IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

Power-Efficient Semiconductors for AI at the Edge: Enabling Scalable Intelligence in Wireless Systems

Goutham Kumar Sheelam

IT Data Engineer, Sr. Staff, ORCID ID: 0009-0004-1031-3710

Abstract: Artificial Intelligence (AI) based applications are increasingly being deployed on the edge of the network due to the faster response times required by many of these applications [1]. At the same time, low latency operation becomes important as real-time edge datasets may not be amenable to storage on the cloud. However, this leads to a larger number of processors that are more broadly distributed over the edge of the network. Power constraints on these edge processors or AI accelerators present a challenge since these devices will usually need to operate on battery power. This may also lead to an increase in the carbon footprint of AI systems as more AI accelerator networks are deployed.

AI applications typically consume more than 50% of the power in infrastructure devices or the edge of the network. Within these devices, the AI accelerator which performs AI inferences consumes a large amount of power (typically over 50%). Subsequently, reductions of both active and idle power of AI accelerators will significantly improve the energy efficiency of infrastructure devices and overall AI systems. Various innovative methods can be used to lower the active power including quantization, pruning, RMSE loss minimization, hybrid architectures which combine different types of AI accelerators, and sparsity.

A significant portion of the power consumption of AI accelerators is due to idle power caused by the scheduling and repeated wake-up of the AI accelerators to perform AI inference tasks [2]. Redundant carry-free designs and asynchronous designs can be used to lower the active power. After compiling, sparse-structured weights can be used to reduce memory access delay, a primary source of power and delay in AI inferences. In order to better analyze the power savings due to sleek operation, a power in the fabricated unit of workloads should be explored.

Keywords: Power Efficiency, Semiconductors, Edge AI, Low Power Design, Scalable Intelligence, Wireless Systems, Embedded AI, Edge Computing, AI Accelerators, Energy Efficiency, Real-Time Processing, IoT Devices, Neural Network Hardware, Smart Sensors, Signal Processing, AI Chips, 5G Integration, Tiny ML–, Hardware Optimization, System-on-Chip (SoC)

I. INTRODUCTION

To address the growing demand for edge intelligence, there is a growing need for energy-efficient, high-performance intelligent chips for AI inference at the edge. The current electronic chips in edge devices need to be designed to provide intelligent processing while reducing energy consumption and latency. Therefore, the research and advancement of novel semiconductor technologies will dramatically change the existing SC technology and will accelerate the next wave of intelligence in the area of smart edge devices. Neuromorphic computing with brain-inspired architecture, material, and device designs will enable seamless and efficient AI applications at the edge. Many of such applications currently rely on data center-based AI technologies, which are notoriously power hungry, slow, expensive, and privacy-violating [1]. Such a development and integration of new SC materials and techniques will open a range of new AI applications beyond the limits of CMOS electronics and conventional semiconductor SCs in information technologies. These applications are critical to the need for high speed, high bandwidth, low latency, low power, compact size, and on-chip processing for parallel processing of information.

Edge devices powered by battery sources will proliferate for smart applications in autonomous surveillance and attack detection, smart homes and offices, healthcare and personal health monitoring, and environmental and biosensing. While promising, these applications will consider high demands for ultra-low power and energy-efficient signal processing and machine learning algorithms to ensure user privacy. However, conventional von Neumann computing architectures can hardly meet these stringent requirements. Edge AI is viewed as the local intelligence that can lower the latency and energy consumption, while improving the data privacy. Despite great potential, challenges remain, particularly in the energy- and power-efficiency of the current primary hardware implementations [2].

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

State-of-the-art low-power edge AI is currently implemented with specialized digital chips, which, suffering the von Neumann bottleneck, cannot match the aforementioned performance scalability. Power-efficient chip-scale implementation of non-von Neumann computing for edge AI applications and devices is thus of utmost importance for the future of the thriving edge intelligence era.

II. OVERVIEW OF AI AT THE EDGE

The tremendous advancements in data analytics and machine learning algorithms have propelled artificial intelligence (AI) capabilities into a wide range of applications over the last decade. In recent years, there is a growing recognition of the need for protecting the data privacy, security, and reliability in AI. The emerging trends of AI in wireless systems or at the edge enable low-latency and privacy-preserving data intelligence while having to tackle stringent resource constraints. AI at the edge can be understood as progressively pushing intelligence – machine learning algorithms (ML) to extract useful information from data – from the cloud centers to devices that are hosting wireless interfaces. There are the opportunities and challenges across the technological stack in the wireless systems from a communication-centric perspective.

Eqn.1: Dynamic Power Consumption of CMOS Circuits

$$P_{dynamic} = \alpha C_L V_{dd}^2 f$$

- α: switching activity factor
- C_L: load capacitance
- V_{dd}: supply voltage
- f: clock frequency

The primary challenge in intelligence at the edge for wireless systems is that unlike traditional cloud-assisted AI systems where the known AI algorithms and the optimization objective are globally shared across devices, the learning model would be unknown to the resource-constrained devices. The wide diversity in the device capabilities would also create challenges in AI model training and deployment [3]. This necessitates the urgent need for innovative and scalable communication strategies for data and information dissemination, AI model training and update, and inference acceleration. On the other hand, there are a wide range of AI-related problems across the communication, circuits, and devices domains either with large datasets or iterative greedy solvers. The imperative energy, latency, or monetary privacy constraints brought by the edge settings can have different value propositions for the use of ML tools [2]. Moreover, this could also gain new perspectives on unfolding the interpretability and robustness issues in AI black boxes. AI has empowered a myriad of intelligent systems across diverse applications, including autonomous driving, augmented reality, finance, IoT, and smart agriculture. There is a clear trend towards moving the intelligence of those applications from the cloud to the edge with the aspirations of enhancing data privacy, reducing latency and bandwidth consumption. Edge intelligence is a convergence of Edge computing and AI. Edge computing offloads data across a spectrum of devices to deliver bandwidth, computation, latency, and energy efficiencies. Edge computing encompasses a whole toolbox of computation paradigms from heterogeneous computing devices to robust frameworks for distribution and collaboration.

III. IMPORTANCE OF POWER EFFICIENCY

The rapid growth of AI applications has resulted in an immense increase of computational loads, which drives the need for deploying AI and ML algorithms at the wireless edge. The edge is a natural location for some degree of processing since it can leverage the proximity to resource-rich cloud-based computing centers and distribute workloads with fluid and flexible workloads. Wireless edge AI has received increasing attention from both academia and industries, leading to various technical solutions across the networks. In the new era of AI, beyond higher learning accuracy, achieving energy efficiency (EE) is equally crucial. AI workloads are massive, in part due to the ulterior number of weights, and effectively waste many computing resources or battery life. Recent studies reveal that as AI algorithms approach their performance limits, the energy consumption required to achieve further intelligence improvement increases super-linearly.

Advances in device technology have been driven by the explosive increases in the number of internet-connected devices, mobile traffic, and computational demands for large-scale deep learning on cloud systems. The requirement for large power and energy consumption has introduced serious concerns regarding environmental pollution and climate crises. The energy consumption growth has been orders of magnitude faster than the population growth, asking for energy-efficient (EE) designs to enable sustainability of information technology.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

To optimize edge AI systems for EE, it is of paramount importance to holistically consider three critical energy components when measuring the total system power consumption: the energy consumed in each EUD for collecting information, power-hungry on-board computation chips for running AI models, and communication units for transmitting the inferences or reports over wireless connections.

A promising approach to improve the EE of edge AI systems is through co-optimization of all three energy components. To strike a balance between the achievable intelligence and consumed energy, the interplay between them must be well addressed across the entire edge AI pipeline. The vision of green and energy-efficient edge AI is new and poses significant challenges. It requires not only innovation in algorithms but also re-architecting the existing mobile networks so that the EUDs and MEC servers can cooperatively offload their tasks in the new multi-cluster heterogeneous network architecture. There has been widespread consideration of EE in the operation of encompassing computation and communication, including strategies for efficient bandwidth, power allocation, resource allocation, and clustering [1].

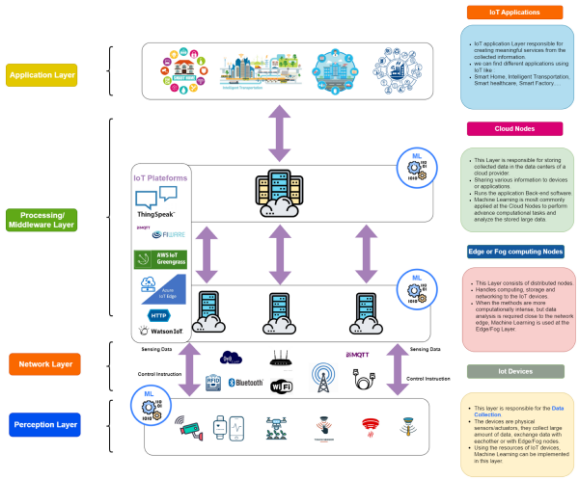

Fig 1: A Survey of Machine Learning

IV. SEMICONDUCTOR TECHNOLOGIES

A strong trend exists to perform statistical learning closer to the data, where physically interpretable and domain-specific assumptions on the data patterns often exist. To this end, eDNNs are constructed with different classes of nonpolynomial functions. These options are extensively investigated both in simulation and in design examples based on several types of eDNNs such as resonance graphs, recurrent and stochastic networks. For the first time, it is demonstrated that via systematic network cascading together with an optimization of locally convex layer parameters, extreme improvements in performance can be obtained automatically.

The edge is commonly understood as the border of a system or a structure. For big server farms and data centers, the edge is envisioned as a demarcation point, where data does not need to go farther to reach the information sink. For other systems, the edge can be a more ambiguous concept, where data can still be processed farther by cloud servers, but at the cost of unavoidable lags and excess financial expenditures. Edge servers may take on various hardware and hosting forms, such as small cells and routers. Regardless, AI algorithms can be applied on edge servers to process raw data and extract actionable insights. When it comes to semiconductors and circuits, edge servers are regularly realized as application-specific chips, which may take the form of either device-integrated or monolithically integrated chiplet modules. These chips are aimed to maximize efficiency to handle the specific statistics and features of data values.

