IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111258

Semiconductor Innovation for Edge AI: Enabling Ultra-Low Latency in Next-Gen Wireless Networks

Goutham Kumar Sheelam

Programmer Analyst, Staff, ORCID ID: 0009-0004-1031-3710

Abstract: Artificial Intelligence (AI) is experiencing a paradigm shift. AI began with large, computationally and energy-intensive server farms doing training and then moved to the cloud to distribute inference. But in the next generation, the compute resources needed for real-time, low-power, localized analysis and decision-making is shifting to the edge. This is enabled by a wide variety of new semiconductor technologies including domain-specific hardware accelerators, custom chips built for efficiency and novel architectures leveraging Many Integrated Core, chiplets, neural processing units, and in-memory compute architectures. The emergent capabilities make possible new services and use cases in the wireless networks, the devices in relation to the networks, and the applications that run on the devices. However, these use cases introduce exponentially higher demands for throughput, latency, training, and power efficiency compared to previous generations. To be effective, semiconductors must support energy efficient low latency inference processing, decentralization of training workloads, and proper device and network integration that recognize and have solutions for the device to data, data to model, and model to data transitions.

This document presents several focus topics spanning wireless networks, chip technology for devices, systems and applications for edge AI. These components themselves are but a subset of the next generation challenges, but they provide a good roadmap to drive focus and prioritization. Thematically as AI transitions to the edge, what does that mean in terms of AI services and workloads? What technological capabilities and requirements drive progress in use cases? What devices need be created? And how do we innovate and create solutions in a cooperative way at the pace required to realize these progress and capabilities?

Keywords : Global semiconductor market, growth forecast, chip industry expansion, integrated circuits, AI chips, 5G semiconductors, automotive electronics, consumer electronics, industrial IoT, advanced packaging, silicon wafer demand, fabrication capacity, semiconductor manufacturing, chip shortage recovery, global supply chain, foundry services, semiconductor innovation, R&D investment, chip exports, emerging markets, leading semiconductor companies, technological advancement, market dynamics, semiconductor trends, microelectronics, global demand drivers.

I. INTRODUCTION

Smart devices powered by Artificial Intelligence (AI) are revolutionizing the way we interact with technology. These devices use AI to continuously learn and adapt to our needs, enabling us to offload a variety of complex tasks. From the numerous IoT devices in our homes to autonomous vehicles, we are surrounded by smart devices. In the near future, we will depend on them increasingly for safety, security, and comfort. All these devices rely on rapid exchange of data, followed by ultra-low latency inference for their operation. One of the important functions of a self-driving car is to detect any potential obstacles on the roads, such as pedestrians, moving objects, or any other vehicles. This function needs to happen in real-time so that the necessary actions can be taken to avoid accidents.

Recent advancements in AI hardware architecture, availability of public AI models, and the ability to fine-tune models are facilitating the deployment of AI software on the Edge. AI was first introduced in the 1950s. However, its renaissance actually began in 2012 with the advent and implementation of Deep Learning. The extent of research and advances in the field, along with the rising interest in its practical applications due to large-scale public datasets and the rapid improvements of computing hardware and algorithms, is beyond what is possible in any other field. However, despite the rapid progress in recent years, the next stage of growth will fracture into multiple versions, with semiconductor technology for AI becoming a key differentiator. Edge devices are expected to perform ultra-low latency AI inference in real-time based on models that have been either trained in the Cloud or on the Edge.

To sustain the intended momentum for performance and energy efficiency, novel Edge-specific hardware accelerators are needed so that state-of-the-art large models can be delivered to the Edge device. In addition, novel computer system architecture, including the software stack, is needed to support those accelerators and the intended applications.

