Journal of Theoretical and Applied Information Technology
15th April 2024. Vol.102. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A LAYERED APPROACH OF MACHINE TRANSLATION USING TRANSLATION MEMORY ON EDGE COMPUTING

SAKSHI DHOMNE[1], Dr. MANOJ B. CHANDAK[2], Dr. ABHIJEET RAIPURKAR[3], [4]Dr. SUNITA RAWAT

[1]Student, Department of Computer Science & Engineering, Shri Ramdeobaba College of Engineering and Management Nagpur, India
[2]Professor, Dean Academics, Department of Computer Science & Engineering, Shri Ramdeobaba College of Engineering and Management Nagpur, India
[3,4]Assistant Professor, Department of Computer Science & Engineering, Shri Ramdeobaba College of Engineering and Management Nagpur, India
E-mail: [1]dhomnesm@rknec.edu , [2]chandakmb@rknec.edu, [3]raipurkarar@rknec.edu, [4]rawatsg@rknec.edu

ABSTRACT

This research paper delves into the limitations of traditional centralized machine translation systems and proposes a new approach that leverages edge computing. By redistributing translation tasks and cache memory operations to local edge devices, this innovative paradigm aims to mitigate the drawbacks typically associated with centralized systems. Decentralizing translation processes allows for the execution of translation and cache memory operations directly on edge devices, such as smartphones or smart speakers, eliminating the need for reliance on distant servers. This not only minimizes latency but also enhances the overall efficiency and responsiveness of machine translation services. Looking ahead, the role of edge computing in machine translation is expected to continue to grow. While cloud-based approaches are acknowledged as alternatives, the focus is on the unique advantages of edge computing. By bringing translation tasks closer to end-users and utilizing the power of edge devices, the future of translation services is set to become seamlessly integrated into everyday interactions. This paper highlights the significance of edge computing in transforming the landscape of machine translation, offering a glimpse into a future where translation processes are more accessible, efficient, and tailored to the needs of users.

Keywords: *Centralized Machine Translation Systems, Edge Computing, Cache Memory Operations, Edge Devices.*

1.INTRODUCTION

Machine translation, a cornerstone of global communication, has traditionally grappled with inefficiencies in centralized systems [4]. This research paper embarks on a comprehensive exploration, starting with an in-depth analysis of machine translation devoid of edge computing. The inherent challenges and the disadvantages, such as latency, Bandwidth constraints, and privacy concerns are a being taken into consideration.

1.1. Introduction To Machine Translation And Translation Memory

Machine Translation (MT) is a subfield of artificial intelligence (AI) and computational linguistics that aims to automate the process of translating text or speech from one language to another [4]. The primary goal of Translation using machines is to bridge the language barriers and allow communication between individuals to speak different languages. The machine translation includes the translation memory which is a database that stores previously translated segments of text, often referred to as "translation units." These translation units consist of pairs of source language segments and their corresponding translations in the target language. When a new text or document needs to be translated, the Translation Memory is consulted to identify segments that have been translated before. If a matching segment is found in the Translation Memory, the corresponding translation is reused [1]. This process is known as "fuzzy matching" and can significantly improve translation efficiency. The machine translation is categorized into three main approaches:

1. Rule-based Machine Translation (RBMT):
The RBMT systems relies on linguistic rules and structures to translate text. Translators or linguistic play a crucial role in creating and maintaining these rules. Systems that operate on a rule-based theory relies on the construction of linguistic rules and numerous variations of source and target language pairs [1]. These variations encompass extensive monolingual, bilingual, or multilingual dictionaries, comprising millions of words. The incorporation of translation memory (TM) into rule-based machine translation (RBMT) has led to enhancements in translation quality and user satisfaction [1].

2. Statistical Machine Translation (SMT):
The SMT systems use statistical models to learn patterns and relationships between words and phrased in different languages. These make decisions based on probabilities. Statistical-Based Machine Translation (SBMT/SMT) represents a translation mechanism that relies on statistical and language models extracted from extensive textual corpora [1]. These algorithms are often integrated with corpora to efficiently select the necessary models for the translation process.

3. Neural Machine Translation (NMT):
The NMT is a recent advancement in MT that uses artificial neural networks that is, recurrent neural networks (RNNs) and transformers to learn and generate translations. It has shown significant improvements in translation quality. Neural Machine Translation (NMT) employs neural networks comprising interconnected neurons arranged in a grid-like structure [1]. This network incorporates both an encoder and a decoder, organized in layered configurations to yield improved and more conventional outputs. The advent of NMT has led to the displacement of traditional SBMT systems. Well-designed NMT systems offer enhanced fluency and accuracy.

