to severe technical, economic, human, and social implications. Existing congestion control mechanisms fail to dynamically adapt to varying traffic loads, resulting in high latency, excessive CPU utilization, and suboptimal bandwidth usage. This inefficiency leads to packet loss, degraded Quality of Service (QoS), and system bottlenecks, making it difficult for cloud service providers to maintain optimal performance. From human perspective, excessive network delays lead to user dissatisfaction, reduce productivity, and negatively impact remote work, e-learning, and cloud-based healthcare services that depend on reliable network performance. Socially, inefficient congestion control affects critical services such as online banking, smart grids, and cloud-driven emergency response systems, potentially leading to disruptions in essential public services. Hence, this study presents a review on the various techniques that can be applied for the mitigation of congestion in cloud computing platforms. This review will guide future studies to ascertain some of the effective, adaptable and reliable approaches which can be adopted for the mitigation and control of congestion around systems and to determine the directions for the improvement in future research works.

2. LITERATURE REVIEW

Hamzah and Athab (2022) studied congestion control of Transmission Control Protocol (TCP) using Artificial Intelligence (AI) in a 4G network. While the work explored how AI techniques can be utilized to address congestion problems, the research gap lies in the limited consideration of multiple constraints such as throughput and load factor in the congestion detection process. The study highlighted the potentials of applying AI in solving congestion issues in wireless networks, indicating the need to incorporate these techniques in the principal research on congestion control. However, the specific research gap lies in the lack of solutions that utilize the DCCP approach to detect congestion while considering multiple constraints simultaneously.

Kanellopoulos (2019) provided an overview of congestion control in mobile ad-hoc networks, focusing on enhancing the transmission control protocol over wireless networks. However, the research gap lies in the absence of comprehensive exploration of congestion control schemes that consider various factors like congestion control in the media access control, load-balanced congestion adaptive routing, and congestion control in multipath routing control. The study suggested future research directions to address this gap and adopt diverse congestion control schemes. Therefore, the research gap is the limited investigation of congestion control mechanisms that incorporate multiple constraints and consider different layers of the network architecture.

Haile *et al*. (2021) conducted research on congestion control in 4G cellular networks, utilizing an end-to-end approach to optimize throughput and minimize delay. While the deployment of the Congestion Control Algorithm (CCA) improved network fairness and reduced congestion, the research gap lies in the limited exploration of other congestion control mechanisms beyond the user-to-user interaction. Future studies could further investigate additional congestion control techniques to enhance network performance. Hence, the research gap is the need for exploring alternative congestion control mechanisms that address congestion beyond the user-to-user interaction, considering factors such as network architecture and traffic characteristics.

Magbo*et al*. (2019) surveyed various congestion control mechanisms for networks operating on contention-based and contention-free random-access procedures. However, the research gap lies in the complexity of implementing the proposed collaborative and distributive Q-learning based algorithm, which may hinder practical improvement and implementation. Therefore, the research gap is the need for developing simpler and more practical congestion control mechanisms that offer improved system performance and can be easily implemented in real-world scenarios.

Kashefi and Pourmina (2019) evaluated the effects of LTE network congestion on Voice over IP (VOIP) communication. While the study highlighted the poor quality of service experienced by VOIP due to congestion, the

research gap lies in the limited exploration of congestion management solutions specifically tailored for VOIP communication in LTE networks. The application of an intelligent Radio Access Network (IRAN) framework with congestion management system showed promising results, indicating the potential for improved quality of service. Thus, the research gap is the need for further investigation into congestion control mechanisms designed specifically for VOIP communication in LTE networks, aiming to enhance the quality of service for such applications.

Albanna and Yousefizadeh (2020) conducted research on minimizing congestion in LTE networks using a deep learning approach. The work integrated optimization techniques with deep learning technology to develop a self-organizing tool for congestion interference in actual cellular networks. However, the research gap lies in the limited exploration of alternative deep learning algorithms or architectures that can further enhance congestion reduction and network performance. Therefore, further research is needed to investigate other deep learning approaches and their effectiveness in minimizing congestion in LTE networks.

Adesh and Renuka (2019) proposed a congestion feedback mechanism to avoid queue overflow and reduce delays in queuing of an eNodeB LTE network. Although the implemented mechanism successfully increased the packet delivery fraction and maintained network throughput, the research gap lies in the limited investigation of the mechanism's performance in scenarios with dynamic and fluctuating congestion levels. Therefore, further research is required to evaluate the effectiveness of the congestion feedback mechanism under varying and unpredictable congestion conditions.

Paranjothi*et al*. (2020) surveyed various congestion techniques applied in networks, such as event-driven, priority-based, and measurement-based approaches. While the study identified challenges and recommended optimization techniques to improve performance, the research gap lies in the limited evaluation of these techniques in real-world network environments. Therefore, further research is necessary to assess the practical applicability and performance of the recommended optimization techniques in diverse network scenarios. Mouna (2022) researched the mitigation of congestion and management of network selection in a heterogeneous C-ITS communication architecture. The study proposed a RAT selection framework called Distributed Context Aware Radio Technology selection (DICART) to improve decision-making during connectivity. However, the research gap lies in the limited investigation of the framework's performance in scenarios with high traffic loads and congestion. Therefore, further research is needed to evaluate the effectiveness of the DICART framework under heavy congestion conditions and explore potential optimizations to handle such scenarios more efficiently.

Balasingam *et al*. (2019) utilized collaborative and adaptive signaling to control congestion in wireless networks, integrating it with a reinforcement learning approach. While the collaborative and adaptive signaling controller demonstrated improved reduction in computational time compared to the conventional system, the research gap lies in the limited exploration of the controller's performance in scenarios with highly dynamic and fluctuating network conditions. Further research is required to evaluate the effectiveness and adaptability of the collaborative and adaptive signaling controller in such dynamic network environments.