Detection schemes in wireless systems can be thought of as quadratic data processing systems even though their hardware implementations are often linear to retain efficiency. Once the wake-up signal is correctly detected, all of the energy-efficient signal processing and ML can then be performed within the power-efficient chiplet module. In contrast, all edge inference has to be performed in application-specific chips designed for general gradient-based learning algorithms. So far, the development of wireless and semiconductor technologies has been largely decoupled from each other, leading to inefficient AI systems on the edge.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

4.1. Silicon-Based Semiconductors

based short-channel field-effect transistors (FETs) down to a few-nanometer scale are main-stream devices today. Being a group-IV semiconductor, silicon allows the growth of high-quality silicon on any substrate and is also CMOS compatible. The framework of traditional complementary metal oxide semiconductors (CMOS) yields great versatility for auxiliary and mixed-signal circuit designs. Nevertheless, native bandgap in silicon restricts the use of silicon devices in high-frequency and optoelectronic applications. Beyond silicon research, there has been increasing interest in alternative semiconductors displaying appealing properties such as wide bandgap, energy band alignment, and atomic thickness. Promising candidates include 2D single-element black phosphorous and group-VI TMDC of MoS2, MoSe2, WSe2, and more. Because of its direct bandgap, WSe2 FETs are regarded as a potential building block for high-frequency and optoelectronic applications. Being engineered chemically, mechanically, or electrostatically, TMDC FETs are also intensively investigated as realizations of tunneling transistors, low-voltage/multi-state FETs, light-emitting devices, etc. Emerging nanomaterials have distinct advantages while they also face tremendous challenges on scaling up with varying factors such as polymer dielectric traps, reactivity, scatterers, electron mobility, etc. The emergence of hetero-structures/stacks integrating two or more semiconductors raises a new paradigm of device design with additional degrees of freedom in trajectory, geometry, and phase engineering. Recent progress includes black phosphorene and h-BN hetero-structure, WSe2 and MoS2 FET, TMDC photodetector, and more. Starting from CNFETs and going through a variety of other nanomaterials, characteristics of such devices spanning from traditional to emerging/heterogeneous semiconductors conclude the fine balance of optimizer simplicity, accuracy, efficiency, and interpretability.

A large number of recent studies focus on demonstrating fissile or reconfigurable synapses at the circuit and device levels, with varied approaches based on electrothermal behaviors, ferroelectricity, phase-change chalcogenides, and TMDC technique. Hereby is presented an experimental demonstration of a portable and reconfigurable time-dependent input weight architecture based on WSe2-alloyed BTA synapse devices. The architecture is equipped with a compact analog spike-timing-dependent plastic learning scheme and fast digital-to-analog-converter circuits permitting efficient training of multi-feature images. In addition, organic semiconductors and other organic soft materials display great potential for neuromorphic event-driven beyond von Neumann networks. The synapses neuromorphic platform is shown to perform effective edge enhancement of x-ray images with predicted errors and power consumption efficiently compatible with real-time processing at IoTs.

4.2. Wide Bandgap Semiconductors

However, performance improvements are expected to saturate soon. This is due to the technology used for further improvement, which is Si-based, and as Si is reaching its physical limitations. Therefore, new materials with physical properties exceeding those of Si are needed. Two strategies for the development of devices based on novel semiconductor materials can be considered: new materials more suitable for different applications, such as wide bandgap (WBG) semiconductors for high-power and RF electronics and ultra-wide bandgap (UWBG) materials for higher power levels; and new approaches to device design, e.g. 3D architecture [4]. Specifically, the discussion on new WBG – SiC and GaN – and UWBG – Ga2O3 and diamond – semiconductor materials for high-power applications is needed. Ga2O3 is emerging as the only material for which large and high-quality single crystals are foreseeable thanks to appropriate crystal growth techniques. WBG semiconductors offer distinct advantages over Si such as a wider bandgap, a higher breakdown electric field and a higher thermal conductivity [5]. The first two features offer higher on-state drain voltages at the same drain current density and much lower off-state drain leakage currents at the same on-state restriction voltages. The consequence is much lower conduction and blocking losses, while the higher thermal conductivity provides better cooling efficiency of the devices. Ga2O3 is a UWBG semiconductor with a 4.8 eV bandgap. It has been known for long thanks to several applications in transparent conducting oxides for low voltage ends in displays. Recently, a renewed interest spurred on Ga2O3 devices for high-power applications.

Power electronics is subject to the same technology advancement and market pressures as many other semiconductor device areas. Although there will always be a need for low power density (60 W/in3) ICs as found in wearable and ambient intelligent systems, silicon (Si)-based devices are reaching their performance limits in portable applications. With electrical energy constituting 40% of the total primary energy usage in the USA, a more responsible way of using the bought-in electrical energy is needed. Concerning electronic energy processing, the only way to achieve higher energy efficiency is through the use of higher voltage-rated (integrated) components in combination with higher carrier saturations velocities. Concerning components rated for larger voltages, new materials/tools that are better suited for high power (at lower parasitic capacitance/packaging costs) can be foreseen. As the Si material limit is reached with the eventual demise of Moore's law, this quest for materials are no longer hypothetical.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

4.3. Emerging Materials

The rising utilization of human-centered technologies is driving the advancement of wireless edge devices with computing capabilities. A variety of artificial intelligence (AI) algorithms are being deployed on smaller and power-restricted edge devices ranging from smart cameras for intelligent surveillance to profile ECG monitoring for at-scale health care. Edge-AI has numerous advantages over cloud-AI, such as sustainable AI with limited carbon footprints and enhanced privacy with on-device and secure AI. Unfortunately, with the explosive growth of AI model sizes in terms of both parameters and computations, the edge devices are now facing severe power challenges [6]. Therefore, it is vital for the semiconductor industry to develop energy-efficient computing technology architectures as well as new materials and devices that excel traditional silicon-based technology to extend Moore's law for better performance, energy efficiency, and scalability for neural networks, and thus, AI models.

The goals and requirements for edge intelligence are diverse, requiring specialized processors, memory architectures, and soft/hardware co-design for efficient denser storage, high-speed data transmission/retrieval, energy-efficient computing, and efficient interconnect systems to meet millisecond-class end-to-end latency challenges [2]. Also, as the network architecture becomes larger and more complex, the scale of AI is expanding rapidly toward trillions of parameters, requiring at-least exaflops-level computations during training. This paper highlights emerging materials and devices that leverage unique physical principles for efficient edge-AI. Emerging memory-class devices such as ferroelectric and resistive memories exhibiting unique switching mechanisms have the potential for in-memory computing technology architecture to minimize the I/O overhead. Diverse neuromorphic devices based on both emerging materials and technologies faithfully mimic biological neural networks at low power consumption. High-power-efficient architectures based on both traditional von Neumann architecture and emerging hybrid architectures are highlighted with the co-design of devices/operations. Finally, an outlook for future materials and devices with new functionalities for AI at the edge is provided.

V. AI ALGORITHMS AND THEIR REQUIREMENTS

Deep learning-based advanced artificial intelligence algorithms can track the joint location, velocity, and state of the target in real-time. In most applications, under energy constraint, it is required to understand the trade-offs between the accuracy, latency, and energy of the algorithms. In edge applications, the bandwidth and speed of Internet-of-Things nodes are limited. This leads to the emergence of sub-optimal AI algorithms that consume minimal latency and energy while guaranteeing a minimum accuracy. Direct computation of the joint probability density function and smooth approximation of the optimal energy-efficient path need high throughput, which may not be feasible at all energy budgets. Probabilistic graphical models-based inference algorithms have the potential to compute the low dimensional approximation of the probability distribution maintaining optimal trade-offs. While probabilistic graphical models have been illustrated to be highly effective in achieving complexity reduction in inference too large deep neural networks, probabilistic graphical models for state-space models encompassing a hidden generative model has for long been intractable for larger networks with higher order latency. The upsurge in deep neural network-based edge and cloud visual-SLAM raises the urgent need for AI algorithms that maintain real-time efficiency, resilience, and privacy at the edge, while maintaining accuracy.

State-space models, particularly in the Bayesian framework, offer first-principles formation of generative models in sparsely connected graph-manifold. Smooth approximation of the distribution and posterior has been proposed with infinitely large regressive structures at the frontier of AI algorithms. This limits the tractability of AI algorithms, with existing works hardly extending such intractability to the platform AI. To bridge the vast gap between intractable AI algorithms and the existing hardware, it is first shown that while sparsity helps obtain the base tree of learning, an introduction of connection water-filling theory of random graphs run on greedy-sparse coding architecture enables the tractable implementation of arbitrary probabilistic graphical models and state-space models on localized edge AI platform. The proposed theory explains the major differences in existing competing non-localized transformer-like recurrence architectures.

Eqn.2:Total Energy Consumption per Inference

$$E_{inf} = (P_{compute} + P_{memory} + P_{I/O}) \cdot t_{inf}$$

- $P_{compute}$: compute power for neural network inference
- P_{memory}: power for memory access (often dominant)
- $P_{I/O}$: power for data communication (e.g., with sensors or RF modules)
- t_{inf}: time per inference

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

5.1. Machine Learning Models

Machine Learning (ML) algorithms have been successfully employed in a number of applications, from classification and detection to prediction problems. These algorithms are increasingly being designed and trained off-line during a process called training. The trained models can then be transferred to devices for execution, referred to as deployment. The deployment of ML models for inference tasks has continuously led to the creation of the so-called Edge Artificial Intelligence (AI) paradigm. While increased awareness of privacy and security is the driving force behind the Edge AI paradigm, the resource scarcity in terms of computational power and energy of these resource-constrained devices needs to be duly contemplated as well.