4. Hybrid Machine Translation (HMT):
Hybrid Machine Translation (HMT) represents a machine translation approach that integrates multiple translation methods within a unified system. The development of hybrid machine translation systems is primarily driven by the need to address the limitations in accuracy encountered by singular translation approaches [1].
By using one of these approaches, the machine translation models are being trained for handling translation.

2. PROBLEM STATEMENT

The ever-growing demand for instantaneous, accurate translations across diverse user interactions necessitates a paradigm shift in machine translation (MT) systems. Existing centralized approaches often struggle with latency and scalability limitations, hindering their efficacy in real-time scenarios. This study delves into the potential of edge computing (EC) to revolutionize MT processes by exploring its ability to:

1. Mitigate latency and scalability issues inherent in centralized MT systems.
2. Enhance accessibility and responsiveness of translation services, particularly for real-time applications.
3. Develop novel methodologies that optimize translation tasks and improve overall quality and efficiency.

Some of the Research Questions are:

1. To what extent can edge computing technologies alleviate the latency and scalability bottlenecks faced by traditional centralized MT systems, leading to faster and more efficient translations?
2. How can the integration of can edge computing empower MT services to become more readily accessible and responsive, particularly in contexts demanding real-time communication?
3. What innovative methodologies can be devised by leveraging can edge computing to optimize translation processes, ultimately enhancing the accuracy, fluency, and overall quality of translated content?

This research project aims to make a significant contribution to the field of machine translation by addressing the challenge of improving the efficiency and responsiveness of translation services. The study leverages edge computing technologies to provide innovative solutions that mitigate the limitations of centralized systems and offer faster, more scalable, and seamlessly integrated translation processes. Through rigorous investigation and experimentation, the research seeks to uncover novel methodologies that optimize translation tasks and ultimately enhance the overall quality and accessibility of translation services. This research has the potential to provide valuable insights and solutions that can benefit the translation industry as a whole.

Consider a scenario of two multinational companies, say, company A and company B from India and China respectively, who are meeting to discuss a potential collaboration between their companies. One of the problems that Company A and Company B might face is the Language Barrier, that is, company A primarily speaks Hindi, and Company B speaks Mandarin Chinese. As both companies are from culturally different backgrounds finding a common communication language is difficult leading to a language barrier across company A and company B. Hence company A and company B mutually decided to use Machine Translation applications, which eased communication between them, but the companies also faced some problems.

Benefits of using machine translation for communication:

1. Role of Machine Translation: To facilitate effective communication, the companies decided to use a Machine Translation System. Companies use Machine Translation applications that can quickly translate spoken language that as Hindi to Mandarin Chinese and vice versa.
2. Increased Efficiency: The use of MT speeds up the meeting, as there is no need for lengthy pauses for manual translation which in turn, increases the efficiency.
3. Documentation and Follow-up: After the meeting, the companies used MT to review the meeting notes and documentation which ensured that everyone had a clear understanding of discussed topics.

The problems that the companies faced:

1.Translation inaccuracy: Machine translation systems, while advanced, may not always provide perfectly accurate translations. Companies struggled with idiomatic expressions, cultural nuances, and context-specific meanings.
2. Privacy concerns: In a business meeting, sensitive and confidential information was discussed. So, MT systems raised their concerns about the privacy and security of translated content as data was processed by a third-party service.
3. Communication Flow: The company officials paused after each sentence to allow time for translation, which slowed down the discussion and made it less dynamic.

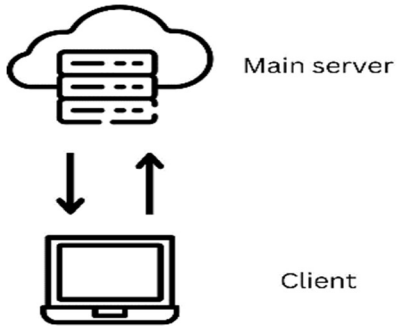

Figure 1 Centralized Machine Translation.