II. OVERVIEW OF EDGE AI

Edge AI refers to the distributed form of Artificial Intelligence that brings the intelligence of AI applications close to the end-target. It enables the opportunity to train or infer intelligent capabilities at the point of data collection or at a nearby location as opposed to sending the data to the cloud for processing. Achieving utmost responsiveness with ultra-low latency on a critical layer in the data pipeline is pivotal for new-age enterprise solutions, as well as consumer applications. The emergence of edge AI is driven by the increasing productization of machine learning and boosts demand for low-latency applications that require smartness at the end-device or at an edge-box appliance. Applications such as smart cameras, auditory scene analysis, location-based services, product recommendation, and security analytics are already being deployed into enterprise or consumer spaces. Delivering real-time insight from a slew of data at the edge has become a necessity. Besides ultra-low latency insight, the amount of data being stored in the cloud can be colossal, and hence its processing can often be cumbersome. Using algorithms with localized processing can assist in avoiding significant cloud storage requirements. Since edge nodes can analyze data in real-time and send pertinent insights to the cloud in a summarized form, edge inference can further lessen the load of post-cloud processing.

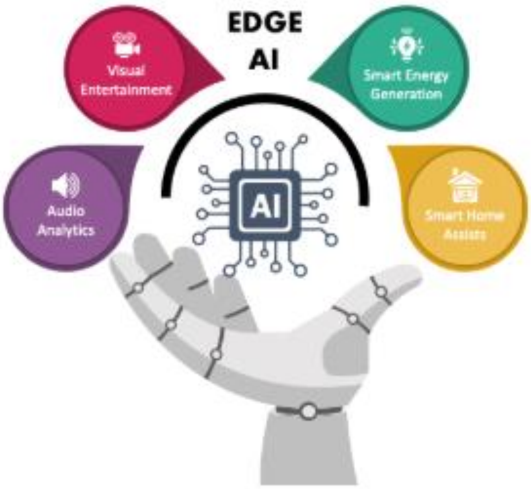


Fig 1: Edge AI (Artificial Intelligence)

In recent years, major players in the IT industry have placed investments as well as strategic designs into pushing AI capabilities at the edge. However, deploying machine learning on edge devices or at edge-boxes supports different requirements from those run in the cloud. Aside from resource constraints, dynamic environmental conditions necessitate that edge AI be more power-, energy-, and performance-efficient. Herein lies the motivation for a dedicated computational infrastructure at the edge that addresses the needs of AI workloads, enables improved efficiencies, and helps seize upon the advantages of edge AI. The ability to nibble away at the limitations of edge devices and push at the frontier for next-generation ultra-low latency applications is imperative. A dedicated edge infrastructure that serves machine learning models tailored for inference refinement, learning compressions, or small incremental updates can enable the much-needed low-latency and localized learning capability. The deployment of specialized AI accelerators across the edge computing infrastructure can maximize processing throughput while trimming down compute- and data-related latencies, thereby catering to the needs of the next-gen wireless framework.

III. IMPORTANCE OF ULTRA-LOW LATENCY

Although wireless connectivity is ubiquitous, there is a growing demand for wireless performance especially for mobile computing systems because of the increasing size and diversity of wireless workloads. Growing wireless capacity is thus required to support the increasing wireless traffic demand across all types of applications. In addition to expanding capacity, wireless networks and the mobile devices need to support ultra-low latency as well since ultra-low latency is critical to the performance of a class of innovative mobile and IoT applications enabled by edge AI.

Mobile edge computing and AI-based algorithmic innovation have become increasingly intertwined and have created a perfect storm for ultra-low latency applications. As edge AI as a data and outdoor vision recognition enabled by fast and energy-efficient edge AI is becoming a reality, advances in wireless networks are needed to support the compute and bandwidth bursts of data transfer towards the edge. This is especially true for the newly emerging use cases in robotic and drone applications. The ability to offload and uplink large-size AI task data and receive device control action responses, all with ultra-low latency to-and-from the mobile devices work in seamless coordination with edge AI response time. Fast response time requires low latency data pipelining, which is the DNA and secret sauce of the upcoming 6G wireless transition. Latency-sensitive applications have attracted significant interest and have played a large part in the quest to develop a new generation of mobile and IoT applications. The list of applications is long and covers various segments. One of the earliest applications and long-time motivation is augmented and virtual reality, in both entertainment and business perspective. Blockchain technology has existed for a few years and still attracts a wide audience and users as a new investment opportunity, but it is latency-sensitive, and a more efficient service via edge computing is required. Applications in industrial, especially robotics and drone-based industries have received increased interest recently because of advancements in AI and robotics.