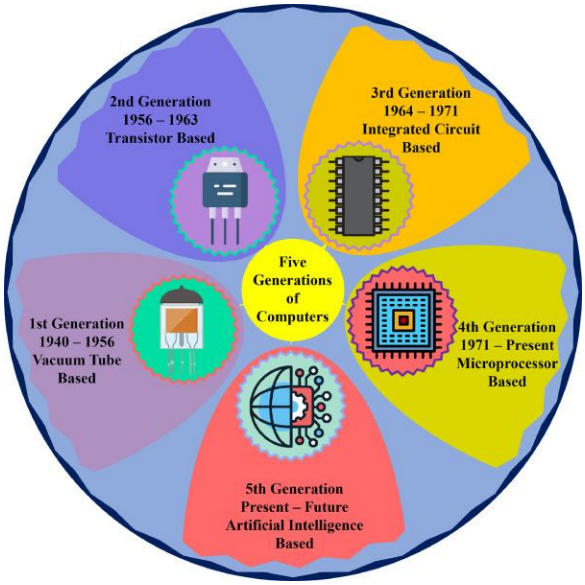

Fig 2: Overview of emerging electronics technologies

Common ML Layer types minimize the amount of dedicated hardware resources while also avoiding the usage of Floating-Point (FPT) number representations. This drives hardware vendor competition for radical improvement of the performance and efficiency of these device types, in order to catch-up with the software only performance improvements in higher precision implementations. On the other hand, quantization strategies can significantly impact edge-device-wide ML model design [8]. FPGA-based direct-feed techniques are also acknowledged for model-architecture aware Dataflow generation and scheduling methodologies that favor the high throughput execution and low power consumption of DNNs on FPGAs. FPGA-based high-level programming language are envisioned to target even more devices and include custom architectures, allowing for a much wider span of space and power optimizations.

5.2. Deep Learning Architectures

Deep learning technology enables scalable intelligence by reducing the AI barrier for new use cases in wireless systems. Sparse 1D convolutional neural networks (CNNs) are investigated to address the challenges of implementing deep learning in on-device training while being low power and energy. Massive connectivity is the key to connecting many sensors in a wireless system, allowing them to share their data with a low-cost processor. Intelligent devices that utilize deep learning are progressively developed to extract intelligence from massive data locally. However, low-power scaling in silicon technology nodes generates tighter timing constraints on device operation, making it challenging to implement deep learning while meeting the device power and energy budget. In addition, there is a growing demand for on-device training to improve model accuracy while protecting user privacy and preventing data breaches. In edge AI applications, edge devices usually have less power, throughput, and die area than cloud solutions. Implementing deep learning with a low number of computations is vital for enabling scalable intelligence with low-power, low-energy AI in edge devices.
The analysis of performance and complexity shows that sparse architectures outperform dense architectures. Sparse architectures with up to 93% of weights pruned can achieve similar accuracy while being significantly lower in operations and energy. With such sparsity in weight and sparsity in activation, the performance of sparse 1D CNNs is further improved. This architecture employs weights spanning multiple cycles, further decreasing power and energy. A dedicated hardware architecture is proposed to exploit the uniqueness of such sparsity without adding significant hardware overheads. It is shown that wireless sensors can effectively utilize sparse 1D CNNs designed with the proposed hardware-compatible methodology in low-power, low-energy on-device training while achieving a more than 3 dB accuracy

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

improvement to dense architectures [9]. These approaches extend the use of deep learning in wireless systems with stringent power and energy constraints.

Graph neural networks (GNNs) are investigated for scalable edge AI in wireless systems where the computation cost per node is lower than that of empirical data sampling and model local training on each node [10]. In evolutionary 5G networks involving both the complexity of new technologies and the time delay for deployment, distributed GNNs with parallel edge AI are further taxed for higher performance with much lower latency. In this framework, ubiquitous computation-unconstrained cooperative nodes are distributed across potentially all base stations, moving deep learning workloads, data, and parameters. One potential pitfall of such distributed GNNs is being captured by asymptotic homogeneity due to decentralized model training, leading to a strong performance drop. The proposed local training convergence is achieved by building information homogeneity in a gradual manner, achieving upper and lower bounds on its performance drop.

VI. CHALLENGES IN POWER EFFICIENCY

With rapid growth in AI workloads, such as deep learning models with billions of parameters, there is an increasing demand for power-efficient computing solutions to execute these workloads. As a result, significant research is underway to investigate new computing solutions and algorithmic techniques. Meanwhile, growing concerns over climate change, energy crisis, and societal issues are prompting a widespread interest in pursuing more sustainable energy use. The substantially growing energy footprint of AI models aggravates this problem. New approaches are needed to address the risk of a looming energy crisis and rising energy cost, aggravating ongoing climate change. Today's approaches to AI are becoming fundamentally misaligned with sustainable energy use. AI hardware and systems design today largely focuses on hyper-scalability, high computing throughput, fast learning, and simplistic co-design. Individually increasing these numbers is yielding diminishing returns in energy efficiency and deliverable intelligence. An interdisciplinary effort focused on sustainable energy use is emerging as an urgent priority for the research community. While there are a host of approaches to pursue power-efficient AI, the societal impact of this field is primarily determined by the power-efficiency of edge AI in owning devices, sensors, and networks, pre- and post-processing, transmission, and learning that recede from it. These proposed approaches are broadly placed into two categories: cost or power-efficient. Cost incentives, such as uncommon chip architectures, should realize substantial edge AI power-efficiency improvement. Equally important, simple cost-cutting measures can yield low-hanging fruit power gains. Some surveyed works explore these hard-chip approaches. However, overcoming chip-level challenges is futile without recourse to software, systems, and societal technologies to enable chip-scale intelligence. Various foundational directions to enable cross-stack power-efficiency improvement are highlighted.

Power-efficient AI is a universally important research and application problem. However, this topic is uniquely challenging for edge AI. Edge’s power-budget and performance-determining metrics are crucially divergent from those at cloud-scale. Today’s neural network training algorithms and edge AI hardware are crudely misaligned, rendering breakthrough performance improvements in AI learning and edge platforms very much in their infancy. Incomplete edge-chip theory, technology, architecture, and circuit development combined further challenge high performance. Edge AI societies is an important new research space with profound theoretical and algorithmic opportunism largely unexplored. Edge power-efficient AI techniques are becoming broadly applicable, such as new datasets and sparsity techniques that can be implemented in edge AI platforms. A major research direction at the intersection of circuit design, architecture, hardware-software co-design, and AI learning is beginning to emerge but has significant challenges ahead.

Fig 3: AI, wireless and security

6.1. Thermal Management

There exist several heat management issues associated with the pen resource management of semiconductor devices. For example, harmful thermal cycling of hardware in semiconductors could occur if short active periods are filled with idle

DOI: 10.17148/IJIREEICE.2022.101220

periods, close to the exciting device's thermal time constant. This is particularly dangerous for applications that involve unpredicted peaks of resource usage.
The spikes, which could be caused by new clients joining in applications or fluctuating demands from devices, might lead to a higher temperature than a steady workload of the same amount of average resource usage. Therefore, semiconductors require sophisticated thermal design and management approaches that involve careful material selection, design of system architecture, and temperature monitoring and management in operational environments.

Microscale thermal engineering of electronic systems has received renewed attention due to the challenges posed by fundamental scaling limits and increased power density. Traditionally, the packaging level in electronic systems has focused on providing a thermally passive enclosure for the chip. As a result, most existing chips are packaged without actively assisting with thermal control. The rapid rise in local power densities within the die has surpassed the ability of the passive thermal conduction path to remove heat. This has resulted in significantly higher thermal gradients, which not only degrade reliability and noise characteristics but also reduce performance by introducing latencies.

Thermal management using micro-scale and nano-scale structures for heat spreading, extraction, and storage devices is critical for the development of future electronic systems. The projected exponential increase in discrete device power and power density levels will exhaust the available thermal management solutions unless novel heat transfer techniques are invented and utilized. Because the number of transistors per unit area is approaching its practical limit, enhancement of circuit speed has shifted to multi-core architectures with an exponentially increasing number of devices per die. Simultaneously, the technology clock speeds have saturated and registered channel lengths shrunk to about ten times the characteristic pitch, which means that power density must increase in proportion to device numbers to maintain throughput. At the same time, reliability demands on high-performance portable electronics are increasingly stringent, and noise considerations have pushed the use of lower voltages.

6.2. Voltage Scaling

Reducing Vdd has a pronounced effect on energy – if only the voltage is changed, about an 80% drop in energy is seen with a 130mV drop in voltage. However, this reduction also leads to a large drop in reliability. Large amounts of checks are added, ensuring that a reduction in energy during computation does not compromise correctness of state, ultimately maintaining reliability. Despite the cost of these checks, enough energy savings can still be realized so that the net savings are substantial. This paper demonstrates this methodology of both reducing Vdd and adding checks to maintain reliability through simulations on a simple logic circuit. Voltage reduction allows net energy savings of 60% compared to operation at the nominal voltage, but also causes a significant drop in reliability. By adding lightweight checks that cost only a fraction of the base computation, reliability is maintained as voltage is reduced and about 50% energy savings is still achieved [11].

Reducing Vdd has a pronounced effect on energy – if only the voltage is changed, about an 80% drop in energy is seen with a 130mV drop in voltage [12]. However, this reduction also leads to a large drop in reliability. The motivation behind determining how to add checks to a circuit to maintain reliable operation is to find a practical method in an age of increasing concerns over power. New implementations of circuitry must keep energy consumption in check, yet a continuing reduction of the characteristic voltages of logic circuits leads to an ever-increasing number of new challenges. There is, however, a point where it is better to scale various aspects that are less dependent on outside sources and that lead to less degradation of the information they manipulate. The methodology described here addresses the validity of low power computations by ensuring the correctness of state by evaluating the output of a computation against lightweight checks to produce a guarantee of correctness similar to that of parity-based checking.