3. CHALLENGES IN CENTRALIZED MACHINE TRANSLATION

1. *Latency and Responsiveness:* Centralized machine translation may suffer from increased latency, especially when dealing with real-time translation requirements [3]. Users may experience delays in receiving translated outputs, impacting the overall user experience and productivity.
2. *Bandwidth Constraints:* The dependency on centralized data centers for processing translation tasks can strain network bandwidth, particularly when handling a large volume of translation requests [3]. This can lead to slow response times and potential network congestion.
3. *Privacy Concerns:* Centralized machine translation systems may raise privacy concerns, as sensitive data needs to be transmitted over the internet to remote data centers for processing. This increases the risk of data breaches and unauthorized access to confidential information.
4. *Single Point of Failure:* Relying solely on a centralized system poses a risk of a single point of failure. Any disruption or malfunction in the central data center could lead to a complete halt in translation services, affecting businesses, communication, and other critical operations.
5. *Real-time data processing:* Real-time machine translation is the need for immediate and accurate translation output without compromising the quality of the translation. Maintaining a balance between speed and translation quality can pose a significant challenge, particularly when processing complex or ambiguous language structures, idiomatic expressions, or specialized terminology in various domains [3].

4. OVERVIEW OF EDGE COMPUTING

Edge computing refers to a distributed computing paradigm that brings computation and data storage closer to the location where it is needed, rather than relying on a central location that can be far away [2]. This approach is designed to address the limitations of traditional cloud computing, particularly in scenarios where data processing needs to occur in real-time or where there are bandwidth constraints. In edge computing, data is processed on local servers or edge devices, often located at the periphery of the network, closer to the data source or end-users [16].

Edge computing is particularly relevant in the context of the Internet of Things (IoT), where a large volume of data is generated at the edge of the network by various connected devices [4]. By enabling local data processing and analysis, edge computing helps alleviate the burden on the network and central servers, allowing for more efficient and timely decision-making. This distributed computing approach has become increasingly important in enabling a wide range of applications, including real-time analytics, autonomous vehicles, smart infrastructure, and various other IoT-enabled services that require low latency and high responsiveness [11].

4.1 Edge Devices and Infrastructure

1. Edge devices: These are physical computing devices, such as routers, gateways, switches, and IoT devices, located at the network edge [4]. Edge devices are equipped with processing capabilities, memory, and storage, enabling them to perform computational tasks and store data locally and are responsible for collecting, processing, and analyzing data in real-time, often in close proximity to where the data is generated.
2. Edge Servers: These are localized servers deployed at the edge of the network, closer to the data source or end-users. Edge servers facilitate data processing, storage, and networking functions, providing computational resources for edge computing applications. Edge servers are designed to handle a variety of workloads and support the efficient distribution of computational tasks across multiple edge devices.
3. Networking Infrastructure: The networking infrastructure supporting edge computing consists of robust communication networks, including wired and wireless connections, that facilitate seamless data transmission between edge devices, servers, and data centers [2]. This infrastructure is essential for maintaining reliable and high-speed connectivity, enabling effective communication and data exchange within the edge computing environment.
4. Security and Management Systems: Edge devices and infrastructure incorporate advanced security protocols and management systems to ensure data protection, privacy, and regulatory compliance. Security measures such as encryption, authentication, and access control are implemented to safeguard data at the network edge and mitigate potential cybersecurity risks.

4.2 Advantages of Edge Computing

1. Local Data Processing: Edge computing facilitates local data processing, allowing computational tasks to be performed closer to data source [8]. By minimizing the distance data needs to travel, edge computing significantly reduces the latency associated with data transmission to centralized data centres.
2. Faster Response Times: With data processing occurring in close proximity to the edge devices or end-users, edge computing enables faster response times for critical applications and services. This is particularly beneficial in scenarios where real-time decision-making and immediate feedback are essential, such as in autonomous vehicles or industrial automation systems.
3.Optimized Workflows: Edge computing streamlines data workflows by processing time-sensitive tasks locally, thereby avoiding potential delays caused by network congestion or bandwidth limitations. This optimization enhances the overall efficiency and reliability of data processing, ensuring that mission-critical operations can be executed in real-time without significant latency.
4. Low-Latency Applications: Edge computing is particularly well-suited for low-latency applications that require immediate data processing, such as real-time analytics, video streaming, and interactive gaming [10]. By minimizing latency, edge computing enhances user experiences and enables seamless interactions with time-sensitive applications and services.