Guo and Yuan (2021) analyzed various routing calculation and optimization algorithms to address congestion in a congested network. However, the research gap lies in the limited investigation of the scalability and efficiency of these algorithms in large-scale and complex network scenarios. Therefore, further research is necessary to evaluate the performance and scalability of the routing calculation and optimization algorithms in more realistic network environments.

Ilie*et al*. (2019) proposed a congestion control scheme based on various TCP schemes. While the study demonstrated the balanced performance of the Cubic TCP scheme, the research gap lies in the limited exploration of other TCP schemes and their comparative performance in congested network scenarios. Further research is needed to evaluate the effectiveness of different TCP schemes in handling congestion and improving network performance.

Umoh *et al*. (2020) presented a call admission control technique for uncertainty and congestion elimination in 4G networks using Interval Type-2 Intuitionist Fuzzy Logic (IT2IFL). Although the technique showed better performance compared to the conventional system, the research gap lies in the limited investigation of the technique's scalability and performance in scenarios with a high number of concurrent users and complex network

conditions. Further research is necessary to evaluate the scalability and effectiveness of the IT2IFL-based technique in large-scale and dynamic network environments.

Figuera *et al*. (2019) developed a low-complexity mechanism for congestion notification in rural IPSec-enabled heterogeneous backhaul networks. While the mechanism achieved high accuracy in backhaul notification, the research gap lies in the limited evaluation of the mechanism's performance in scenarios with varying network topologies and traffic patterns. Further research is required to assess the mechanism's robustness and accuracy in diverse rural network environments.

Tshilongamulenzhe *et al*. (2020) presented an algorithm for traffic congestion management in a wireless sensor network, integrating routing congestion control and traffic rate adjustment algorithms. However, the research gap lies in the limited investigation of the algorithm's performance under dynamic and fluctuating congestion conditions. Further research is required to assess the algorithm's effectiveness and adaptability in scenarios with varying levels of congestion.

Shen and Wu (2022) proposed a robust dynamic tariff method for congestion management in distribution networks. However, the research gap lies in the limited investigation of the method's performance under various network constraints and scenarios. Further research is required to evaluate the effectiveness and robustness of the dynamic tariff method in diverse distribution network environments. Khan *et al*. (2022) presented a robust multi-objective congestion management technique in distribution networks, focusing on load flexibility scheduling. While the technique showed potential in alleviating congestion and reducing power consumption, the research gap lies in the limited exploration of the technique's performance under uncertain and dynamic load conditions. Further research is needed to evaluate the technique's effectiveness in scenarios with varying load patterns and uncertainty.

3. RESEARCH METHODOLOGY

The Systematic Literature Review (SLR) approach used in this paper guaranteed an adequately planned, hybrid, objective analysis of the available body of research on congestion-control in wireless and cloud computing. Unlike a traditional review, the research has a clear methodology, which included the clearly elaborated research questions, the definition of inclusion and exclusion criteria, the dedicated search within the popular databases (IEEE Xplore, ACM Digital Library and Science Direct), and intense screening and data collection. This process helped in locating pertinent researches which dealt with the different congestion-control measures such as AI-based approach, DCCP, and Markov decision algorithms. SLR showed that, out of a number of investigations looking into congestion control over wireless networks, very few have found to evaluate congestion control issues only in cloud computing conditions, especially in the presence of transient traffic flows, resource contention, large request rates, and network bottlenecks. As such, the SLR provided a strong evidence-based foundation for identifying this gap and justifying the need for an adaptive and intelligent congestion control framework tailored to the dynamic nature of cloud service environments.

4. CONGESTION IN CLOUD COMPUTING ENVIRONMENTS

Congestion in cloud computing environments is a significant challenge that affects the performance and reliability of cloud services. As more users and applications rely on cloud infrastructure for data storage, processing, and service delivery, the risk of network congestion increases (Prateesh et al. 2022). This congestion occurs when the demand for resources, such as bandwidth, processing power, or storage, exceeds the available capacity, leading to delays, packet loss, and reduced quality of service. The dynamic nature of cloud environments, where resources are shared among multiple users and applications, exacerbates this issue, making congestion management a critical aspect of cloud computing (Gibson et al., 2022). Figure 1 presents the architecture of a cloud computing network.

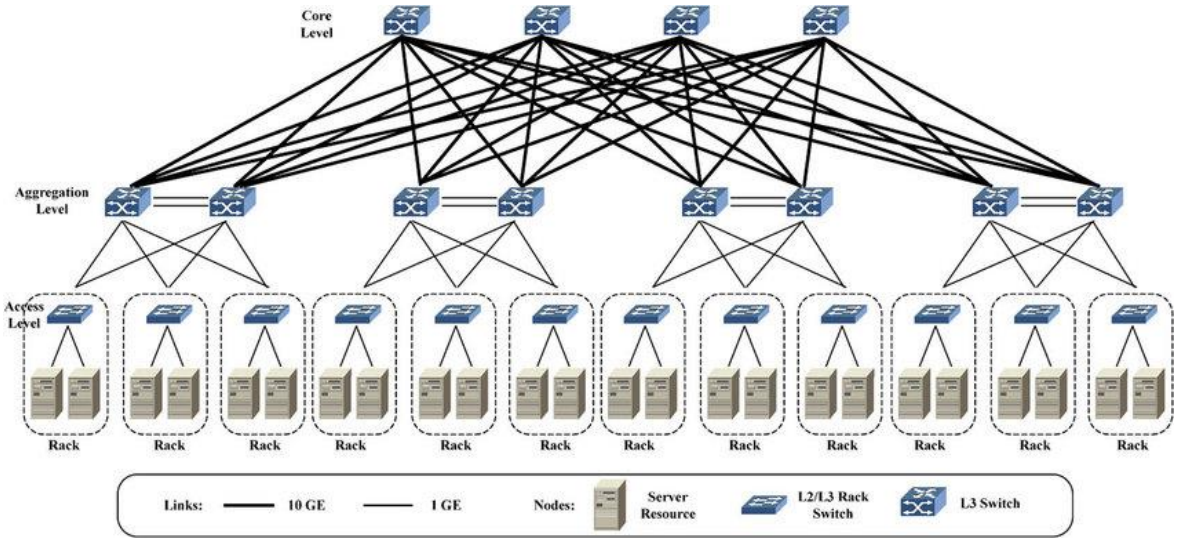

Figure 1: Cloud computing architecture (Gibson et al., 2022)