6.3. Energy Harvesting

For many Internet-of-Things (IoT) systems, the long-term operation of the sensors is critical, but when deploying without batteries, this becomes problematic. For this reason, energy harvesting systems are of escalating interest, in which the sensors scavenge energy from several energy sources and store this energy in electromechanical capacitors or batteries [13]. To reliably harvest energy, converters need to have a very low startup voltage (ideally in the mV range), high input-to-output voltage conversion ratios, and extremely low quiescent power. Current designs are typically bulky circuits and utilize discrete components. Because, with the pervasive growth of cellular and wireless Internet signals, signals that are too small to harvest energy become abundant, miniaturized integrated systems that can scavenge and use this energy to power ultralow-power devices are desired [14]. The efficient scavenging of ambient RF power is especially intriguing for the new class of mobile, inexpensive RF-powered devices: A source of energy whose signals are abundant and pervasive and that are often stronger and always higher in frequency than the RF signals used in present cellular and wireless internet communications. Very high efficiency and very low quiescent power operations enable this high-density

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

harvesting of spatially uniform transmitters like cellular and TV signals: Although waves of energy can be observed propagating in many forms from many antenna patterns, most of this energy cannot be used.

The future is a continuous ecology of energy sources that are designed to be wireless and will need micro anticipatory extraction, purification, and conveyance systems. To design the scavenger circuits, it was needed to establish the voltage scaling limit for long term operation of ultralow-power sensing circuits, and bound the constraints on including RF and vibrational harvesting circuits in the same package with a PCB fabrication technology.

VII. DESIGN STRATEGIES FOR POWER-EFFICIENT SEMICONDUCTORS

With a focus on the emerging multilingual capabilities of AI systems in a newly constructed edge setting, this section intends to visualize the extensive need for, challenges of, significant potential impacts of, and possible avenues for AI maturation in this new environment. Emphasis on new applications for AI in wireless system design and operation at the edge of the network. Potential positive impacts, challenges, and recent advances in this fresh setting that predicts a promising, scalable path for AI in wireless systems.

Within current network architectures, AI-assisted wireless systems already enhance operations within base stations, at the edge, and in the cloud. Ultra-large scale wireless systems with millions of connected devices are the intent of next generation designs. They will rely on power-efficient, scalable, and plentiful AI-assisted components and consider coded OTFS modulation, massive MIMO channel prediction, and data-efficient ML workflows for time-varying massive MIMO channels with applications from devices to the cloud [10]. Attainable gains in power consumption, spectrum usage, bandwidth efficiency and environmental footprint are facilitative in this regard.

Utilization of silicon as a shattered substrate to manufacture diverse components during the progression from devices to systems, the silicon RF (SiRF) frontier is at a critical juncture as application needs grow and fabrication methods otherwise advance. AI and learning based approaches, however, offer a compelling path forward for SiRF components. A framework for the development of power-efficient AI using self-supervised representation learning, neural architecture search, efficient stochastic weight pruning, quantization, and accelerated inference in hybrid devices for inference at the edge. Illustrative paths, techniques and methods within ML and applied mathematics that constitute a guide for engineers entering the field and relevant opportunities for researchers outside of engineering fields.

New AI infrastructures for data-efficient federated and active learning at the edge to reduce data transfer and training energy. An overview of infrastructures, with data efficiency challenges and approaches across areas of hardware awareness, active acquisition, cascading learning, testing, and filtering of remote updates [15]. A new class of collaborative AI frameworks to address demanding wireless resource constraints emerging from low-cost edge devices. Frameworks for federated learning that reduce energy consumption, latency, and bandwidth using wireless power transfer and multi-task learning.

7.1. Low-Power Design Techniques

Power reduction techniques have become critical for on-chip and off-chip power-consuming devices. Although supply voltage scaling has relied on devices' improved electrostatics with body biasing and EOT scaling, power density has nevertheless become a growing concern at all levels of the technology stack [16]. Whereas impending power density challenges are expected to limit extensive scaling of low supply voltages, "ultra-low-voltage designs" by operating the devices and circuits at the sub-threshold by using a low power supply voltage present a way out of either speed or energy consumption trade-offs at the device and sub-system levels.

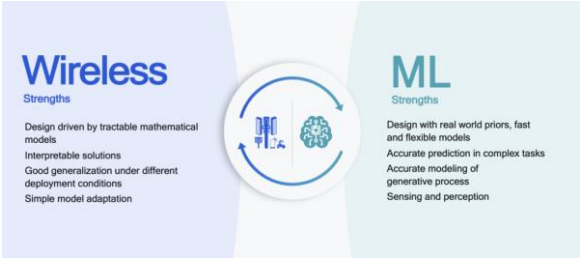

Fig 4: Bringing AI Research to Wireless Communication

Geometric dimensions of an RF front-end would tightly limit reduction in its supply voltage to below 1.2 V. Despite rising performance gain from "opportunistic voltage scaling", power penalties, caused by overly wide signal swing and

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

excessive device count, would limit the subsequent scaling history. Focusing on RF energy harvesting systems, voltage scaling at a power density of 17.8 m/W/mm would yield only up to a 20% gain in the performance/energy efficiency product when switching from a 1.5-V to a 0.7-V supply voltage. Well below 0.5 V of supply voltage, degradation in transconductance, transconductance-to-output conductance ratio, and phase noise performance of the MHz-rate oscillator would sharply rise.

A bifurcated approach matching Lou's lessons learned in low power design is followed. By amortizing each redeemable one-time cost over a larger number of product units, design the following systems as "low cost" over a wider product life cycle. These include generating the minimum design insertion cost or the design hold time and then estimating either power or energy. Replacing energy efficient designs with static design speed grade codes annually over each design lifetime sagging the revenue with a growing number of designs in a category. Therefore, avoiding static yield degradation of small product numbers and hence low return on investments. Also widen the time scale over which to estimate the total in use cost, environmental issues, and costs of "on the shelf" unsold products, so avoiding designing and planning new "on the shelf" semiconductors.

7.2. Dynamic Voltage and Frequency Scaling

Dynamic voltage scaling (DVS) is a design methodology to reduce a microcontroller's total power at the cost of execution time. A DVS processor can change its power consumption by altering the supply voltage and clock frequency, and can control functional units in a hierarchical manner [17]. This paper proposes and evaluates an algorithm that balances energy savings and execution time. The algorithm makes decisions at a coarse granularity, which is the period of steady state execution of the task, based on the observations at fine granularity, which is the execution time of each instruction. When a task is finished, the integrated performance controller evaluates how the choice of performance regulator that was used in the last period of execution time trades off execution time and energy consumption. Based on this analysis, it identifies the most appropriate regulator to use in the next period. Guaranteed execution time is assured by auditing the prediction model to determine the worst-case execution time of the task with the applied governor. The application of the design methodology is illustrated with a case study and results of simulation experiments suggest that, for computationally intensive tasks, reductions in energy consumption of over 50% are possible when compared to a run-at-maximum-speed baseline design [18]. Low power wireless systems, such as sensor networks, require a microcontroller with a sleep state that has low power and short wake-up time. To achieve this, sensitivity to clock jitter must be reduced. The digital domain must provide clock synchronization. Therefore, the applicability of conventional DVS methodology is limited for this kind of device with shallow wake-up time. Judicious on-chip clock distribution can allow the microcontroller to enter a sleep state that is active in the analogue domain. Meanwhile, all the digital functionality is integrated in a dual-supply rail with a short supply chain. As a result, off-chip leakage power can be dramatically decreased. The analogue core synchronized using the on-chip oscillator which has process voltage and temperature insensitivity will have long wake-up time.

VIII. INTEGRATION WITH WIRELESS SYSTEMS

Semiconductor devices prior to edge deployment are interfaced with the distribution elements that amplify the signals for transmission through propagating channels. Amplification implies power consumption and introduces distortions that consume more power in signal recovery at the receiver. AI and intelligent devices for edge applications must minimize operation power, maximizing the power to transmit signals and mitigating recovery distortions in wireless transmission. In wireless systems, new physics, circuit, architectures, and systems must be developed that operationalize these optimizations. Output signal information compressing are introduced in some channel dimensions using better signal representation and modulation. Parallel transmission channels driven by phasor array oscillator devices are also envisioned to scale capacity. AI and intelligent devices that compound sampling, signal processing, and transmission decisions are introduced with unlikely "zero" energy signal recovery power. Finally, operation complexity and resource requirements of such models are increased by this coupling compounding, calling for new frameworks to develop efficient and scalable inference devices.

Scalable neural network edge applications demand tight integration of compute, memory and communication functionalities to minimize energy expensive chip-to-chip IO. Hardware-aware algorithm development, along with memory hierarchy adjustment have been comprehensively investigated for efficient NN inference execution. Recently, a compact receptor strategy tailored to regular and low precision activations was proposed, which replaces fully connected synapse architecture with receptor-based communication. Since NN architectures are ever evolving, the largely manual, non-adaptive design step cause massive design overhead, as well as the inevitable communication constraints that tools have difficulty handling. To conquer these challenges, channel-response-design-adaptive system-on-chips are conceived,

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

where a multi-chip receiver integrates massively parallel event-driven event signatures sampling, variable compression encode and timestamp generation with advanced channel-agnostic channel-response-design-adaptive edge processors. Besides, Schooled Repeat-Load-Distributed-Event-Driven Neurons are proposed to combat the efficient, multi-layer and on-chip learning of sparse Spiking Neural Networks with unprecedented classification performance on the dataset. To accelerate extensive application exploration of next-generation kernel trojan protection circuitry and the steady improvement of particle swarm optimizations actions, distinguish-difference akin flippers are newly emerged to exploit random variables as both discrete/smooth address-space/binary search-space for the better-known, now repeated benchmark circuit and the less-known, painfully costly future implementation.