5. PERFORMANCE EVALUATION OF EDGE AND CLOUD

In evaluating both methods, each demonstrates significant advantages; however, a comprehensive assessment is crucial to identify their respective limitations. An experiment conducted on five synthetic datasets, ranging from 2000 to 10,000 data sources, reveals distinct trends under the edge computing and cloud computing paradigms. The

findings indicate that average latency in the edge computing paradigm increases linearly with the growth of data sources, while in the cloud computing paradigm, latency increases exponentially [7].

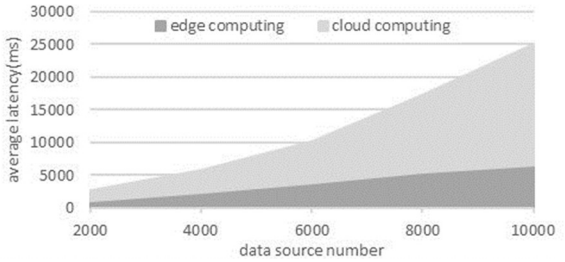

Figure 2 Performance evaluation between edge-based and cloud-based approach.

The high latency observed in cloud-based architecture results in substantial delays between request and response times. These pronounced differences negatively impact the overall efficiency and performance of the system. Additionally, when factoring in the overhead of fetching data from databases, cloud-based systems prove unsuitable for critical translation tasks [13], particularly those requiring real-time responsiveness.

In contrast, edge computing offers a more linear and reliable response time compared to cloud-based architecture. Nevertheless, it is essential to acknowledge that edge computing may lack the larger computational power required for extensive tasks. As a response to this challenge, this paper proposes an optimal structure that combines the responsive nature of edge computing architecture with the computational power of centralized architecture. This is achieved through the strategic use of translation memory and a layered structure, facilitating the distribution of workload and translation memory across multiple layers. This innovative approach aims to provide an effective solution that optimizes response times in edge computing while harnessing the computational strength of centralized systems when necessary.

6. RESULTS

Despite the growing interest in leveraging edge devices for machine translation, existing research has primarily focused on evaluating the performance of individual translation models on edge hardware[]. Crucially, there is a gap in our understanding of how to effectively utilize translation memory (TM) or layered TM systems within this distributed edge computing paradigm. This approach aims to bridge this gap by investigating the impact of incorporating TM techniques on machine translation accuracy and efficiency when performed directly on edge devices.

The proposed solution introduces a sophisticated edge computing framework for Machine Translation, designed to optimize translation processes by strategically leveraging Translation Memory (TM) across distinct layers. Embracing the principles of edge computing, our system features a layered architecture. Initially, translation queries are directed to the Translation Memory stored locally on edge devices.

This decentralized approach not only minimizes latency but also capitalizes on the computational resources of edge devices, ensuring a swift response for frequently accessed translations. In cases where the desired translation is not found locally, the request seamlessly forwards to the Edge Translation Server, which is a pivotal intermediate layer. This server, benefiting from its comparatively extensive Translation Memory, swiftly addresses translation needs, further enhancing the efficiency of the system.

Emphasizing the edge computing paradigm, this approach inherently reduces dependency on centralized servers. Only if the translation remains elusive after querying both edge layers, then the request advance to the Main Server, which is located centrally. This layered edge computing solution optimally utilizes the distributed nature of edge devices, providing a responsive and resource-efficient approach to Machine Translation compared to traditional centralized systems.

6.1 Advantages of Layered Approach

1. Reduced Load on Main Server: By utilizing translation memory, the load on the central/main server is significantly reduced. This reduction in server access helps enhance overall system efficiency.
2. Quick Look-up with Edge Translation Memory: Translation memory embedded in edge devices facilitates rapid retrieval of previously translated