Eqn 1: Jitter (Latency Variation)

$$\text{Jitter} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i - \bar{D})^2}$$

Where:

- D_i: Delay of packet i
- $\bar{D}$: Average delay

IV. SEMICONDUCTOR TECHNOLOGIES FOR EDGE AI

Artificial Intelligence (AI), and its branch of edge AI, will require high-performance, ultra-low latency computing chips that can be deployed on the edge of the network. These chips must enable ultra-low cost and energy-efficient operations to process and analyze massive streams of data, such as image, video, sound, or language, that IoT devices will continuously generate. Motion and sound detection at the boundaries of the human experience - audio and visual - and other computer perception functions will be needed to monitor the hyperspace of edge devices connected through 5G/6G networks. At the telecommunications edge of the AI initiative, semiconductor innovations powered by hardware accelerators will be at the forefront of what new AI models can deliver. Beyond hardware accelerators, other semiconductor blocks will also be used throughout the entire path of analytic model offerings, ranging from small AI Engines to Floating Points Processors, Digital Application-Specific Integrated Circuits, FPGAs, mobile System-on-Chip, and wireless modems in sensors and base stations. Demand will be strong as chip design tools make it easier to co-design and prototype chips containing different types of semiconductor blocks.

Beyond the many types of chips to build AI models and algorithms, the continuous improvement of chip design methods will also accelerate innovations in AI applications. Electronic Design Automation tools are crucial for design reusable and module-based chips, a method which lowers the cost of design for companies wishing to enter AI by only putting together ready-to-deploy and secure AI modules. Co-design will also feed the growing importance of soft chips, which make increased flexibility in chip usage possible via Ubiquitous Computing tools and sensor models, Manager System Chip, or smart-sensing Design Electromagnetic Computing. Also, photonic-layer technology and new chip can increase chip performance by overcoming technological bottlenecks in current electronic-layer chips. Hence, with these advancements, future models might calculate the perception with the minimum possible amount of light necessary, and by doing so, the future chip might help keep the sensors at the edge of the communications resources-efficient and battery saver, exactly as it is appropriate for Edge AI models.

4.1. Types of Semiconductors

Because the physical and operational characteristics of a semiconductor are determined by its distinct atomic structure, semiconducting materials and devices proprietary to the semiconductor industry are made using specific materials. These are often key components in equipment ranging from computers to radars. The material used for a semiconductor is typically different from a metal, insulator, superconductor or a beacon. Since other types of devices cannot perform the specific task that semiconductors perform in integrated circuits, semiconductors are called the backbone of modern electronics. Based on the atomic structure, semiconductors can be classified into two categories: elemental and compound semiconductors. The most common type of elemental semiconductor is silicon. Other elemental semiconductors include germanium and tin.

Compound semiconductors are made from two or more elements, including two of the elemental semiconductors. Compound semiconductors often outperform elemental semiconductors because of their modified band structure, but they are usually more expensive to manufacture. Commonly used compound semiconductors are silicon carbide, gallium arsenide, and gallium nitride. Compound semiconductors are better than elemental semiconductors mainly because of their physical properties including energy band gap, thermal conductivity, thermal resistance, etc. They are used in a wide range of high-speed and high-power electronics applications including telecommunications, power amplifier, microwave and millimeter wave technology. On the contrary, elemental semiconductors are usually used in microelectronics applications.

4.2. Advancements in Chip Design

The rapid churn of semiconductor innovation has made possible several alternatives to the von Neumann execution model of computer design, allowing for tailored architectures that match the algorithmic demands of specific workloads. In this section, we present a broad overview of the possibilities today in chip architecture, from general-purpose compute engines with energy-efficient adaptations, through specialized acceleration engines, to dedicated systems. The first risk of chip innovation, of course, is that the only customer for the innovation is the chip itself. The programmable architectures based on FPGA technology do a much better job of scaling and producing chips for edge AI. While FPGAs are still prohibitively expensive for almost everyone, soft and hard implementation options for FPGAs, chips, and both seem to be finally finding their sweet spot, allowing more people to afford FPGAs on sheet or on discount after prices tank. Programmable chips to accelerate machine learning workloads have also become an area of concentration for chip startups in the past few years. Despite high technology risk, the potential payoff may also be a factor of 10 higher than the solutions offered by dedicated chips.