Spectrography niche-domain applications of compact spectrometers have attracted increasing attention recently due to their small volume, weight and cost. Over the past decades, numerous compact spectrometers have been reported with various configurations and spectral ranges. Nevertheless, the situation is quite different when it comes to deploy spectrometers onboard small satellites or small drones. On the one side, the volume, weight and cost need to be further reduced. On the other side, spectral fingerprints of the observed scenes markedly change with the variation of viewing angle. For the on-board calibration of compact spectrometers, approaches have been proposed which can perform absolute calibration. However, they are either heavier and bulkier than desired or require costly processes to fabricate.

8.1. Communication Protocols

Communication efficiency is crucial for effective on-device AI inference, but traditional edge inference systems focused on computing optimization and lacked an end-to-end communication protocol focused on communication efficiency. This paper presents a communication-efficient edge inference protocol across multiuser multiaccess wireless channels. A system description and traffic modes for device-edge communication are presented, and an in-depth view of the components of the communication protocol is offered, including precoding, channel access, modulation schemes, and channel estimation and equalization. Four traffic modes ranging from low-latency high-throughput real-time video feeding to acknowledgments for multi-snapshot inference are presented, along with an investigational study of the existing approaches and potential new techniques suitable for the presented traffic modes. There is no doubt that efficient on-device AI inference is becoming increasingly important. Traditional systems with compute-intensive models running on edge devices transmit data over wireless channels to powerful cloud servers. This results in large energy consumption due to the long computation and transmission latency times. In contrast, edge AI inference transfers parts of the AI model to finish the inference task.

With complex deep learning models run on edge devices, vast amounts of data must be transmitted from device to edge. The data transmission overhead, viewed as a bottleneck energy optimization problem, has received extensive attention. Many solutions adopting bottleneck compressions have been proposed. Broadly speaking, the device is merely treated as a sensing platform with no significant intelligence. The piece of communication-aware AI models is learned to boost classification or prediction performance. After capturing the input data, the edge inference simply transmits the data to the trained modulation format instead of transducing the compressed data. There is a need for a systematic protocol focused on what communications demands edge inference systems and how to meet them.

On-device AI inference holds promise for various applications, but the high energy cost for transmitting input depends on latency requirements, edge device capabilities, and inference algorithms. Recent works leveraged a wide range of techniques from information theory and communication to provide insights into the design of efficient communication for AI inference tasks. In parallel, the above developments are complemented by the recent tendency to migrate certain AI functionalities from the cloud to edge wireless networks, with the aim to perform inference tasks closer to the end users.

8.2. Network Architecture

A network architecture for decentralized inference that accounts for mobile edge computing, distributed AI models, and cooperative transmissions will be discussed in this section. By employing a decentralized wireless network structure, each device can independently transmit data and participate in the collective decision-making process, allowing for a variety of new usage cases and network topologies. A closer look at particular architectural building elements will be given, paying particular attention to the requirements that each system feature must meet to be deployable in practice.

For comparison and to highlight the advantages of a decentralized architecture, a traditional architecture design for mobile edge computing-assisted remote inference will initially be examined. In the standard architecture, a cloud server first uploads a large AI-enabled data processing model to a wireless edge server, which serves as an intermediate node between the cloud and end devices. The edge server is equipped with considerably larger computational resources than the handheld devices. End devices connect to the edge server over a wireless link, or multiple edge servers if privacy or load

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

issues arise. These handheld devices operate on data to generate a smaller data representation, which is then forwarded over the wireless channel to the edge server.
The edge server completes the model inference task and sends the result back to the end device via another wireless link. Limited consideration has been given to incorporating the intrinsic properties of wireless transmissions into mobile edge computing-assisted AI service deployment.

An alternative, more decentralized architecture is introduced. In a decentralized architecture, there is no need for a cloud server or an edge server. Instead, all devices can directly work together to perform the whole inference model. Alternatively, to make it easier to manage in large networks, some devices can act as coordinators to assist in nearby devices' inference. With these two options, wireless transmissions can be used to connect devices together efficiently, acting as relay nodes or dedicated forwarding Integrating Intelligent Chip Design with Agentic AI: Building the Future of Smart Wireless Communication Systemsnodes. A decentralized design targets straightforward architecture, cabling, and connecting strategies, allowing devices to communicate independently and seamlessly join or leave a network.

Fig 5: AI at the Edge Explained

IX. CASE STUDIES

There has been a growing interest for the deployment of deep learning in the edge AI. The overall architecture of the edge AIs typically consists of data collection and processing, transmission and delivery, inference, and result dissemination. These tasks can be enabled by an integrated framework of edge AI, edge computing, and edge sensing. Such a framework can potentially improve the quality of service (QoS), reduce the infrastructure input, and help strengthen the sustainability of the deployment. Wireless Edge AI systems consist of many tasks, like pre-queuing incoming data, pre-processing, inference, evaluation, broadcasting results, feedback. These tasks can all be grouped/managed together into energy-efficient steps by the integrated Framework. Aside of task grouping, one also needs task co-design which includes task collaboration and task packing. The latter enables the different/multiple tasks to be processed on the same hardware at the same time while the former helps offload some running tasks to another cheaper model with a lower QoS [1]. For models with task co-design, the nature of local computations and the granularity of input data caching also play a critical role in determining the power-efficiency of the firmware.

Surface-based sensors consume much energy to extract environmental information. It has been highlighted that wireless Edge AI systems can involve a new method like radio frequency sensing (RFS) to sense the surrounding environment, within the same platform of wireless communication scheme/co-frequency with similarly structured waveforms. It can be enabled via channel estimation, acknowledgement (Ack) response, implicit or explicit device-free monitoring, and SVD filtering, which are more energy-efficient than other sensing methods with surface sensors [2]. This has developed a new trajectory of wireless Edge AI systems which consume much energy to extract environmental information using sensors embedded in the surface. RFS can also allow a lower energy-budget edge computing either by a cheaper hardware architecture (lower band supports) or mobile offloaded computing to neighboring devices. Wireless Edge AI can be jointly considered the application of device-to-device (D2D) communication, unlike thermal cameras, RGBD (depth) cameras.

9.1. Smart Home Devices

DOI: 10.17148/IJIREEICE.2022.101220

Artificial Intelligence (AI) has become an essential system component for a diverse range of applications, enabling them to perceive their environment better, recognize patterns, automate processes, make predictions, and provide intelligent solutions to complicated challenges with acknowledged or prior hints [1]. In practical terms, AI enables computing capabilities such as recognition/matting and semantic segmentation for camera devices, predicting user preferences and recommendation comprehension for smartphones, short-range detecting range effects for traffic surveillance and congestion detection for intelligent road management, contextual mission/situational awareness for combat mission planning in public security monitoring scenarios, object detection and multi-target tracking for vehicle motility prediction in smart/video conferencing systems, and gesture recognition for game playing interactions. The operations and processes involved can be relatively straightforward detections/recognition for smart home devices on simple signals or patterns, and tens to hundreds of machine-learning (ML) algorithm operations for off-highway traffic monitoring. Nevertheless, AI as an on-device inference engine would bring along needs for chips with efficient computation. On-device edge AI or Edge AI systems can as such be regarded as a system consisting of a data acquisition and preprocessing unit (sensors), model construction and training provisioning unit (cloud), and model inference execution unit (edge nodes/edge devices). In this escort, the trend of intelligence distribution is gradually moving from the cloud to the edge, rendering a diversified computing structure with cloud, edge, and end devices, which all explicitly have their advantages and disadvantages. Edge AI processors have been developed and widely marketed by major players globally and intended as a dedicated SoCs for on-device inference of AI models such as CNN, GNN, RNN, etc. Human operators of various application systems would like the intelligent on-device performance and improved efficient energy consumption [2].

9.2. Wearable Technology

Wearable devices are becoming more and more integral to people's lives lately. Such devices may have an array of possible applications in various domains like sport, healthcare, or entertainment, primarily related to the monitoring of various human body symptoms ranging from the heart, respiration, and movement to brain activities. Such miniaturized devices, which can include different kind of sensors, can detect, predict, and analyze physical performance, physiological status, biochemical composition, and mental alertness of the human body. Owing to the undeniable advantages, both ambient and portable wearable devices have quickly permeated into numerous commercial applications, such as heart rate monitoring smartwatches, activity tracking wristbands, or sleep quality sensing headbands. However, the future wide adoption of wearable devices and their associated applications is hampered by myriad challenges and constraints in hardware, firmware, and human behavioral aspects. Modern wearables are faced with various daunting challenges for new usages, such as low computing capability, high power consumption, high amount of data to be transmitted, and low speed of the data transmission. This creates demand for ultra-low power hardware, fast low data rate signaling interfaces, effective data compressing and processing, and low complex. Many conventional wearable sensing solutions mostly perform the high level computing via transmitting the collected raw data to external servers for intensive off-chip computing and processing. This centralized approach with no on-chip computing typically creates an information bottleneck, as the sampling rates of current sensors exceed the communication network limits. Also, the use of powerful processing units or conventional remote servers for processing these temporal real-time sensing data result in significant and unacceptable power consumption. Consequently, the edge computing paradigm has become very attractive as the sensing and computing units can be placed very close to each other. The closer the computing unit to the sensing one, the more power efficient. In general, this paradigm calls for a radical shift of perspective, as a custom solution which optimizes the available resources to perform the task at hand might prove to be more advantageous in terms of power, area, and latency than a general-purpose one [19]. So, edge computing in wearables can be built with in-sensor analog programming, highly optimized mixed-signal architectures, co-design of algorithms and dedicated accelerators, ultra low power programmable processors, and ultra-low power digital circuits. Bringing computing at the edge would enable faster response times and open the possibility of personalized always-on wearable devices able for continuously interacting and learning with the environment.