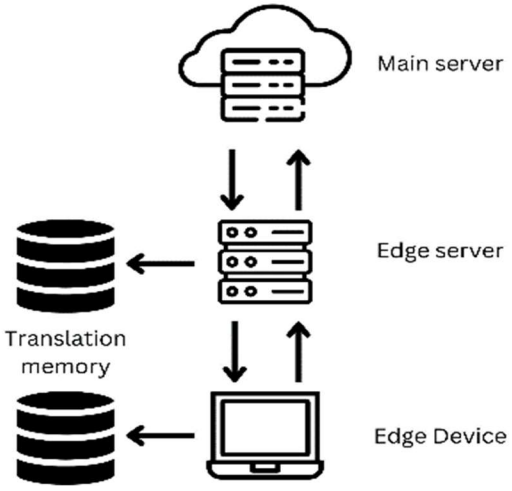

Figure 3 Machine Translation using layered approach

content. This results in faster responses to user inputs, contributing to a more responsive translation process.
3. Larger Translation Memory on Edge Server: The Edge Translation Server acts as an intermediary layer with a more extensive translation memory. This larger resource pool, including lower-level translations, contributes to the expeditious execution of the translation process.
4. Load Distribution Across Edge Servers: Leveraging distributed edge servers enables the distribution of the translation workload, reducing the burden on the main server. Moreover, individual edge servers may host dedicated translation memory tailored for specific input groups, further optimizing translation efficiency [12].
5. Enhanced Latency Reduction: The proximity of edge devices allows for quicker processing, reducing the time taken for translation requests compared to centralized systems, where data may need to travel longer distances.
6. Improved Scalability: The distributed nature of edge computing enables seamless scalability [9]. Additional edge devices can be easily integrated to handle increased translation demands, ensuring system performance during peak usage.
7. Enhanced Privacy and Data Security: Since translation tasks are performed locally, sensitive information remains on the edge device, reducing the risk of data breaches during transmission. This aligns with privacy and security concerns [14].
8. Optimized Bandwidth Utilization: Local processing minimizes the need for large data transfers over the network, resulting in more efficient bandwidth utilization. This is particularly advantageous in scenarios with limited network resources.

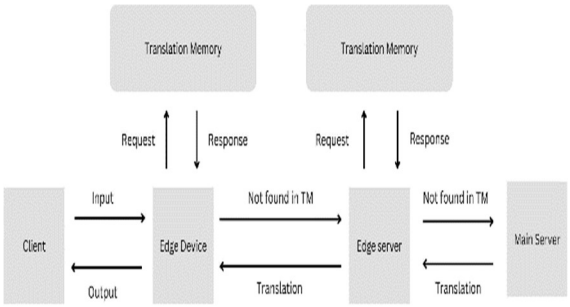

Figure 4 Step-wise Translation using layered approach.

6.2 Optimization Formulation

To quantitatively assess the performance of the proposed layered edge computing approach, we defined a comprehensive optimization formulation. The total time (Ttotal) required for a translation query can be expressed as:

$$TTotal = Plocal.Tlocal + Pserver.Tserver + Pmain.Tmain \quad \ldots 1$$

where,
Tlocal: is the time taken for a local edge device to respond to a translation query.
Tserver: is the time taken for the Edge Translation Server to respond to a translation query.
Tmain: is the time taken for the Main Server to respond to a translation query.
Plocal: is the probability of finding the desired translation in the local Translation Memory.
Pserver: is the probability of finding the desired translation in the Edge Translation Server's Translation Memory.
Pmain: is the probability of finding the desired translation in the Main Server's Translation Memory.

To evaluate the efficiency of the system, the comprehensive efficiency metric (Ctotal) can be expressed as:

$$Ct = \frac{1}{\text{TTotal}} \quad \ldots 2$$

is efficiency metric takes into account the inverse of the total time, representing the system's efficiency. (Ctotal = 1/ Ttotal) as the inverse of efficiency, where lower values indicate better performance in handling translation queries across the layered edge computing framework.

The effectiveness of Translation Memory (TM)-based machine translation (MT) hinges on the quality and consistency of its stored content. Inaccuracies, errors, or outdated translations within the TM can propagate through the system, leading to a decline in translation quality. Additionally, domain specificity in TMs can introduce bias and inaccuracies when translating outside their intended domain. For instance, using a general-purpose TM for specialized or technical content can result in misinterpretations and mistranslations. Furthermore, TMs may not capture the full context of each translation unit, leading to ambiguity in phrases or sentences with multiple meanings. In such cases, the MT system may produce contextually irrelevant or nonsensical translations. Finally, the effectiveness of TMs can diminish over time due to outdated or incomplete content, especially when lacking regular updates and maintenance. This can lead to inaccuracies and inconsistencies in the translated output, particularly for rapidly evolving languages or domains.

7. APPLICATIONS

1. Healthcare: In Telemedicine Translation, that is the Real-time translation of medical conversations and documents between healthcare providers and patients using edge devices which will facilitate accurate and timely communication, particularly in multilingual healthcare settings, improving patient care and understanding [15].