One of the primary causes of congestion in cloud computing is the high volume of data traffic generated by modern applications (Poutievski et al., 2022). With the rise of big data, IoT devices, and multimedia content, the amount of data being transmitted across cloud networks have grown exponentially. This surge in data traffic can overwhelm network links and processing nodes, leading to bottlenecks that slow down the flow of information. In cloud environments, where multiple tenants share the same physical infrastructure, the competition for bandwidth and computing resources can result in congestion, particularly during peak usage periods (Joshi et al., 2023).

Another factor contributing to congestion is the distributed nature of cloud computing. Cloud services are often delivered across geographically dispersed data centers, requiring data to travel long distances over the internet or private networks (Li et al., 2023). As data packets traverse these networks, they may encounter congestion points, especially at network interconnections or data center gateways. Additionally, the variability in network conditions, such as latency, jitter, and packet loss, can exacerbate congestion issues, leading to inconsistent service quality and degraded user experiences (Arslan et al., 2023).

To mitigate congestion in cloud computing environments, various strategies can be employed. One common approach is the use of traffic management techniques, such as load balancing and traffic shaping. Load balancing distributes incoming requests across multiple servers or network paths to prevent any single resource from becoming overloaded. Traffic shaping involves controlling the flow of data to ensure that network links are not overwhelmed by sudden spikes in traffic. By implementing these techniques, cloud providers can reduce the likelihood of congestion and maintain a higher level of service quality (Bolanowski et al., 2023; Zaher et al., 2021).

Another effective method for managing congestion is the use of scalable architectures and elastic resource provisioning. Cloud environments are inherently flexible, allowing resources to be dynamically allocated based on demand (Zaher et al., 2021). By scaling up resources during periods of high demand and scaling down during low usage, cloud providers can better match resource availability with traffic loads, minimizing the risk of congestion. Additionally, the deployment of edge computing resources can help reduce congestion by processing data closer to the source, thereby offloading traffic from the central cloud infrastructure (Scazzariello et al., 2023).

Overall, congestion in cloud computing environments poses a significant challenge to maintaining service quality and reliability. The growing volume of data traffic, combined with the distributed nature of cloud services, increases the likelihood of congestion-related issues. However, through the use of traffic management techniques, scalable architectures, and edge computing, cloud providers can effectively manage congestion and ensure that their services remain responsive and reliable. As cloud computing continues to evolve and expand, addressing congestion will remain a critical priority to meet the demands of increasingly data-intensive applications and services.

a. Causes of congestion in Cloud computing

Congestion in cloud data centers can stem from various factors that impact the flow of data and the availability of resources. One of the primary causes is the high demand for bandwidth and processing power due to the increasing number of applications, services, and users relying on cloud infrastructure. As more data is generated and processed, especially in real-time applications like streaming, online gaming, and big data analytics, the network and computational resources in a data center can become overwhelmed. This overload can lead to bottlenecks, where the capacity of certain network links or processing nodes is exceeded, causing delays, packet loss, and reduced performance (Xiao et al., 2019).

Another significant cause of congestion is the shared nature of cloud environments. In multi-tenant data centers, multiple users and applications often share the same physical resources, such as servers, storage devices, and network bandwidth (Agarwal et al., 2023). This sharing can lead to resource contention, especially during peak usage times when many users are simultaneously accessing or transmitting large amounts of data. Without effective resource allocation and management strategies, this contention can result in congestion, where the demand for resources exceeds their availability, leading to degraded service quality for all users involved (Namkung et al., 2023).

The architecture of data centers also plays a role in causing congestion. Traditional hierarchical network designs, that rely on a few central switches, or routers to manage traffic, can become bottlenecks as data volumes increase. In such architectures, all data must pass through these central points, which can quickly become congested when traffic loads are high. Additionally, the physical distance between servers and storage devices within a data center can affect the speed and efficiency of data transfers, further contributing to congestion, especially in large-scale data centers where data needs to travel longer distances (Kundel et al., 2021).

Finally, inefficient data management practices, such as improper load balancing, can exacerbate congestion in cloud data centers. If traffic is not evenly distributed across the available resources, certain servers or network links may become overloaded while others remain underutilized. This imbalance can lead to hotspots of congestion, where certain parts of the network or data center are consistently overwhelmed, while others have excess capacity. Furthermore, delays in scaling up resources in response to increased demand can also contribute to congestion, as the data center may not have sufficient capacity to handle sudden spikes in traffic.

b. Methods of congestion management in Cloud computing

Congestion management in cloud computing is essential for maintaining the efficiency and reliability of cloud services, particularly as the demand for these services continues to grow. As cloud environments become more complex and data-intensive, various strategies have been developed to prevent and mitigate congestion. These methods include load balancing, traffic shaping, scalable architectures, edge computing, Quality of Service (QoS) management, and Content Delivery Networks (CDNs) (Kundel et al., 2021). Each of these approaches play a crucial role in ensuring that cloud infrastructure can handle high traffic volumes without degrading service quality. Below, each method is introduced and discussed in detail. Figure 2 presents the method of congestion management in cloud computing.