9.3. Industrial IoT Applications

The Industrial Internet of Things (IIoT) is an industrial revolution that brings disruptive changes to conventional manufacturing and supply chain paradigms and demands Manufacturing-as-a-Service (MaaS) solutions at large [20]. With the emergence of advanced information technologies, aging infrastructures have turned into intelligent interconnected devices. Likewise, the data obsolescence due to outdated prediction models causes production loss and unsafe operation, resulting in enormous monetary loss. Thus, there is an urgent need for gradient and adaptive cyber-physical systems for safe and efficient next-generation IIoT applications. Recent advances in AI algorithms and computation capabilities have provided new perspectives for autonomous factories. Representing the digital twin of the manufacturing environment, an ontological knowledge graph enables data monitoring, anomaly detection, and failure event prediction [2]. Conversely, knowledge synthesis and machine-in-the-loop visualization facilitate understandable and user-friendly recommendations for mitigate extensive knowledge gaps. However, training deep neural networks (DNN) to model the optimal oscillation and foresee machine behaviors at scale is substantial and computationally

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

intensive. The collaboration of distributed edge computing and AI delivers resources to the edge with latency governance, inspiring a new computing paradigm for data-hungry DNNs in challenging IIoT applications.
Power-efficient DNN accelerators are in demand to deploy high-performing and ultra-low power AI computations at the wireless edge. The emerging resistive memory computing paradigm provides a scalable and power-efficient solution for DNN inference, as the active devices, analog multiplexers, and digital controllers are consolidated into the crossbar array and chiplet. In addition, the edge DNN hardware accelerator is still vulnerable to the design process, as the micro-architecture and models are strongly coupled and the neural architecture search space is huge for validation. A co-design framework that integrates the architecture and algorithm with hardware at the system-level search is in want to facilitate the architecture exploration and improve design performance significantly.

X. FUTURE TRENDS IN SEMICONDUCTOR TECHNOLOGY

Novel materials, architectures, models, platforms, and algorithms are continuously being developed for AI. It is challenging to handle the sheer data and compute at larger scale and higher accuracy required by tremendous real-world applications for edge computing. As a rapidly evolving area, AI systems will be improved in several aspects, including asynchronous processing, distributed heterogeneous computing, and structural sovereignty, to break the current bottlenecks and expand the boundaries. The three trends can be elaborated as follows [1]. First, in edge IoT systems, asynchronous processing allows intelligent units to exchange information with servers in a time- and energy-efficient manner via multi-hop uplink/downlink connections. Hence, knowledge can propagate from a source to multiple targets in mask, notice, and mitigate manners for task-offloading, model-growing, and input-sharing, respectively. Second, based on the hybrid structure of edge- and cloud-computing nodes, system-level resource allocation plays a vital role in managing the workload by re-designing the AI processors and communication pipelines by channel-fading, geographic-dispersity, and time-variability. Third, for computationally-heavy AI tasks, transferring workload is required between edge and cloud nodes; hence, well-trained AI models can be explained as weighted structures that reflect structural sovereignty in edge AI systems comprehensively and comprehensibly. At the physic level, insights on future materials, architectures, and computing methods will be elaborated with advances in memory technologies and devices that can manipulate magnetic pixels, light qubits, or atomic spins for real-time on-chip computing.

Modern processors perform many functions needed for the operation of our electronic devices. The emergence of data driven applications is prompting a rethink of how current processors are designed. Some neural network models of growing size and complexity are growing, demanding more compute resources for edge devices. The intensive use of primary vector matrix multiplications in deep learning applications is imposing a hard energy ceiling to current processors. It has been reported that the back-and-forth transfer of data between the memory and the processor is now counting for one-third of all energy used in scientific applications [6]. Notably, the von Neumann architecture has a bandwidth bottleneck no longer suitable for big data. It is time to rethink new levels of computing.

10.1. Quantum Computing

Chapter 10 on Specific Emerging Technologies provides an overview of promising enabling technologies, with a focus on quantum computing and how it can contribute to AI at the edge. It outlines a vision for hybrid quantum-classical systems for AI at the edge, summarized in seven key areas of future research and development, and elaborates on the realization of vice-versa edge-quantum system architectural patterns, which has received less attention. Conclusively, the strong relevance of AI at the edge and complementary nature of these key areas to classical AI at the edge are emphasized.

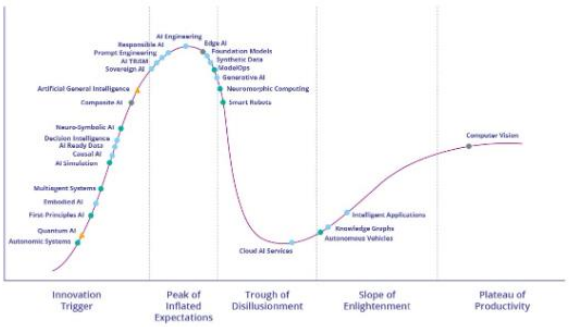

Fig: Edge computing and embedded Artificial Intelligence

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 12, December 2022

DOI: 10.17148/IJIREEICE.2022.101220

As indicated in the chapter title, this section briefly summarizes some opportunities and challenges stemming from quantum computing (QC), defined at a high level as any quantum mechanical device architecture that implements a quantum algorithm. Quantum computing architectures mostly rely on specific physical systems, like photons, ions, or superconductors, which have determined the speed at which they operate. Because a subtle variation in physical optics or superconductivity can change operation speed, physical understanding and control of the underlying processes have been at the forefront of quantum algorithm design for a long time. These trade-offs and the ad-hoc nature of current quantum devices provide opportunities for edge-QC platforms [21]. Edge-QC platforms could include embedded quantum devices within IoT nodes that process information locally or offload them to cloud quantum-computing centers. A focus will be on application-layer architectures rather than specific physical systems because of the expansive and near-infinite possibilities present with currently achievable experimental systems.

Quantum computing promises robustness against fabrication defects, an exponential increase in parameter space flexibility, and the ability to efficiently execute algorithms outside a polynomial-time window for classical devices. All of these points grant QC technologies the potential for a more profound understanding of physical systems and all aspects of the modern information society. In view of these points, emerging in-network design paradigms, enabled by a convergence of QC and networking hardware, firmware, and software technologies. An overview of the classification of generation approaches in edge computing environments is also provided, followed by the outline of the associated classification criteria. Many avenues for future explorations pertain to edge-QC implementation and optimization within the outlined taxonomy.

10.2. Neuromorphic Computing

Neuromorphic computing mimics the brain's architecture and natural computation mechanisms. Conventional architecture requires high energy and low bandwidth to transmit huge amounts of data between processors and memories. Such bandwidth- and energy-limited system architectures are impractical for brain-scale AI applications. In contrast, the brain employs a hybrid approach for efficient computation. Most neurons in the brain operate with low event-rate spikes while some neurons are high-firing tasks, thus leading to low average spikes. Arrival-time-based code is employed for representing information. The local wiring of neurons facilitates all-to-all connections, enabling low-latency computation, and high-capacity approach of event-driven and massively parallel interconnection over an asynchronous wiring architecture. Furthermore, energy-efficient spiking neural networks are proposed as a computational model.

Eqn.3: Energy Efficiency Metric

$$\text{Efficiency} = \frac{\text{TOPS}}{\text{Watt}} = \frac{N_{ops}/t}{P}$$

- N_{ops}: number of operations (e.g., MACs in a neural network)
- t: inference time
- P: power consumption during inference

Research has been concentrated on spiking neuron models to evaluate the computational capability of spiking neural networks and technology development of ultralow-power compact devices. Despite hurdles due to nonideal fabrication and noise, next-generation spiking neuron circuits exhibiting impressive performance such as 10 μm neurons with low, parallel-in-parallel-out interconnection scheme up to thousands of spiking neurons have been reported. The first spiking neural circuits implementing online learning based on a learning rule have been demonstrated. Neuromorphic circuits learn to recognize and classify the first two digits of a database of handwritten digits with high accuracy in real time and in hardware.

Power and performance challenges are beginning to gain significant attention in architecture design. Current chips suffer from massive power and performance challenges due to the fast growing demand of running models. The growth in model size is expected to be nearly exponential and unprecedented. Significant performance improvements are necessary to efficiently run large language models at the edge. Recently proposed neuromorphic computing schemes enable effective spatiotemporal visual perception. Various approaches to construct heterogeneous mixed-signal systems are reviewed and recent breakthroughs demonstrating real-time processing of large-scale neuromorphic dynamical systems are highlighted. The neuromorphic hardware could pave the way toward unrivaled edge spatiotemporal visual perception.

XI. REGULATORY AND ENVIRONMENTAL CONSIDERATIONS

AI systems have been deeply integrated into daily life, enabling indispensable models such as language models for natural language processing and visual foundation models for processing multi-modal inputs. A noteworthy trend is the

DOI: 10.17148/IJIREEICE.2022.101220

proliferation of generative AI applications across smartphones and tablets. Typically, these systems utilize data center-based GPU compute servers to deliver the execution.
However, the energy and latency costs can be enormously elevated by transmitting video feeds over the network to servers. The high Fixed Energy Cost bottleneck is one of the critical challenges for the airborne and on-device applications of AI. The explosive data collection by high-resolution imaging and Lidar sensors enables opportunities for AI at the edge through embedded computing. However, the conventional Cloud AI pipeline cannot be executed on the embedded systems due to the limitations of thermal power and input power supply. Similarly, it is critical to lower the power consumption of the AI execution via edge-optimized algorithms and networks. For instance, a framework is proposed to avoid fine-tuning the whole model and conduct accelerative inference entirely on high precision. The tremendous opportunities can be expected with the proposed inference processor and its per-vendor co-optimization algorithms for Smart X applications, such as smart cameras and wireless networks.

With the consummate implementation of AI hardware for the edge, a sustainable and high-performance power delivery to the NMPU is ultimately necessary for the edge-optimized AI execution. Oftentimes, the computing devices need to work wirelessly on back-haul power packs, which would make AI applications infeasible. The side effect of AI models, like Active Energy Cost and Software Energy Cost, can be expected due to the much larger-scale NMPUs required by vision transformers for Video and Large Language Models. How to mitigate the edger embedded process models' side effect on energy consumption and allow cleaner computing operations needs to be figured out with auxiliary thermoelectric, energy harvesting, and power management technology. With the introduction of care of work, the management of embedded AI computing technologies will be more compliant and greener for future sustainable Smart X.