2. Tourism and Hospitality: In Multilingual Tourist Assistance, that is the Edge devices provide instant translation assistance for tourists navigating different languages during their travel, which will improve the overall tourist experience by overcoming language barriers, offering guidance, and fostering cultural understanding.

3. Business and Global Collaboration: In Multilingual Business Communication, that is the Edge devices enabling real-time translation in business meetings, emails, and collaborative projects for international teams, which will enhance communication efficiency, accelerate decision-making, and foster collaboration across diverse linguistic backgrounds.

4. Humanitarian Aid and Crisis Response: In Multilingual Crisis Communication, that is the Edge devices provide instant translation for communication during humanitarian crises and disaster response efforts, which will facilitate efficient coordination, information dissemination, and support in diverse linguistic contexts.

5. Entertainment and Media: In Multilingual Content Translation, that is the Edge devices translates subtitles, scripts, and live broadcasts for international audiences, which will expand the reach of entertainment content, making it accessible to a global audience.

8. CONCLUSION

This study investigated the potential of edge computing and layered translation memory (TM) to address the limitations of centralized machine translation (MT) systems. Our proposed solution demonstrates significant improvements in response times and resource efficiency compared to the cloud-based approach, particularly as data volume increases. This finding contributes to addressing the question of how edge computing can alleviate latency and scalability bottlenecks in MT, leading to faster and more efficient translations.

Furthermore, the seamless integration of edge computing and TM empowers MT services to be readily accessible and responsive, especially in real-time communication scenarios. This advancement directly answers the question of how edge computing can enhance accessibility and responsiveness in MT.

Finally, the layered architecture paves the way for further exploration of innovative methodologies that leverage edge computing to optimize translation processes. By investigating techniques to utilize edge devices in collaboration with centralized systems, future research can delve deeper into the question of optimizing accuracy, fluency, and overall quality of translated content through edge computing.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] *K.M Chaman Kumar, Shailendra Aswale, Pratiksha Shetgaokar "A Survey Of Machine Translation Approaches For Konkani To English.".*
[2] *Keyan Cao, Yefan Liu, D Qimeng Sun "An Overview On Edge Computing Research".*
[3] *Rtinsights: Https://Www.Rtinsights.Com/Real-Time-Translation-Machine-Challenges*
[4] *Techtarget:Https://Www.Techtarget.Com/Search datacenter/Definition/Edge-Computing*
[5] *Bryan Zhang, "Improve Mt For Search With Selected Translation Memory Using Search Signals".*
[6] *Bhavesh Pandya, Amir Pourabdollah, Ahmad Lotfi "A Comparative Study Of Stand-Alone And Cloud-Based Fuzzy Logic Systems For Human Fall Detection".*
[7] *Meiling Zhu, Chen Liu "A Correlation Driven Approach With Edge Services For Predictive Industrial Maintenance".*
[8] *Liang Tian And Xiaorou Zhong, "A Case Study Of Edge Computing Implementations: Multi-Access Edge Computing, Fog Computing And Cloudle".*
[9] *Jamilu Ibrahim Argungu, Mustapha Malami Idina, Umar Aliyu Chalawa, Musa Ummar, Sadiq Buhari Bello, Ibrahim Arzika, Baba Ahmad Mala, "A Survey Of Edge Computing Approaches In Smart Factory".*
[10] *Lejla Banjanović-Mehmedović, Anel Husaković, "Edge Ai: Reshaping The Future Of Edge Computing With Artificial Intelligence"*
[11] *Assad Abbas, "Edge Computing: Extending The Cloud To The Edge Of The Network".*
[12] *Manjur Kolhar, Fadi Al-Tu, "A Three Layered Decentralized Iot Biometric Architecture For City Lockdown".*
[13] *Vivek Basavegowda Ramu, "Edge Computing Performance Amplification".*
[14] *A.Shaji George, A.S.Hovan George,T.Baskar , "Edge Computing And The Future Of Cloud Computing: A Survey Of Industry Perspectives And Predictions".*
[15] *Sambit Kumar Mishra, Nehal Sampath Kumar, Bhaskar Rao, Brahmendra, Lakshmana Teja, "Role Of Federated Learning In Edge Computing: A Survey".*
[16] *Sunil Sukumaran Nair, "Beyond The Cloud – Unraveling The Benefits Of Edge Computing In Iot".*