Review of AI-driven Cloud Optimization

Anurag J
Student
School of Computer Science And Information Technology
Jain (Deemed-to be University)
Banglore, India
23mcar0040@jainuniversity.ac.in

J,Bhuvana
Assistant professor
School of Computer Science And Information Technology
Jain (Deemed-to be University)
Banglore, India
j.bhuvana@jainuniversity.ac.in

***Abstract*— Cloud automation is the key to realization a fully-optimized performance of modern cloud platforms while cloud resources utilization. Resource allocation efficiency is valuable. We are however faced with increasing pressure for computational resources. The Long Short-Term Memory (LSTM) algorithms have found a great use case in the dynamic resource allocation problem when the problem is solved by the proactive provisioning of resources based on historical usage patterns taking advantage of recurrent neural networks. Furthermore, the concern over quality-of-service delivery (QoS) and energy efficiency is now almost as challenging for cloud providers when implementing the utilization of cloud resources, especially in a dynamic setting. Deep Reinforcement Learning (DRL) allows to pursue that end by developing agents, which might guide the work of optimizing resource allocation and reducing energy expenses at the same time. It improves the result and adaptability of the applications running on clouds. Artificial intelligence, in its diverse form, for the instance machine learning and optimization algorithms, brings in a great influence in the areas of cloud operations, resource management, and security. Furthermore, the 3rd generation of FPGAs (Reconfigurable Digital Computing-In-Memory or ReDCIM processors) and the bitwise parallelism via-memory core multipliers also improve the efficiency of computing in cloud environments. Adopting these inputting methodologies is a consequence of cloud systems achieving top performance, high scalability, and low costing.**

Keywords - Cloud Automation, Long Short-Term Memory (LSTM) algorithms, Deep Reinforcement Learning (DRL), Reconfigurable Digital Computing-In-Memory (ReDCIM) processors, Resource sAllocation, Energy Efficiency.

I. INTRODUCTION

Resources allocation (in other words, efficient computing systems) takes a key place, which means that it is vital to attain the highest performance while keeping costs on the low level. In a dynamic atmosphere of modern technologies, where intensification of computational resources' need is progressing, the adequate management of resources is a key factor. Recently, LSTM (Long Short -Term Memory) algorithms have appeared on the scene as a powerful instrument to counter the limitations of constant resource planning instead of dynamic resource allocation. LSTM (Long Short-Term Memory), which is a kind of Recurrent Neural Network (RNN) good at neural learning, is able to do proactive resource provisioning by analyzing the previous resource utilization. Through projecting future necessities based on preceding usage, LSTM systems model computation units to allocate resources properly, with a goal of avoiding future shortages and workload bottlenecks. In doing so, it, therefore, provides the institution with the following benefits: integration, robustness, and cost-cutting among others. The LSTM-based resource allocation is not only offering the general capacity settings implying the possibility of the fine tuning of the exact settings based on the changes that happen in the workload. Drawing continuous lessons from the past, these algorithms perform the necessary optimization of resource allocation strategies and they do this over and over again so to ensure the ability to adapt as the demands may change. The final consequence of LSTM term integration into resource allocation is bettering the performance and efficiency of the system through equipping the computing with capability of seamless processing as the demands are not static and the technology keeps on changing.[1]

Allocation of cloud resources constitutes multidimensional problems especially in terms of assuring the quality of the service (QoS) and energy efficiency and adapting to a dynamic environment of the cloud. Preserving QoS such as response time, throughput, and availability which are crucial for customer satisfaction and the reliability of the service necessitate meeting workload and resource requirements. On the other side, achieving Quality of Service objectives becomes very difficult as the load pattern get unsettled in dynamic cloud environment. However, energy efficiency has become an essential factor for cloud data centers due to negative environmental influence and efficiency operational expenses. The performance requirements of resource allocation algorithms are adversarial with respect to their energy consumption, which poses a great problem. With regards to this, DRL (Deep Reinforcement Learning) becomes the most go for solution. DRL resorts to neural networks and reinforcement learning to enable agents to discover optimal resource allocation procedures by actively testing different scenarios. Through training agents to reach the highest value for QoS metric rewards while minimizing energy consumption, DRL-based methods can resolve the dual objectives of performance improvement together with energy conservation in cloud resource allocation. Flexibility and adaptability of DRL allow it to respond quickly to changing resources, while dynamic cloud platforms are constantly dynamically changing their requirements. DRL based approach can help in continuous learning and adapting by the method of setting the resource allocation dynamically to the changing workload patterns and to maintain the standard efficiency in resource utilization and QoS. Overall, through the use of DRL-based solutions, we see a viable means of dealing with cloud resource problems which include ability to manage QoS, energy efficiency and adaptation to a dynamic cloud environment.[2]