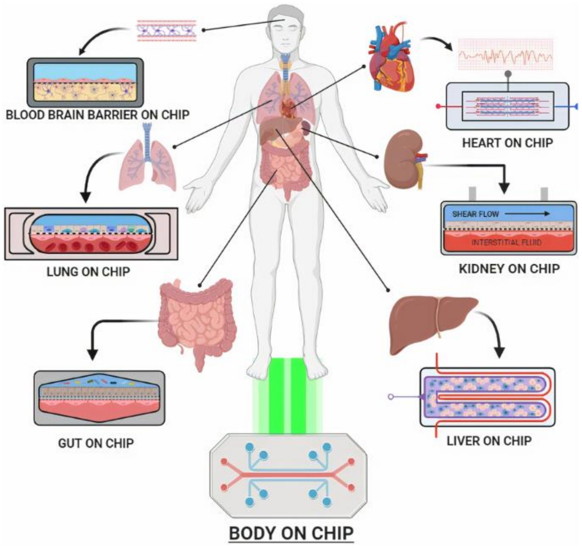


Fig 2: Organ-On-A-Chip Technology: An In-depth Review of Recent Advancements and Future of Whole Body

But implementing silicon for an accelerator chip can take so long that it's not a good bet even for large enterprise companies. For these reasons, we will focus on architecture innovations for accelerator chips and inference and training engines that deliver quasi-dedicated silicon implementations, backed by hardwiring technologies, but remain current enough to boot up on a new algorithm, either by logic, associative content addressable cache application, or neural structure. For AI workloads that naturally fit into this category, the ideal accelerator is a hardwired logic engine that emphasizes associative lookup, memory, and lookup technology. However, inference tasks also lend themselves remarkably well to application-specific general-purpose compute engines or hybrid chips, programmed to a large degree by software, with either FPGA-like execution flexibility of special units or pipeline stages of fixed units operating in a synchronous fashion.

V. WIRELESS NETWORK ARCHITECTURES

Designing the right architecture for wireless networks is critical for achieving ultra-low latency and deterministic connectivity and for enabling real-time AI applications at the edge. Existing wireless architectures and protocols today focus on increasing wireless cell capacity and are designed with a best-effort paradigm, thus cannot guarantee ultra-low end-to-end latencies, in the order of milliseconds, nor delivery assurance, which are essential for edge AI applications.

In this section, we will analyze the inherent limitations of the existing designs and protocols in the context of wireless architecture, particularly for next-gen 5G and beyond networks. We will also explore emerging designs based on principles of resource reservation, radio task offload at the physical layer, joint physical layer – networking layer optimization, and distributed AI, which promise to overcome these limitations.

Low latency and reliability are among some of the ambitious goals of the next-gen wireless networks, 5G and beyond, which is driven by the technological advancement of AI, the increasing adoption of IoT, and the rise of new applications such as autonomous driving, smart cities, and remote healthcare. The envisioned ultra-reliable low latency wireless connectivity has the potential to transform our society, by revolutionizing how we live today, and making what seemed impossible, possible. However, meeting these goals at scale, has numerous challenges, which focus on the following two aspects: (i), how low is low-latency when it comes to end-to-end latencies, and (ii), how do we manage to pack the huge volume of data, driven largely by the growing popularity of video, virtual reality, augmented reality, into large groups of edge-nodes. Indeed, the answer to the first question is in the domain of milliseconds, while for the second question, bringing the computation at the edge closer to the users is one compelling solution, which needs to be validated and standardized. These two questions essentially influence the wireless architecture and protocol design of 5G and beyond networks.