4.2.1. Load Balancing: Load balancing is a critical congestion management technique in cloud computing that involves distributing incoming network traffic across multiple servers or data paths to ensure no single resource is overwhelmed. By dynamically adjusting the distribution of requests based on current loads, load balancing prevents bottlenecks and ensures that resources are utilized efficiently (Chen et al., 2023)

4.2.2. Traffic Shaping: Traffic shaping, also known as packet shaping, is a congestion management method that regulates the flow of data to ensure that the network is not overwhelmed by sudden spikes in traffic (Wang et al., 2023a). This technique involves controlling the rate at which data packets are transmitted over the network, prioritizing certain types of traffic while delaying others to avoid congestion (Kumar et al., 2020)

4.2.3. Scalable Architectures: Scalable architectures are designed to adjust resource allocation dynamically based on real-time demand, helping to manage congestion effectively. In cloud computing, this often involves elastic resource provisioning, where additional servers, storage, or bandwidth are automatically added to the system as demand increases and scaled down when demand decreases. This elasticity ensures that the cloud infrastructure can handle varying loads without becoming congested (Wang et al., 2023b).

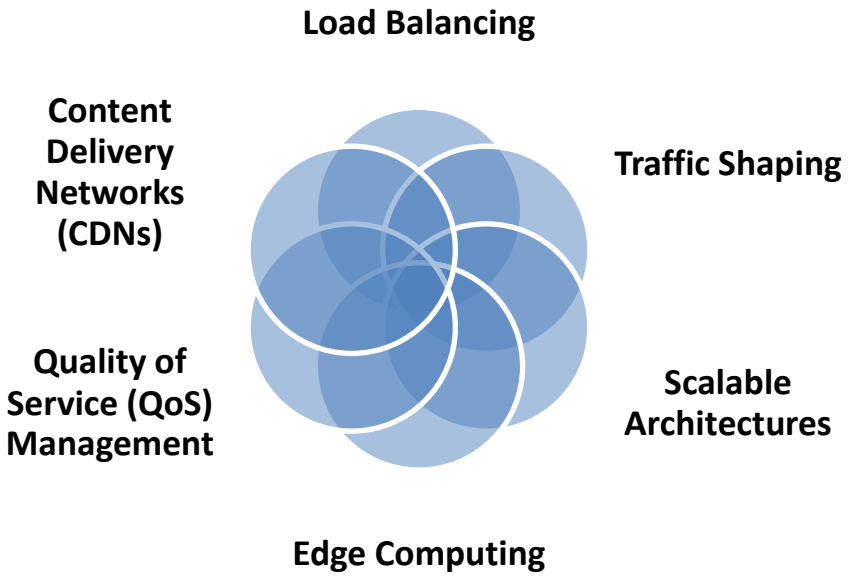

Figure 2: Methods of congestion management in cloud computing

4.2.4. Edge Computing: Edge computing is an approach that helps alleviate congestion in cloud environments by processing data closer to the source, reducing the need to send large volumes of data back to central cloud data centers. By offloading tasks to edge devices or local servers, edge computing minimizes the amount of data that needs to traverse the network, thus reducing potential congestion points (Lim et al., 2023).

4.2.5. Quality of Service (QoS) Management: Quality of Service (QoS) management is a congestion control technique that prioritizes certain types of traffic based on their importance or performance requirements. In cloud computing, QoS settings can be used to ensure that critical applications receive the necessary bandwidth and resources, even during times of heavy congestion. This method involves setting rules and policies that dictate how different types of traffic are handled, such as giving priority to video conferencing over file downloads (Lim et al., 2023; Alizadeh et al., 2020).

4.2.6. Content Delivery Networks (CDNs): Content Delivery Networks (CDNs) help manage congestion by distributing content across a network of geographically dispersed servers, allowing users to access data from a location that is closest to them. This reduces the load on the central cloud data centers and minimizes the distance data needs to travel, which in turn reduces latency and the likelihood of congestion (Zhang et al., 2021).

c. Optimization techniques for congestion management

Optimization techniques are essential for managing congestion in cloud computing environments, where the efficient allocation and utilization of resources can significantly reduce bottlenecks and improve overall system performance. One of the most effective optimization techniques is resource allocation optimization, which involves dynamically adjusting the distribution of computing power, storage, and bandwidth based on real-time demand. By using algorithms that predict usage patterns and allocate resources accordingly, cloud providers can prevent overloading any single part of the infrastructure. This approach minimizes the risk of congestion by ensuring that resources are available where and when they are needed, leading to more balanced and efficient operation (Hauser et al., 2023).

Another powerful optimization technique is load balancing through machine learning. Machine learning algorithms can analyze historical traffic data and predict future congestion points, allowing cloud systems to proactively balance loads before congestion occurs. These algorithms can learn from past network behaviour, identifying patterns that typically lead to congestion and adjusting the routing of data accordingly. By continuously refining

these predictions, machine learning-enhanced load balancing can optimize the distribution of network traffic, reducing the likelihood of congestion and improving the responsiveness of cloud services (Gibson et al., 2022).
Finally, network optimization through Software-Defined Networking (SDN) offers a highly flexible approach to congestion management. SDN allows network administrators to dynamically configure network paths based on current traffic conditions, optimizing the flow of data across the cloud infrastructure. By decoupling the control plane from the data plane, SDN enables more granular control over network resources, allowing for real-time adjustments that can alleviate congestion. This technique is particularly useful in large-scale cloud environments where traditional network management methods might struggle to keep up with rapidly changing traffic demands. Through SDN, cloud providers can optimize network performance, reduce latency, and enhance the overall user experience by effectively managing congestion (Prateesh et al., 2022).