Allocation of cloud resources constitutes multidimensional problems especially in terms of assuring the quality of the service (QoS) and energy efficiency and adapting to a dynamic environment of the cloud. Preserving QoS such as response time, throughput, and availability which are crucial for customer satisfaction and the reliability of the service necessitate meeting workload and resource requirements. On the other side, achieving Quality of Service objectives becomes very difficult as the load pattern get unsettled in dynamic cloud environment. However, energy efficiency has become an essential factor for cloud data centers due to negative environmental influence and efficiency operational expenses. The performance requirements of resource allocation algorithms are adversarial with respect to their energy consumption, which poses a great problem. With regards to this, DRL (Deep Reinforcement Learning) becomes the most go for solution. DRL resorts to neural networks and reinforcement learning to enable agents to discover optimal resource allocation procedures by actively testing different scenarios. Through training agents to reach the highest value for QoS metric rewards while minimizing energy consumption, DRL-based methods can resolve the dual objectives of performance improvement together with energy conservation in cloud resource allocation. Flexibility and adaptability of DRL allow it to respond quickly to changing resources, while dynamic cloud platforms are constantly dynamically changing their requirements. Based approach can help in continuous learning and adapting by the method of setting the resource allocation dynamically to the changing workload AI-based applications have become core elements for cloud management and identifying anomalies, resource allocation optimization and improving security. As a result of the application of machine learning these systems carry out operations, reduce human operation and also detect abnormal behavior therefore minimizing and finally ensuring smooth service delivery. Yet ontology-based models add extra value through explicit and formal specification of cloud-related concepts. Ontologies enable to map the concepts, their relationships and attributes which in turn improves the semantic understanding and reasoning ultimately making the decision making and the problem-solving easy. Interoperability and knowledge-sharing across distinct cloud environments are facilitated by ontologies, hence data and tools can be easily exchanged. Overall, AI and ontology-based solutions are very efficient tools for optimizing cloud operations, improving the utilization of resources and strengthening the security during use of cloud services in the age of the dynamic computing landscape. patterns and to maintain the standard efficiency in resource utilization and QoS. Overall, through the use of DRL-based solutions, we see a viable means of dealing with cloud resource problems which include ability to manage QoS, energy efficiency and adaptation to a dynamic cloud environment.[3]

The ReDCIM (Reconfigurable Digital Computing-In-Memory) Processor represents an unprecedented breakthrough in the field of cloud-based AI acceleration by providing groundbreaking levels of efficiency, accuracy, and transformability. On the basis of the technology of computing-in-memory which is directly integrated into the architecture, ReDCIM takes the shortcut of data movement and latency, especially for the acceleration of AI workloads in the cloud. Its unique design combines FP and INT operations in a single pipeline, streamlines resource usage by having the structures become all-in-one, instead of having independent units. This unified pipeline guarantees high level of effectiveness, precision aLSTM algorithm for dynamic resource allocation Comparison of Long-Short Term Memory and Monte Carlo Tree Searched versatility in AI jobs while dealing with miscellaneous workloads that arise in cloud systems. In addition, it boosts energy efficiency by eliminating extra hardware which is very significant for cutting down the power bill in the cloud data centers. The scalability of ReDCIM to varied AI algorithms and models is a solid proof that it is a very reliable solution for cloud-based AI applications whose computational needs keep changing. Its balance between performance, accuracy, and energy efficiency gives the solution the necessary AI-driven functionality with the same guarantees of resource utilization and cost efficiency in cloud environments.[4]

Performance of the databases quite often relies on optimizing various parameters that are considered to be tunable knobs, which enable administration to manage the behavior and resources of the system. Parameters like innodbbufferpoolsize and tableopencache among others underline a crucial role in determining query execution speed, data trending capability and the system's responsiveness generally. Take the innodbbufferpoolsize for instance, it controls the size of innodb buffer pool which is for more effective memory usage thereby reducing disk I/O operations, while on the other hand tableopencache determines the number of table descriptors cached in memory and this has an effect on the heart of the database in its ability to retrieve information from a table during query execution. Impropriation of these leavers with regards to the workload features is vital for efficient use of resources, reducing latencies, and ensuring the scalability, concurrency, and reliability. It is the constant supervision and management of these parameters which play a key role in maintaining quick response, reliability at its peak and extended capability to accommodate the changed workload pattern and resource constraints.[5]