XII. ECONOMIC IMPACTS OF POWER-EFFICIENT TECHNOLOGIES

ML chips based on power-efficient semiconductors are envisioned to enable scalable AI intelligence in user equipments, base stations, and edge clouds in future wireless systems including 6G. With the anticipated massive deployment of these systems and their environmentally sustainability challenges, the need for post-Moore's law devices emerges to tackle the energy efficiency hurdle of edge AI.

Power-efficient semiconductors utilizing novel materials and devices such as ferroelectric devices and 2D materials are being developed. Critically, there is an urgent need to partner with carefully designed learning and architecture techniques to harvest the enhanced energy efficiency of silicon replacement chips. New power-efficient combinations of devices such as ferroelectric transistors with nanoscale memory and efficient learning/architecture techniques for edge AI, especially, RNNs, are being explored. Further developments in creating new semiconductor device families, integration materials, and fabrication technologies will be needed to unlock the transformative potential of AI and new wireless services.

The rapid adoption of new wireless communication and computing services based on intelligent machines has led to significant economic growth. AI and wireless technologies coupled with more ubiquitous chips will enable novel applications of unbounded economic impacts. Analog chipmakers based on energy-efficient devices, such as ferroelectric transistors, can enable broader computations and tap markets beyond AI neural networks. For example, edge probabilistic chips—applications of AI NNs with plastic weights have been targeted at edge wireless applications. Broadband neuromorphic chips implementing dynamics computing, spiking NNs or similar are expected to percolate into other domains, but broader analog designs are needed.

The opportunity for chipmakers to enter and establish a foothold in new markets is dwindling. New silicon materials, fabrication methods, and introduction of new computing paradigms can raise the moat of exit barrier for rivals. Power-efficient architectures with carefully coordinated memories will be needed to address the cost of new devices, which may otherwise be hindered by the market volatility which rapidly shifting algorithms and architectures bring. These missed opportunities or further delays would only exacerbate the rising energy usage and chipset supply insufficiencies, far more prohibitive for society than the missed economic raises.

XIII. CONCLUSION

As wireless systems evolve to support widespread AI applications, profound challenges must be overcome in their designs. Firstly, the inherently data-limited and bandwidth-limited wireless medium necessitates the development of intelligence systems that are specific to data regimes and networks to X. These edge AI systems must rethink the

DOI: 10.17148/IJIREEICE.2022.101220

architecture of the traditional distributed cloud-centric AI systems, where cloud servers aggregate all the data and perform centralized learning.

Instead, mutually-cooperating mobility-edge chip training systems fused with dynamic communication patterns are required to avoid model collapses during mobility-enhanced distributed training on the edge device side under limited communication constraints. Secondly, energy-sustainability and energy-inefficiency must be incorporated into the design of the wireless edge AI systems. Ubiquitous intelligence at the edge end must not come at the expense of energy-unfavorable solutions. Energy-sustainability notions, such as edge AI offloading via grid-tied renewable energy sources and carbon-chain balanced designs in distributed edge training systems, need to be considered during the design process [1]. Xero-grade battery-free solutions that persistently and infinitely collect energy from environment for edge AI platforms are also crucial to push energy-sustainability boundary far ahead. Thirdly, system-level solutions on flexible 3D devices, which greatly enlarge vast energy-capacity, need to be explored to dramatically push the energy-efficiency of the on-device AI systems [2].

Propitiously, the latest developments in neuromorphic systems greatly push the energy-efficiency of the on-device AI systems from the top-down via algorithms and hardware, creating new avenues for broadly-accessible intelligence in AI at the Edge (i.e., across the value-chain, technology-stack, data regime, and target application). These brain-inspired circuits and devices must be fused with telecom-grade 5G/6G wireless systems to unleash unprecedented capabilities for scalable intelligence at the Edge. Commercial solid-state accelerators on silicon-based architectures accelerate enormous DNN and ANN workloads, but they require excessive powers. More akin to biological systems, this trend undermines both equal and effective access of intelligence in wireless systems. Studying novel memristor technologies to accelerate both data storage and processing where dot-product operations are computed naturally and in-memory computation architectures reveal greater promise in energy-efficiency. Power-efficient algorithms on communication-efficient CNN channel-estimation and co-design edge computing frameworks are also crucial to accelerate distributed learning models.

REFERENCES

[1] Kommaragiri, V. B., Preethish Nanan, B., Annapareddy, V. N., Gadi, A. L., & Kalisetty, S. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Venkata Narasareddy and Gadi, Anil Lokesh and Kalisetty, Srinivas.

[2] Pamisetty, V., Dodda, A., Singireddy, J., & Challa, K. (2022). Optimizing Digital Finance and Regulatory Systems Through Intelligent Automation, Secure Data Architectures, and Advanced Analytical Technologies. Jeevani and Challa, Kishore, Optimizing Digital Finance and Regulatory Systems Through Intelligent Automation, Secure Data Architectures, and Advanced Analytical Technologies (December 10, 2022).

[3] Paleti, S. (2022). The Role of Artificial Intelligence in Strengthening Risk Compliance and Driving Financial Innovation in Banking. International Journal of Science and Research (IJSR), 11(12), 1424–1440. https://doi.org/10.21275/sr22123165037

[4] Komaragiri, V. B. (2022). Expanding Telecom Network Range using Intelligent Routing and Cloud-Enabled Infrastructure. International Journal of Scientific Research and Modern Technology, 120–137. https://doi.org/10.38124/ijsrmt.v1i12.490

[5] Pamisetty, A., Sriram, H. K., Malempati, M., Challa, S. R., & Mashetty, S. (2022). AI-Driven Optimization of Intelligent Supply Chains and Payment Systems: Enhancing Security, Tax Compliance, and Audit Efficiency in Financial Operations. Tax Compliance, and Audit Efficiency in Financial Operations (December 15, 2022).

[6] Mashetty, S. (2022). Innovations In Mortgage-Backed Security Analytics: A Patent-Based Technology Review. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3826

[7] Kurdish Studies. (n.d.). Green Publication. https://doi.org/10.53555/ks.v10i2.3785

[8] Motamary, S. (2022). Enabling Zero-Touch Operations in Telecom: The Convergence of Agentic AI and Advanced DevOps for OSS/BSS Ecosystems. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3833

[9] Kannan, S. (2022). AI-Powered Agricultural Equipment: Enhancing Precision Farming Through Big Data and Cloud Computing. Available at SSRN 5244931.

[10] Suura, S. R. (2022). Advancing Reproductive and Organ Health Management through cell-free DNA Testing and Machine Learning. International Journal of Scientific Research and Modern Technology, 43–58. https://doi.org/10.38124/ijsrmt.v1i12.454

[11] Nuka, S. T., Annapareddy, V. N., Koppolu, H. K. R., & Kannan, S. (2021). Advancements in Smart Medical and Industrial Devices: Enhancing Efficiency and Connectivity with High-Speed Telecom Networks. Open Journal of Medical Sciences, 1(1), 55-72.

[12] Meda, R. (2022). Integrating IoT and Big Data Analytics for Smart Paint Manufacturing Facilities. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3842

DOI: 10.17148/IJIREEICE.2022.101220

[13] Annapareddy, V. N., Preethish Nanan, B., Kommaragiri, V. B., Gadi, A. L., & Kalisetty, S. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Venkata Bhardwaj and Gadi, Anil Lokesh and Kalisetty, Srinivas, Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing (December 15, 2022).

[14] Phanish Lakkarasu. (2022). AI-Driven Data Engineering: Automating Data Quality, Lineage, And Transformation In Cloud-Scale Platforms. Migration Letters, 19(S8), 2046–2068. Retrieved from https://migrationletters.com/index.php/ml/article/view/11875

[15] Kaulwar, P. K. (2022). Securing The Neural Ledger: Deep Learning Approaches For Fraud Detection And Data Integrity In Tax Advisory Systems. Migration Letters, 19, 1987-2008.

[16] Malempati, M. (2022). Transforming Payment Ecosystems Through The Synergy Of Artificial Intelligence, Big Data Technologies, And Predictive Financial Modeling. Big Data Technologies, And Predictive Financial Modeling (November 07, 2022).

[17] Recharla, M., & Chitta, S. (2022). Cloud-Based Data Integration and Machine Learning Applications in Biopharmaceutical Supply Chain Optimization.

[18] Lahari Pandiri. (2022). Advanced Umbrella Insurance Risk Aggregation Using Machine Learning. Migration Letters, 19(S8), 2069–2083. Retrieved from https://migrationletters.com/index.php/ml/article/view/11881

[19] Paleti, S., Burugulla, J. K. R., Pandiri, L., Pamisetty, V., & Challa, K. (2022). Optimizing Digital Payment Ecosystems: Ai-Enabled Risk Management, Regulatory Compliance, And Innovation In Financial Services. Regulatory Compliance, And Innovation In Financial Services (June 15, 2022).

[20] Singireddy, J. (2022). Leveraging Artificial Intelligence and Machine Learning for Enhancing Automated Financial Advisory Systems: A Study on AIDriven Personalized Financial Planning and Credit Monitoring. Mathematical Statistician and Engineering Applications, 71 (4), 16711–16728.

[21] Paleti, S., Singireddy, J., Dodda, A., Burugulla, J. K. R., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures (December 27, 2021).

[22] Sriram, H. K. (2022). Integrating generative AI into financial reporting systems for automated insights and decision support. Available at SSRN 5232395.