5.1. 5G and Beyond

The fifth generation mobile network (5G) is a culmination of multiple radio interfaces, advanced networking concepts and enabling technologies, such as Software defined Networking, Neural Networking, Information Centric Networking, Ultra Reliable Low Latency Communications, Intelligent Reflective Surfaces, Physical Layer Slicing, Mixed-Focus Arrays, light-fidelity, Deep-Space Communications, Massive Type Communications, Optical Wireless Communications, and Tactile Internet, for providing diverse basic services, such as Ultra-Reliable and Low Latency Services, Enhanced Mobile Broadband, etc., and advanced application solutions across verticals like Industrial Automation, Smart City, Connected and Autonomous Mobility, Tele-medicine, Education, and Entertainment, etc. The aim of 5G is to address unified connectivity on a shared platform for all types of vertical application space worldwide, including the IoT devices ranging from low complexity, low rate, loss tolerant to high complexity, high rate, latency sensitive.

URLLC and eMBB, which have been selectively used as basic services for most of state-of-the-art 5G commercial service deployments worldwide to date, along with Edge Computing, have created segmented traffic, while breaking basic service overload bottlenecks for latency sensitive applications like Augmented and Virtual Reality, Driving Safety, Autonomous Driving, Cognitive Healthcare, Industrial Robotics, Wireless Factories, and eMBB heavy content delivery, respectively. Additionally, the use of Multi-Access Edge Computing technology supporting real-time processing of services at the wireless edge, has continued to reduce latency while increasing throughput for applications like Industrial Process Control Automation, Road and Traffic Safety, Evaluation and Management of Physical Health Parameters, Behavior Assessment and Mental Health Monitoring, and multiple Interactive high-definition, Ultra HD, Live broadcasts, respectively, as edge applications have eventually become key drivers for massive MTC traffic explosion.

5.2. Network Slicing and Edge Computing

A major limitation of 4G, LTE-based mobile networks is their focus on providing a single service – the aggregation of data at high rate – and a single edge infrastructure – radio base stations – across all user device use cases. By contrast, use cases currently emerging driven by AIoT demand radically different characteristics when compared to conventional mobile broadband services. To grow mobile networks as a flexible and programmable infrastructure that can host such diverse use cases, 5G and beyond mobile network technologies are creating the concepts of network and service slicing: the possibility to partition an operator network into virtual networks that can each be dedicated to a specific use case and operated in a more tailored manner to activate enhanced service properties.

The concept of a one size fits all mobile network optimized for MBB is rapidly becoming impractical both from a network sustainability and user needs standpoint. The partitioning of mobile infrastructure into virtual networks and multiple parallel domains supporting unique and specific use cases is the only way forward for mobile service operators if investment and operational cost constraints are to be continuously met. However, meeting the need for use case service specific mobile network functionality is just one side of the coin. The other side are mobile service operators' growing requirements for network cost effective remote and in-field operation; a requirement that is hardly met by the relatively high operational costs of conventional mobile network operation today. Slicing that allows mobile services to be operated in a more centralized manner using Edge Cloud infrastructure technology enables flexibility as well as cost optimization of service deployment and pricing.