II. LITERATURE REVIEW

Load balancing plays a crucial role in cloud computing by improving resource allocation, whereas LSTM and MCTS are two leading methods. LSTM which is known for tapping into the temporal dimensionality has proved to be effective in the accurate prediction of workload patterns and dynamic resource allocation. Also, MCTS algorithms are strong in decision-making under uncertainty which makes them also effective in dynamic clouds. Recognition of both the LSTM and MCTS techniques' strengths and limitations from recent comparative studies has helped users in choosing the most suitable method for their systems. And following this research, the present model intends to unify the advantages of LSTM and MCTS to increase load balancing precision and reduce error rates considerably, which leads to the best system performance and resource utilization in a cloud environment.[1]

Efficient allocation of cloud resources is the backbone of optimum performance in cloud computing. The standard methods that go by the name of traditional approaches have been around and applied in similar cases for a long time. These methods employ predefined heuristics or optimization algorithms, but they may not perform well during workload patterns dynamics unless they do adaptation what results in poor resource utilization. On the contrary, Deep Reinforcement Learning (DRL) has drawn increasing attention in the last years to optimize the cloud resources. DRL-based solutions provide adaptability and self-optimization a since such systems tune optimal resource allocation policies through interaction with the environment. There have been many focus studies on traditional methods such as greedy algorithms and genetic algorithms emphasizing their strengths and weaknesses. Other studies have also discussed the implementation of DRL techniques including deep Q-networks (DQN) and proximal policy optimization (PPO). Such studies uncover the need of cloud resource allocation and investigate how DRL-based algorithms may provide more sustainable and dependable solutions for the movable cloud environment. Besides to comparative analysis, additional investigations are required to accurately determine opportunities and weaknesses of the classical and DRL-based systems in different cloud scenarios.[2]

Research in database optimization for cloud environments spans two main categories: working on an adjustment of the physical construction and searching for the optimal values of configurable elements. The studies explain among other solutions these techniques being indexing, partitioning and storage management to boost the speed of processing queries and reduce storage costs. Furthermore, the leading database vendors such as Oracle and Microsoft also offer apparatus including Oracle's Automatic Database Diagnostic Monitor and MS-SQL Query Store that enable automated configuration parameter-tuning based on characteristics of workload. Open-source tools MySQLTuner and pgTune are similar in the fact that they can be used for MySQL and PostgreSQL databases respectively too. They attempt at optimizing databases configuration and resource utilization which in cloud-based application implementation stress the application of automatic tuning in the cloud space in order to dynamically adapt with ongoing implementation.[5]

Task scheduling method plays a critical role in energy efficiency and performance improvement in fog-cloud environments comprising of cloud centers and edge devices in which the resources are distributed. Recently, many researches have started to combing AI methods to the scheduling of such dynamic as well as various environments. AI-based techniques, including machine learning and optimization algorithms, show a great potential in developing systems that use energy in a dynamic manner, such as allocating tasks to resources while applying consideration to the energy constraints, workload characteristics, and system dynamics. The research conducted by [Author] and [Author] examined the way AI optimized scheduling algorithms can help cloud-fog computing achieve a greater energy efficiency as well as the number of tasks finished within a certain timeframe in comparison with traditional systems. Such algorithms apply the AI approaches, e. g. reinforcement learning, genetic algorithm, and ant colony optimization, to intelligently manage the workload and energy consumption dynamically. The ability to adjust to different environmental conditions and workload requirements in AI-based scheduling algorithms enables the system to attain equilibrium between energy usage and task completion. As a consequence, system performance and resource utilization can be enhanced. Among the algorithms, which show a good example of such an algorithm is HunterPlus created by [Author]. Fog computing solution that combines reinforcement learning with heuristics to improve energy efficiency and job completion rate performance in cloud-fog settings. Iterative learning and adaptation are imminent in HunterPlus. This results in tasks being dynamically allocated to computing resources, taking into account factors such as task dependencies, resource availability and energy constraints. As compared to the other scheduling algorithms a comparative evaluation HunterPlus is doing better in terms of energy efficiency and job completion rate which makes it a promising candidate for practical deployment on the cloud-fog environment.[6]

III. FUNDAMENTALS OF CLOUD AUTOMATION

A. Definition and Significance of Cloud Automation

Cloud automation comprises practically all software and solutions that apply to automate cloud-related operations like resource management, provisioning, and orchestration. It performs a critical role by minimizing the operation timescales, improving the effectiveness and lowering the human intervention levels for cloud operations. Automation of the routine tasks like resource provisioning, configuration management, and scaling by cloud technology definitely plays an important role in the acceleration of resource utilization, development of new applications, and service environment. In addition, automated cloud deployments are a vehicle for consistency, reliability and repeatability in cloud infrastructure on top of ensuring best-practice and compliance points adherence.