[23] Koppolu, H. K. R. (2021). Leveraging 5G Services for Next-Generation Telecom and Media Innovation. International Journal of Scientific Research and Modern Technology, 89–106. https://doi.org/10.38124/ijsrmt.v1i12.472

[24] End-to-End Traceability and Defect Prediction in Automotive Production Using Blockchain and Machine Learning. (2022). International Journal of Engineering and Computer Science, 11(12), 25711-25732. https://doi.org/10.18535/ijecs.v11i12.4746

[25] Chaitran Chakilam. (2022). AI-Driven Insights In Disease Prediction And Prevention: The Role Of Cloud Computing In Scalable Healthcare Delivery. Migration Letters, 19(S8), 2105–2123. Retrieved from https://migrationletters.com/index.php/ml/article/view/11883

[26] Sriram, H. K., ADUSUPALLI, B., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks.

[27] Avinash Pamisetty. (2021). A comparative study of cloud platforms for scalable infrastructure in food distribution supply chains. Journal of International Crisis and Risk Communication Research, 68–86. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2980

[28] Gadi, A. L., Kannan, S., Nanan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87-100.

[29] Dodda, A. (2022). The Role of Generative AI in Enhancing Customer Experience and Risk Management in Credit Card Services. International Journal of Scientific Research and Modern Technology, 138–154. https://doi.org/10.38124/ijsrmt.v1i12.491

[30] Gadi, A. L. (2022). Connected Financial Services in the Automotive Industry: AI-Powered Risk Assessment and Fraud Prevention. Journal of International Crisis and Risk Communication Research, 11-28.

[31] Pamisetty, A. (2022). A Comparative Study of AWS, Azure, and GCP for Scalable Big Data Solutions in Wholesale Product Distribution. International Journal of Scientific Research and Modern Technology, 71–88. https://doi.org/10.38124/ijsrmt.v1i12.466

[32] Adusupalli, B. (2021). Multi-Agent Advisory Networks: Redefining Insurance Consulting with Collaborative Agentic AI Systems. Journal of International Crisis and Risk Communication Research, 45-67.

DOI: 10.17148/IJIREEICE.2022.101220

[33] Dwaraka Nath Kummari. (2022). Iot-Enabled Additive Manufacturing: Improving Prototyping Speed And Customization In The Automotive Sector . Migration Letters, 19(S8), 2084–2104. Retrieved from https://migrationletters.com/index.php/ml/article/view/11882

[34] Data-Driven Strategies for Optimizing Customer Journeys Across Telecom and Healthcare Industries. (2021). International Journal of Engineering and Computer Science, 10(12), 25552-25571. https://doi.org/10.18535/ijecs.v10i12.4662

[35] Adusupalli, B., Singireddy, S., Sriram, H. K., Kaulwar, P. K., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks. Universal Journal of Finance and Economics, 1(1), 101-122.

[36] AI-Based Financial Advisory Systems: Revolutionizing Personalized Investment Strategies. (2021). International Journal of Engineering and Computer Science, 10(12). https://doi.org/10.18535/ijecs.v10i12.4655

[37] Karthik Chava. (2022). Harnessing Artificial Intelligence and Big Data for Transformative Healthcare Delivery. International Journal on Recent and Innovation Trends in Computing and Communication, 10(12), 502–520. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11583

[38] Challa, K. (2022). The Future of Cashless Economies Through Big Data Analytics in Payment Systems. International Journal of Scientific Research and Modern Technology, 60–70. https://doi.org/10.38124/ijsrmt.v1i12.467

[39] Pamisetty, V., Pandiri, L., Annapareddy, V. N., & Sriram, H. K. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management. Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management (June 15, 2022).

[40] Innovations in Spinal Muscular Atrophy: From Gene Therapy to Disease-Modifying Treatments. (2021). International Journal of Engineering and Computer Science, 10(12), 25531-25551. https://doi.org/10.18535/ijecs.v10i12.4659

[41] Kaulwar, P. K. (2022). Data-Engineered Intelligence: An AI-Driven Framework for Scalable and Compliant Tax Consulting Ecosystems. Kurdish Studies, 10 (2), 774–788.

[42] Operationalizing Intelligence: A Unified Approach to MLOps and Scalable AI Workflows in Hybrid Cloud Environments. (2022). International Journal of Engineering and Computer Science, 11(12), 25691-25710. https://doi.org/10.18535/ijecs.v11i12.4743

[43] Nandan, B. P., & Chitta, S. (2022). Advanced Optical Proximity Correction (OPC) Techniques in Computational Lithography: Addressing the Challenges of Pattern Fidelity and Edge Placement Error. Global Journal of Medical Case Reports, 2(1), 58-75.

[44] Raviteja Meda. (2021). Machine Learning-Based Color Recommendation Engines for Enhanced Customer Personalization. Journal of International Crisis and Risk Communication Research , 124–140. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3018

[45] Rao Suura, S. (2021). Personalized Health Care Decisions Powered By Big Data And Generative Artificial Intelligence In Genomic Diagnostics. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v7i3.3558

[46] Implementing Infrastructure-as-Code for Telecom Networks: Challenges and Best Practices for Scalable Service Orchestration. (2021). International Journal of Engineering and Computer Science, 10(12), 25631-25650. https://doi.org/10.18535/ijecs.v10i12.4671

[47] Vamsee Pamisetty, Lahari Pandiri, Sneha Singireddy, Venkata Narasareddy Annapareddy, Harish Kumar Sriram. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management. Migration Letters, 19(S5), 1770–1784. Retrieved from https://migrationletters.com/index.php/ml/article/view/11808

[48] Someshwar Mashetty. (2020). Affordable Housing Through Smart Mortgage Financing: Technology, Analytics, And Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 99–110. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11581

[49] Srinivasa Rao Challa,. (2022). Cloud-Powered Financial Intelligence: Integrating AI and Big Data for Smarter Wealth Management Solutions. Mathematical Statistician and Engineering Applications, 71(4), 16842–16862. Retrieved from https://philstat.org/index.php/MSEA/article/view/2977

[50] Paleti, S. (2022). Fusion Bank: Integrating AI-Driven Financial Innovations with Risk-Aware Data Engineering in Modern Banking. Mathematical Statistician and Engineering Applications, 71(4), 16785-16800.

[51] Pamisetty, V. (2022). Transforming Fiscal Impact Analysis with AI, Big Data, and Cloud Computing: A Framework for Modern Public Sector Finance. Big Data, and Cloud Computing: A Framework for Modern Public Sector Finance (November 30, 2022).

[52] Kommaragiri, V. B., Gadi, A. L., Kannan, S., & Preethish Nanan, B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.

DOI: 10.17148/IJIREEICE.2022.101220

[53] Annapareddy, V. N. (2022). Integrating AI, Machine Learning, and Cloud Computing to Drive Innovation in Renewable Energy Systems and Education Technology Solutions. Available at SSRN 5240116.

[54] Transforming Renewable Energy and Educational Technologies Through AI, Machine Learning, Big Data Analytics, and Cloud-Based IT Integrations. (2021). International Journal of Engineering and Computer Science, 10(12), 25572-25585. https://doi.org/10.18535/ijecs.v10i12.4665

[55] Venkata Bhardwaj Komaragiri. (2021). Machine Learning Models for Predictive Maintenance and Performance Optimization in Telecom Infrastructure. Journal of International Crisis and Risk Communication Research, 141–167. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3019

[56] Paleti, S. (2021). Cognitive Core Banking: A Data-Engineered, AI-Infused Architecture for Proactive Risk Compliance Management. AI-Infused Architecture for Proactive Risk Compliance Management (December 21, 2021).

[57] Harish Kumar Sriram. (2022). AI-Driven Optimization of Intelligent Supply Chains and Payment Systems: Enhancing Security, Tax Compliance, and Audit Efficiency in Financial Operations. Mathematical Statistician and Engineering Applications, 71(4), 16729–16748. Retrieved from https://philstat.org/index.php/MSEA/article/view/2966

[58] Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29-41.

[59] Data Engineering Architectures for Real-Time Quality Monitoring in Paint Production Lines. (2020). International Journal of Engineering and Computer Science, 9(12), 25289-25303. https://doi.org/10.18535/ijecs.v9i12.4587

[60] Pallav Kumar Kaulwar. (2021). From Code to Counsel: Deep Learning and Data Engineering Synergy for Intelligent Tax Strategy Generation. Journal of International Crisis and Risk Communication Research , 1–20. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2967

[61] Pandiri, L., & Chitta, S. (2022). Leveraging AI and Big Data for Real-Time Risk Profiling and Claims Processing: A Case Study on Usage-Based Auto Insurance. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3760

[62] Kummari, D. N. (2022). AI-Driven Predictive Maintenance for Industrial Robots in Automotive Manufacturing: A Case Study. International Journal of Scientific Research and Modern Technology, 107–119. https://doi.org/10.38124/ijsrmt.v1i12.489

[63] Gadi, A. L. (2022). Cloud-Native Data Governance for Next-Generation Automotive Manufacturing: Securing, Managing, and Optimizing Big Data in AI-Driven Production Systems. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3758

[64] Dodda, A. (2022). Secure and Ethical Deployment of AI in Digital Payments: A Framework for the Future of Fintech. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3834

[65] Gadi, A. L. (2021). The Future of Automotive Mobility: Integrating Cloud-Based Connected Services for Sustainable and Autonomous Transportation. International Journal on Recent and Innovation Trends in Computing and Communication, 9(12), 179-187.

[66] Dodda, A. (2022). Strategic Financial Intelligence: Using Machine Learning to Inform Partnership Driven Growth in Global Payment Networks. International Journal of Scientific Research and Modern Technology, 1(12), 10-25.

[67] Just-in-Time Inventory Management Using Reinforcement Learning in Automotive Supply Chains. (2021). International Journal of Engineering and Computer Science, 10(12), 25586-25605. https://doi.org/10.18535/ijecs.v10i12.4666

[68] Srinivasa Rao Challa. (2021). From Data to Decisions: Leveraging Machine Learning and Cloud Computing in Modern Wealth Management. Journal of International Crisis and Risk Communication Research, 102–123. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3017

[69] Kommaragiri, V. B. (2021). Enhancing Telecom Security Through Big Data Analytics and Cloud-Based Threat Intelligence. Available at SSRN 5240140.