d. Major components of cloud computing

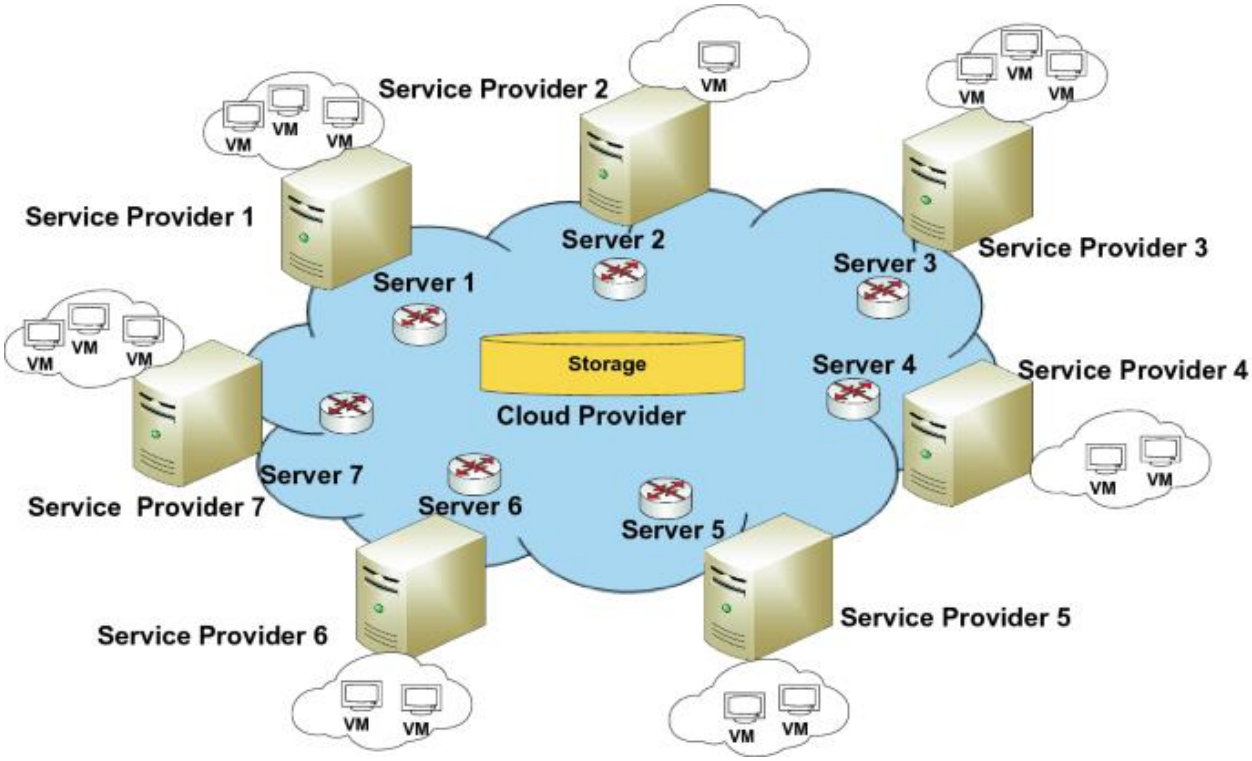

Figure 3: Major Components of cloud computing (Gens, 2019)

a. Application: The application itself, which is the component that end-users will spend most of their time using, is hosted on severs that are remote from the user and can be run in real-time from a thin client that hosts the application through a web browser. The majority of applications that are hosted on clouds are run via browsers.

b. Client: A cloud client is computer hardware and software which relies on the Cloud for application delivery, or which is specifically designed for delivery of cloud services, and which is in either case essentially useless without a Cloud.

c. Infrastructure: Cloud infrastructure (e.g. Infrastructure as a service) is the delivery of computer infrastructure (typically a platform virtualization environment) as a service (Obama, 2019). A cloud infrastructure is the collection of hardware and software that enables the five essential characteristics of cloud computing. The cloud infrastructure can be viewed as containing both a physical layer and an abstraction layer. The physical layer consists of the hardware resources that are necessary to support the cloud services being provided, and typically includes server, storage and network components. The abstraction layer consists of the software deployed across the physical layer, which manifests the essential cloud characteristics. Conceptually the abstraction layer sits above the physical layer (Gens, 2019).

d. **Platform:** A cloud platform facilitates deployment of applications without the cost and complexity of buying and managing the underlying hardware and software layers(Guardian, 2019) e.g. Google App Engine. The cloud platform is referring to the way that applications can be deployed.

e. **Service:** A cloud service (e.g. Web Service) is a software system designed to support interoperable machine-to-machine interaction over a network (Barr, 2019) which may be accessed by other cloud computing components, software or end users directly. It refers to what users can reap from their cloud experience. One of the most popular services in recent years that uses cloud computing would be mapping services e.g. Yahoo Maps.

f. **Storage:** Cloud storage is the delivery of data storage as a service (including database-like services), often billed on a utility computing basis (e.g. per gigabyte per month)(Farber, 2019).

5. DISCUSSIONS

Recent research on wireless and cellular networks, most predominantly 4G/LTE, on congestion control introduces a broad scope of efforts using artificial intelligence (AI), machine learning (ML), and optimization. However, a number of bright observations and pertinent gaps in research have appeared.

5.1. Narrow Addressing of Multi-Network Constraints

Another limitation shared by various studies is the inability to manage the aspect of congestion control comprehensively. Hamzah and Athab (2022) proposed the use of AI-based methods of managing congestion but failed to deliberate on pertinent elements in tandem; throughput, load factor, and fairness. Similarly, Kanellopoulos (2019) has written a paper stressing the necessity of multilayer congestion-control mechanisms without making suggestions that would address all the MAC, routing, and transport layers. The disjointed paradigm tends to limit the management of congestion to specific layers in the network or highly specific metrics of performance.