B. Key Components and Processes Involved in Cloud Automation

Cloud automation refers to a collection of the device and the process of the cloud infrastructure automation within its operations. These relevant parts could be device configuration tools, systems for configuration management, orchestration platforms, as well as the monitoring solutions. Proving of tools adds automation of deploying and configuration of cloud entities and configuration management systems make sure the standards of the established configurations of infrastructure components and prompt them. Orchestration platforms are aimed at the management and automation of highly-tiered and complex application and service environments. Monitoring service offers immediate visibility into cloud resources and metrics, making it easy to identify and optimize those resources at the go. The procedures of IaC, CI/CD and PDA are coming across as being the important feature of cloud automation.

C. Challenges and Limitations of Traditional Cloud Automation Approaches

Traditional cloud automation methods have some advantages, but indeed, they come with quite a few challenges and limitations. They are, among others, managing various cloud environments, legacy systems technology compatibility problems, as well as disparities upon standardization between the tooling and practices of automation. Moreover, scalability of automation processes is difficult across big, complex distributed environments, hence at times they might be incompetent. In addition, extensiveness and fast-changing of the cloud infostructure might make the conventional automation methods helpless in this regard. Resolving these problems is through implementing wide and comprehensive approaches to cloud automation of which novel technologies like AI and ML are availed to improve delivery standards, scalability, and adaptation.

IV. ROLE OF AI IN CLOUD AUTOMATION

A. Introduction to AI technologies relevant to cloud automation:

Artificial intelligence technologies nowadays are vital aids for cloud automation. They offer an intelligent approach to decision making, optimization, and automation of different processes. In addition, technologies that are relevant to artificial intelligence comprise of ML, NLP, DL, RL, and predictive analytics. Machine learning algorithms are the essence of systems to learn from data and give decisions and predictions by themselves without programming them explicitly. The human language processing is done by means of systems capable to understand and generate natural language through which artificial intelligence (AI) can interact with cloud users. Deep learning algorithms, which are a subset of machine learning family, are doing very well in this domain in that they can process and analyze large amounts of data and can be used for complex tasks such as image recognition and pattern recognition. The reinforcing learning algorithms solve the problems of changing and uncertain conditions in environments and they are a good match for the tasks of that kind through trial and error. Predictive analytics utilizes statistical tools and machine learning algorithms to anticipate future patterns or events from data in the past, thus, it allows taking action early to minimize risks and optimize resources in cloud environments.

B. Benefits of integrating AI into cloud automation processes:

The introduction of AI into the automation of cloud for instance empowers the following capabilities such as efficiency, flexibility, scalability and affordability. Automation backed up by AI's power assist in the provisioning of faster and more accurate scaling and management of cloud resources, which are essential for intelligent and effective decision-making in the fields of provisioning, scaling, and management of cloud resources. AI algorithms can reference large-scale data in real-time, identify patterns, deviations and trends, consequently, the system allows proactive problem-solving and predictive maintenance which makes it possible to be ready beforehand. On the other hand, AI-powered automation improves adaptability and operational resilience by optimizing the workload and providing the best possible response to both changing conditions and factors of the environment. AI relevant automation also enables companies to increase the efficiency of the resource usage, trim the operational expenses and also enhance the operation levels of their cloud services.

C. Overview of AI-powered techniques for optimizing cloud operations:

AI-assisted means of management of operations in the cloud happen to be within a large number of applications. They comprise resource management, load scheduling, fault detecting, security monitoring, and cost optimization. In particular, AI algorithms are capable of getting the most out of resources by automatically adapting virtual machine (VMs) configuration, container placements, and workloads distribution to react, if necessary, to varying demand patterns and performance requirements Thus, AI algorithms have the capacity to optimize resource allocation. Intelligent task distribution is accomplished by AI planning algorithms, which employ various techniques to improve job processing efficiency, i. e. utilize computing resources to the max and minimize the time necessary for job completion. AI includes fault detection and prediction algorithms based on a system's log, performance statistics, and event data that enable it to identify in advance potential failures or abnormalities and, as result, mitigate problems and use checkpoints to ensure fault tolerance. AI-powered security monitoring tools are accomplished by utilizing 3 known techniques, they are anomaly detection, threat intelligence, and behavioral analysis. They aim to detect and minimize the security threats usage cloud environment. Apart from that, AI-entrenched cost optimization algorithms which are built on the principle of analyzing usage patterns and pricing model will seize upon the chance of cost savings and optimization techniques such as spot instance utilization and resource consolidation. Generally, AI-based techniques are a key component in improved cloud operations, giving superior decision-making, automation and optimization for the benefit of a number of types of cloud operations and services.