5.2. Lack of Diversity of AI Models

Intelligent algorithms and deep learning are becoming successively used within scenarios of LTE and vehicles networks (Albanna and Yousefizadeh, 2020; Perez-Murueta et al., 2019), still, little research on the extent and variance of algorithms has been performed. They are designed in many cases in a limited form to fit controlled conditions and to be limited in their scaling up to very dynamic large scale environments (i.e. urban vehicular networks or complex sensor networks). Therefore, models that can withstand rigors, be scalable, and can be set easily to situations with different traffic are necessary.

5.3. Theoretical Solutions Lacking Real-World Validation

One notable observation is that of the unraveling of theoretical frameworks of congestion control and the application used in the real world. Although a few papers (e.g., Paranjothi et al., 2020; Ariet et al., 2022) present potential congestion-mitigation schemes, the confirmatory studies of the large-scale and heterogeneous networks are still limited. It is also yet to be considered how such schemes behave in response to dynamic changes in traffic patterns, varied topology and granular models of mobility therefore the need to validate those congestion control techniques in realistic or close to reality simulation.

5.4. Narrow Focus on Application-Specific Scenarios

Another limitation is that, several empirical studies are limited in scope. Some of the works (Kashefi & Pourmina, 2019; Tshimangadzo et al., 2020) are focused on the applications in the specific domains like VOIP over LTE or routing in wireless sensor networks. Despite the achievements of these contributions in terms of knowledge of particular situations, the point of modularity limits their use in more general networking conditions. So the literature needs more open solutions that are not use case specific.

5.5. Complexity Hindering Practical Implementation

Complexity is also revealed as an imperative obstacle. Some solutions based on AI, such as collaborative Q-learning (Magbo et al., 2019) and reinforcement-based controllers (Balasingam et al., 2019) have been proposed to show theoretical potential but at a large computational cost. The complexity that arises prevents their use in real-time or in environments that have limited resources like in th rural networks or Internet of Things (IoT) systems. Thus there is urgency in the demand of lightweight compute efficient congestion control mechanism.

In concert, the above remarks create the need to study congestion control strategies in the realistic context of networks and highlight the importance of conducting more generalisable methodologies that can deal with the diversity of traffic patterns. As examined in the studies, in one case, Majeed et al. (2022) used a Distributed Congestion Control Protocol (DCCP) to minimize congestion, maximize the performance of its buffers by using these proactive tactics, and determine the best mechanism of reducing congestion in a Markov decision algorithm (Mahawish et al., 2022). Both portrayed promising results but the former can only work if the congestion stands out to be persistent and the latter is not able to resolve the congestion on transient spikes, resource contention, or network bottlenecks as well as simultaneous code generation with high request rates. The study thus proposes to address this research gap on future research works by introducing a more flexible, intelligent and resilient congestion control scheme that may be used in the dynamic environment inherent in cloud service environments.

6. CONCLUSION

A conceptual literature review was done to analyze the current congestion control mechanisms in the wireless as well as cloud computing environment and this was especially done with respect to its aptness in response to dynamic and complex congestion events. Scholarly works also had been critically reviewed exposing trends as well as the strengths of those techniques together with the research gaps still remaining to fill in the recently developed methods that use artificial intelligence, decision-theoretic models and network protocols like DCCP. Although the performance in wireless networks using these methods has been significantly enhanced, the research shows that there seems to be a significant lack of emphasis on congestion specifics in a cloud namely; transient bursts of network traffic, contention of resources across shared physical infrastructures, network bottlenecks due to simultaneous data generation, and a high request based on concurrent service demands.

The existing studies, such as those conducted by Majeed et al. (2022) and Mahawish et al. (2022), have demonstrated both proactive and algorithm-based methods of reducing congestion in the cloud environments; they are, however, practically limited in highly dynamic cloud service environments. The given study, hence, outlines an outstanding gap in the research: no in-depth, smart, and adaptive congestion regulation system is clearly specific to cloud environments. Critical examination of these previous papers shows that they are not robust, multi-constraint coping abilities, and are not real-time adaptable yet these are elements that are essential in maintaining a consistent performance in a current cloud system.

To address this shortcoming, the present study puts forward the notion of next-generation congestion control system that integrates intelligent decision-making, dynamic adjustability (i.e. real-time) and cross-layer optimization as a concept which may be studied in the future. The targeted structure is aimed at providing high quality of service (QoS), better resource usage, and reduced latency and hence high availability and reliability requirements in the cloud-based services. In the end, the study can help the development of the discipline since it brings intellectual rigor to the dynamic reality and begins to build a platform that can more effectively manage networks and overcome various failures in future systems by synchronizing traditional constructs of congestion control with innovation in the cloud computing domain.

REFERENCES

Adesh, N., & Renuka, A. (2019). Avoiding queue overflow and reducing queuing delay at eNodeB in LTE networks using congestion feedback mechanism. *Computer Communications, 146*, 131–143. https://doi.org/10.1016/j.comcom.2019.07.015

Agarwal, S., Krishnamurthy, A., & Agarwal, R. (2023). Congestion control. In *Proceedings of the ACM SIGCOMM 2023 Conference* (pp. 275–287). New York, USA.

Albanna, A., &Yousefi'Zadeh, H. (2020). Congestion minimization of LTE networks: A deep learning approach. *IEEE/ACM Transactions on Networking, 28*, 347–359. https://doi.org/10.1109/TNET.2019.2960266

Alizadeh, M., Yang, S., Sharif, M., Katti, S., McKeown, N., Prabhakar, B., & Shenker, S. (n.d.). pFabric: High-speed datacenter transport.

Arieth, M., Anuradha, K., Harika, B., Ali, D., & Chowdhury, S. (2022). Congestion management of CGSTEB routing protocol using K-means algorithm in wireless sensor networks.