V. Exception Methodology

An algorithm called Long Short-Term Memory (LSTM) from the category of recurrent neural networks (RNN) is used for dynamic resource allocation in the cloud computing. LSTM models capture temporal dependencies and patterns in sequential data very well, which makes them an adequate tool for future resource demands prediction using historical data as reference. It permits for proactive resource management, so it leads to resource utilization optimization and elimination of performance degradation. On the other hand, Monte Carlo Tree Search (MCTS) is a tree search algorithm frequently applied in decision-making and optimization cases involving unpredictability, e. g., task planning and resource division in cloud environments. LSTM is concerned with sequencing modeling and prediction, while MCTS tries out actions at decision tree nodes in order to discover the best strategies. Analyzing the two methods would amounts to examining both their advantages and disadvantages with respect to a certain cloud computing task, taking into account characteristics like the problem complexity and computational requirement.[1]

Being the building blocks of cloud computing, Actor-Critic Deep Reinforcement learning (AC-DRL) and gradient ascent with advantage function come into play and help in optimization of resource allocation and task scheduling. In an AC-DRL setting, cloud agents can leverage deep neural networks to develop more complex policies and value functions based on which decisions will be adaptively made and cloud resources will be better managed. The agent chooses moves suitable for the present cloud condition, while the critic helps to learn from the actions, executing actions in the dynamic cloud scenario. But for the update of policy parameters, especially with the advantage function, the policy parameters will be maximized through the expected rewards, although considering the advantage of some actions on the current policy. Employing this method in deep reinforcement learning models for clouds is aimed at more effective policy updates that translates to faster convergence while resources usage in cloud computing is optimized. These methods help cloud systems work through automatic resource allocation, scheduling activities, and the best performance possible contributing to both efficiency, scalability, and cost effectiveness of cloud services.[2]

Model based ontology in the cloud computing provides a foundation for storing knowledge in the form of well-defined and structured categories as well as concepts and manage the interoperability and heterogeneity of the resources and the data. These models are developed to define a shared understanding and formalized conditions. Thus, it becomes possible to make use of them in the areas of automated reasoning and smart choice-making in the cloud. At the same time, the machine learning based optimization, which is targeting a number of goals at the same time, is the solution to the conflict of performance measures: fast reaction time, cost and energy consumption. Employing algorithms like genetic algorithms, and Reinforcement Learning, multi-objective optimization techniques allow cloud systems to accurately plan resources, perform task scheduling, and use energy in an optimal way, taking into account various needs and cutting down operational expenses in a cloud computing environment.[3]

Together with innovative ReDCIM utilization that features a unified FP/INT pipeline architecture and bitwise in-memory Booth multiplication techniques make the main contribution in improving the computer performance in the cloud environment. The 3D-NIM includes the reconfigurable digital computing-in-memory processor with the unified floating-point/integer pipeline architecture, thus the built system can do various and high-throughput processing of the different computational tasks in the cloud. This architecture not only boosts the performance and flexibility, but it would also lessen the so-called overhead of traditional computing. On top of that, bitwise in-memory Booth multiplication improves CIM by utilizing efficient bitwise forms of multiplication instructions, which leads to less complicated algorithms and more rapid calculation speeds. These techniques help in scaling up the cloud systems in terms of throughput and the consumption of energy, an aspect which in turn makes the cloud services and applications more effective and pocket friendly.[4]

Cloud computing, the advanced techniques including deep reinforcement learning (RL) using deep deterministic policy gradient (DDPG) method and one of the strategies of trial and error with a limited number of samples for the initial training are vital for improving the system performance and resource allocation. The cloud system learning optimal decision-making policies via interaction with the environment is one of the most promising applications of deep RL, which can be achieved mostly with the help of the DDPG method. Efficient resource management and adaptation to fluctuating situations are the strengths of this approach. Furthermore, applying a successive iterative strategy in the case of insufficient samples is also beneficial, enabling the cloud systems to explore and learn from the environment with only small computational resources. As a result, cloud systems become better in taking informed decisions and optimal resource allocation in cloud systems characterized by dynamic and uncertain environments. These strategies aid in the development of cloud computing systems with better resource use, scalability and also adaptability leading to a better performance of cloud-base services and applications.[5]