Arslan, S., Li, Y., Kumar, G., &Dukkipati, N. B. (2023). Sub-RTT congestion control for ultra-low latency. In *Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* (pp. 219–236). Boston, USA.

Balasingam, A., Bansal, M., Misra, R., Nagaraj, K., Tandra, R., Katti, S., & Schulman, A. (2019). Detecting if LTE is the bottleneck with BurstTracker. *Academic Medicine*. https://doi.org/10.1145/3300061.3300140

Barr, J. (2019). The emerging cloud service architecture. Retrieved from https://aws.amazon.com/blogs/aws/the-forthcoming/

Bolanowski, M., Gerka, A., Paszkiewicz, A., Ganzha, M., &Paprzycki, M. (2023). Application of genetic algorithm to load balancing in networks with a homogeneous traffic flow. In J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Proceedings of the International Conference on Computational Science* (pp. 314–321). Springer Nature. https://doi.org/10.1007/978-3-031-36021-3_32

Chen, J., Jing, Y., & Xie, H. (2023). Multiple bottleneck topology TCP/AWM network event-triggered congestion control with new prescribed performance. *International Journal of Control, Automation and Systems, 21*, 2487–2503. https://doi.org/10.1007/s12555-022-0522-9

Farber, D. (2019). The new geek chic: Data centers. *CNET News*. Retrieved from https://www.cnet.com/news/the-new-geek-chic-data-centers/

Figuera, C., Morgado, E., Municio, E., & Simó-Reigadas, J. (2019). A low complexity mechanism for congestion notification in rural IPSec-enabled heterogeneous backhaul networks. *International Journal of Communication Systems, 32*(1), 1–12. https://doi.org/10.1002/dac.4082

Gaili, A., & Yigit, G. (2021). Intelligent RAN congestion management could help operators to get more from their 4G networks. *Analysis Mason ENEA Openwave*. https://owmobility.com/trafficmanagement/session-congestion-manager-scm/

Gens, F. (2019). Defining 'cloud services' and 'cloud computing'. IDC. Retrieved from https://blogs.idc.com/ie/?p=190

Gibson, D., Hariharan, H., Lance, E., McLaren, M., Montazeri, B., Singh, A., Wang, S., Wassel, H. M., Wu, Z., & Yoo, S. A. (2022). A unified, low-latency fabric for datacenter networks. In *Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)* (pp. 1249–1266). Renton, WA, USA.

Gibson, D., Hariharan, H., Lance, E., McLaren, M., Montazeri, B., Singh, A., Wang, S., Wassel, H. M., Wu, Z., & Yoo, S. A. (2022). A unified, low-latency fabric for datacenter networks. In *Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)* (pp. 1249–1266).

Guardian. (2019). Google angles for business users with platform as a service. Retrieved from https://www.theguardian.com/technology/2019/apr/17/google-software

Guo, A., & Yuan, C. (2021). Network intelligent control and traffic optimization based on SDN and artificial intelligence. *Electronics, 10*, 601–700.

Haile, H., Grinnemo, K. J., Ferlin, S., Hurtig, P., &Brunstrom, A. (2021). End-to-end congestion control approaches for high throughput and low delay in 4G/5G cellular networks. *Computer Networks, 186*, 107692. https://doi.org/10.1016/j.comnet.2020.107692

Hamzah, M., &Athab, O. (2022). A review of TCP congestion control using artificial intelligence in 4G and 5G networks. *American Academic Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 88*, 172–186.

Hauser, F., Häberle, M., Merling, D., Lindner, S., Gurevich, V., Zeiger, F., Frank, R., & Menth, M. (2023). A survey on data plane programming with P4: Fundamentals, advances, and applied research. *Journal of Network and Computer Applications, 212*, 103561. https://doi.org/10.1016/j.jnca.2022.103561

Iliev, T., Bikov, T., Mihaylov, G., Ivanova, E., Stoyanov, I., & Keseev, V. (2019). Algorithms for congestion control in LTE mobile networks. In R. Silhavy (Ed.), *CSOC 2018: Advanced Intelligent Systems and Computing (AISC), 763*, 474–484. Springer.

Imam-Fulani, Y. O., Faruk, N., Sowande, O. A., Abdulkarim, A., Alozie, E., Usman, A. D., Adewole, K. S., Oloyede, A. A., Chiroma, H., Garba, S., Imoize, A. L., Baba, B. A., Musa, A., Adediran, Y. A., & Taura, L. S. (2023). 5G frequency standardization, technologies, channel models, and network deployment: Advances, challenges, and future directions. *Sustainability, 15*(6), 5173. https://doi.org/10.3390/su15065173

Joshi, R., Song, C. H., Khooi, X. Z., Budhdev, N., Mishra, A., Chan, M. C., & Leong, B. (2023). Masking corruption packet losses in datacenter networks with link-local retransmission. In *Proceedings of the ACM SIGCOMM 2023 Conference* (pp. 288–304). New York, USA. https://doi.org/10.1145/3603269.3604853

Kanellopoulos, D. (2019). Congestion control for MANETs: An overview. *The Korean Institute of Communications and Information Sciences (KICS), 77–83*. https://doi.org/10.1016/j.adhoc.2019.05.001 (if DOI is applicable; else omit)

Kashefi, F., &Pourmina, M. (2019). Evaluate how to minimize the effects of LTE network congestion for VOIP. *Specialty Journal of Electronic and Computer Sciences, 5*, 20–27.