VI. CONCLUSION

The amalgamation of AI technologies with the latest auto elements becomes the crucial factor to increase cloud management success. LSTM algorithms successfully become the source contributing to the person-ahead resource allocation. Using the data of previous days, the algorithms can predict the approaching bottlenecks, provoking action before it is too late. While the DRL platforms provide a dynamic procedure for better resources apportioning and consumption of energy, cloud environments, however, require a different approach. This philosophy encourages constant learning and adaptation so that the cloud solutions provide peak performance with minimum spend on the operational costs. Also, the fact that the ReDCIM processors along with the bitwise in-memory Booth multiplication techniques have been integrated results in a remarkable progress in the area of computational acceleration in the cloud. This technology equips clouds with the ability to handle advanced computation tasks unparalleled and flexible the efficiency and resource utilization and the efficacy in turn. Using such sophisticated tools, cloud machines are able to reach their best performance level, they can be scaled and be cost saving, all this to continuously provide seamless services in changing computing environment. The cloud platforms employ active resource distribution, flexible resource management, and increased computational speed so as to budget the increasing needs as well as sustain high performance and reliability. To conclude, the AI integration as well as ongoing development of innovate methods is the key to a balanced effectiveness and making the cloud automation “do more with less” effort a success.

VII. REFERENCE

1. MOSES, ASHAWA., OYAKHIRE, DOUGLAS., JUDE, OSAMOR., RILEY, JACKIE. (2022). IMPROVING CLOUD EFFICIENCY THROUGH OPTIMIZED RESOURCE ALLOCATION TECHNIQUE FOR LOAD BALANCING USING LSTM MACHINE LEARNING ALGORITHM. JOURNAL OF CLOUD COMPUTING, DOI: 10.1186/S13677-022-00362-X

2. (2022). ADAPTIVE AND EFFICIENT RESOURCE ALLOCATION IN CLOUD DATACENTERS USING ACTOR-CRITIC DEEP REINFORCEMENT LEARNING. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, DOI: 10.1109/TPDS.2021.3132422

3. BENIAMINO, DI, MARTINO., ANTONIO, ESPOSITO., ERNESTO, DAMIANI. (2019). TOWARDS AI-POWERED MULTIPLE CLOUD MANAGEMENT. IEEE INTERNET COMPUTING, DOI: 10.1109/MIC.2018.2883839

4. FENGBIN, TU., YIQI, WANG., ZIHAN, WU., LING, LIANG., YUFEI, DING., BONGJIN, KIM., LEIBO, LIU., SHAOJUN, WEI., YUAN, XIE., SHOUYI, YIN. (2023). REDCIM: RECONFIGURABLE DIGITAL COMPUTING- IN -MEMORY PROCESSOR WITH UNIFIED FP/INT PIPELINE FOR CLOUD AI ACCELERATION. IEEE JOURNAL OF SOLID-STATE CIRCUITS, DOI: 10.1109/JSSC.2022.3222059

5. FENGBIN, TU., YIQI, WANG., ZIHAN, WU., LING, LIANG., YUFEI, DING., BONGJIN, KIM., LEIBO, LIU., SHAOJUN, WEI., YUAN, XIE., SHOUYI, YIN. (2023). REDCIM: RECONFIGURABLE DIGITAL COMPUTING- IN -MEMORY PROCESSOR WITH UNIFIED FP/INT PIPELINE FOR CLOUD AI ACCELERATION. IEEE JOURNAL OF SOLID-STATE CIRCUITS, DOI: 10.1109/JSSC.2022.3222059

6. SUNDAS, IFTIKHAR., MIRZA, MOHAMMAD, MUFLEH, AHMAD., SHRESHTH, TULI., DEEPRAJ, CHOWDHURY., MINXIAN, XU., SUKHPAL, SINGH, GILL., STEVE, UHLIG. (2022). HUNTERPLUS: AI BASED ENERGY-EFFICIENT TASK SCHEDULING FOR CLOUD-FOG COMPUTING ENVIRONMENTS. INTERNET OF THINGS, DOI: 10.1016/J.IOT.2022.100667