Khan, O. G. M., Youssef, A., Salama, M., & El-Saadany, E. (2022). Robust multi-objective congestion management in distribution network. *IEEE Transactions on Power Systems, 10*, 1–11. https://doi.org/10.1109/TPWRS.2022.3200838

Khan, O. G. M., Youssef, A., Salama, M., & El-Saadany, E. (2022). Robust multi-objective congestion management in distribution network. *IEEE Transactions on Power Systems, 10*, 1–11. https://doi.org/10.1109/TPWRS.2022.3200838

Kumar, G., Dukkipati, N., Jang, K., Wassel, H. M., Wu, X., Montazeri, B., Wang, Y., Springborn, K., Alfeld, C., & Ryan, M. S. (2020). Delay is simple and effective for congestion control in the datacenter. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication* (pp. 514–528). https://doi.org/10.1145/nnnnnnn

Kundel, R., Krishna, N. B., Gärtner, C., Meuser, T., & Rizk, A. (2021). Poster: Reverse-path congestion notification: Accelerating the congestion control feedback loop. In *Proceedings of the 2021 IEEE 29th International Conference on Network Protocols (ICNP)* (pp. 1–2). Dallas, USA. IEEE. https://doi.org/10.1109/ICNP52444.2021.9651961

Li, Q., & Flexi, T. C. P. (2023). S: A new approach to incipient congestion detection and control. *IEEE/ACM Transactions on Networking*, 1–16.

Lim, H., Kim, J., Cho, I., Jang, K., Bai, W., & Han, D. F. P. (2023). A case for flexible credit-based transport for datacenter networks. In *Proceedings of the Eighteenth European Conference on Computer Systems* (pp. 606–622). https://doi.org/10.1145/nnnnnnn

Magbo, P., Chukwudi, P., & Abubakar, A. (2019). Congestion control mechanisms for 4G random access channel (RACH): A survey. *International Journal of Engineering and Applied Sciences (IJEAS), 6*, 6–9.

Mahawish, A. A., & Hassan, H. J. (2022). Improving RED algorithm congestion control by using the Markov decision process. *Scientific Reports, 12*, 13363–19674. https://doi.org/10.1038/s41598-022-17528-x

Mouna Patel, S., & Bhoi, U. (2020). Priority based job scheduling techniques in cloud computing: A systematic review. *International Journal of Scientific & Technology Research, 2*(11). Retrieved from http://www.ijstr.org

Namkung, H., Liu, Z., Kim, D., Sekar, V., &Steenkiste, P. (2023). Sketchovsky: Enabling ensembles of sketches on programmable switches. In *Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* (pp. 1273–1292). Boston, USA.

Obama, B. (2019). Remarks by the president at the National Academy of Sciences annual meeting. Retrieved from https://obamawhitehouse.archives.gov

Paranjothi, A., Khan, M., &Zeadally, S. (2020). A survey on congestion detection and control in connected vehicles. *Ad Hoc Networks Journal, 5*, 38–49.

Poutievski, L., Mashayekhi, O., Ong, J., Singh, A., Tariq, M., Wang, R., Zhang, J., Beauregard, V., Conner, P., & Gribble, S. (2022). Jupiter evolving: Transforming Google's datacenter network via optical circuit switches

and software-defined networking. In *Proceedings of the ACM SIGCOMM 2022 Conference* (pp. 66–85). Amsterdam, The Netherlands.

Prateesh, G., Preey, S., Kevin, Z., Georgios, N., Mohammad, A., & Thomas, A. (2022). Backpressure flow control. In *Proceedings of the Symposium on Network System Design and Implementation (NSDI)* (pp. 779–805).

Prateesh, G., Preey, S., Kevin, Z., Georgios, N., Mohammad, A., & Thomas, A. (2022). Backpressure flow control. In *Proceedings of the Symposium on Network System Design and Implementation (NSDI)* (pp. 779–805).

Scazzariello, M., Caiazzi, T., Ghasemirahni, H., Barbette, T., Kostić, D., & Chiesa, M. (2023). High-speed packet processing approach for Tbps programmable switches. In *Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* (pp. 1237–1255). Boston, USA.

Shen, F., & Wu, Q. (2022). Robust dynamic tariff method for day-ahead congestion management of distribution networks. *International Journal of Electrical Power and Energy Systems, 10*, 61–72. *Elsevier Ltd.*

Tshilongamulenzhe, T. M., Mathonsi, T. E., Mphahlele, M. I., & Du Plessis, D. (2020). Traffic-based congestion management algorithm for wireless sensor networks. *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 202–207. https://doi.org/10.1109/CSCI51800.2020.00041

Umoh, U., Eyoh, I., Isong, E., & Inyang, A. (2020). Uncertainty and congestion elimination in 4G network call admission control using interval type-2 intuitionistic fuzzy logic. *Global Journal of Computer Science and Technology, 4*, 34–51.

Wang, W., Moshref, M., Li, Y., Kumar, G., Ng, T. E., Cardwell, N., &Dukkipati, N. P. (2023b). Efficient, robust, and practical datacenter congestion control via deployable INT. In *Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* (pp. 255–274). Boston, USA.

Wang, W., Moshref, M., Li, Y., Kumar, G., Ng, T. S. E., Cardwell, N., &Dukkipati, N. (2023a). Poseidon: Efficient, robust, and practical datacenter CC via deployable INT. In *Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* (pp. 255–274). USENIX Association.

Xiao, L. Y. S., Jun, B. I., Yu, Z., Cheng, Z., Ping, W. J., Zheng, L. Z., & Ran, Z. Y. (2019). Research and applications of programmable data plane based on P4. *Chinese Journal of Computers, 42*, 2539–2560.

Zaher, M., Alawadi, A. H., & Molnár, S. (2021). Sieve: A flow scheduling framework in SDN-based data center networks. *Computer Communications, 171*, 99–111. https://doi.org/10.1016/j.comcom.2021.02.013

Zhang, Y., Liu, Y., Meng, Q., & Ren, F. (2021). Congestion detection in lossless networks. In *Proceedings of the 2021 ACM SIGCOMM Conference* (pp. 370–383). https://doi.org/10.1145/3452296.3472899