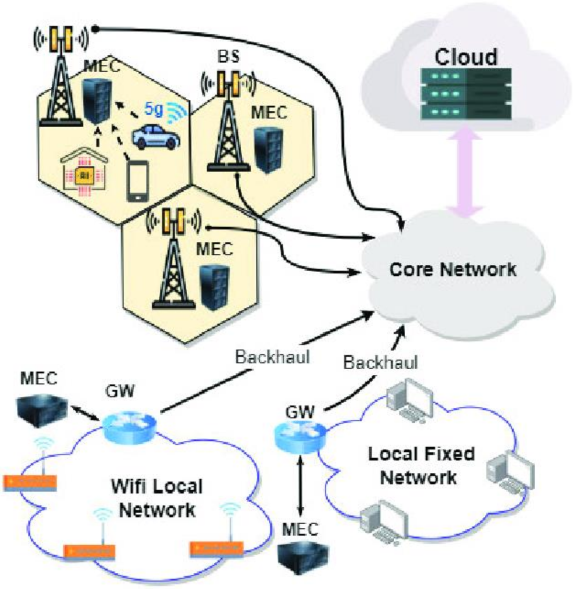


Fig 3: 5G-multi-access edge computing architecture

VI. CHALLENGES IN ACHIEVING ULTRA-LOW LATENCY

Low-latency is one of the most critical design criteria for edge AI and next-gen wireless networks because actionable insights must be available within a short period of time. The overall system latency for the end-to-end process from data acquisition to model inference usually includes the following components: (1) sensor acquisition time: the time needed for the sensor to sample the raw data and transfer it to the processor, which depends on the sensor architecture and the data rate configured; (2) signal processing delay: the time taken by the digital electronics to perform pre-processing of the raw data, such as filtering, compression, and other manipulations; (3) data transmission delay: the bottleneck time during which the processed data travel across various paths to reach the AI processor; (4) model inference delay: the time required by a special purpose processor to perform AI computations; and (5) actuating time: the time taken to carry out the action recommended by the AI model. Although sensor acquisition time and actuating time are usually fixed as they depend on the hardware setting, the other three delays constitute the main focus of this chapter since they contribute more than half of the overall system latency.

Eqn 2: Processing Delay in AI & Baseband Units

$$D_{\text{proc}} = \frac{N_{\text{ops}}}{f_{\text{clk}} \cdot \text{IPC}}$$

Where:

- N_{ops}: Number of operations
- f_{clk}: Clock frequency
- IPC: Instructions per cycle

Signal processing delays can be mitigated in a number of different ways, including reducing the size of the input data sent to the AI chip, computing at lower precision formats such as binary or ternary, and using parallel processing. Reduction of input data size is often accomplished by careful sensor processing of the raw data in the FPGA or customized hardware processors, especially for very high dimensional data like images and videos. Pre-processing units in the acquisition path can compress the high dimensional data by binning or downsampling techniques. Additionally, by resorting to low-precision processing such as binarized neural networks, edge AI latency can be reduced significantly. Further gains in throughput can be achieved by using application-optimized processors that are designed to execute specific deep learning algorithms.

6.1. Signal Processing Delays

Wireless networks are being re-engineered to support latency-sensitive applications in the impact of 5G systems. Although numerous improvements at the architecture level have taken place, focus has mainly been on the system-side optimizations, with insufficient work devoted to exploring the semiconductor processing technology space for edge

DOI: 10.17148/IJARCCE.2022.111258

applications that demand ultra-low latencies. In this chapter, we explore the signal processing latency bottlenecks in wireless communications and discuss the possible silicon technology alternatives to address them. We identify the component building blocks of wireless systems that create latencies in the signal processing chain and reveal their limitations in terms of processing speed, precision, and power dissipation. Based on this analysis, we classify different messages and features in the present and future wireless systems and identify an appropriate processing scheme for each data feature based on its precision and processing time requirements. Signal delay is comprised of various components from multiple signal processing blocks in the wireless system chain. Baseband signal processing involves various signal processing functions such as modulation, pulse shaping, channel estimation, equalization, phase recovery, and detection during transmission, which combines, shapes, and optimally detects the message signal carried on the channel to accurately reconstruct the baseband signal at the receiver. Each of these functions introduces a latency because they are implemented on blocks of silicon chips, and the successive signal processing tasks are highly parallelized. Signal latency may also take place in the power amplifier in the transmitter path and the low-noise amplifier in the vicinity of RF/ADC conversion. These functions introduce a latency in the growth of output signal amplitude, which then induces a delay in the amplification circuit.

6.2. Data Transmission Bottlenecks

A key factor improving latency is processing efficiency but minimizing processing load is even more important. For instance, when detecting objects using a deep learning model, limited memory and compute availability restrict the choice of deployment platform and therefore also the choice of model. Experimental results show that the trade-off between model accuracy and inference time is highly sensitive to the choice of model. A wrong choice can result in substantially increased latency, thereby degrading user experience for real-time applications. The current research trend that leverages model compression to design ultra-lightweight deep learning models can also help by making the edge compute more power- and area-efficient. Additionally, by offloading inference to the cloud during nighttime, the choice of model can be shifted to high accuracy, thereby using the edge platform only for a small subset of data. This would alleviate some of the data transmission bottleneck.

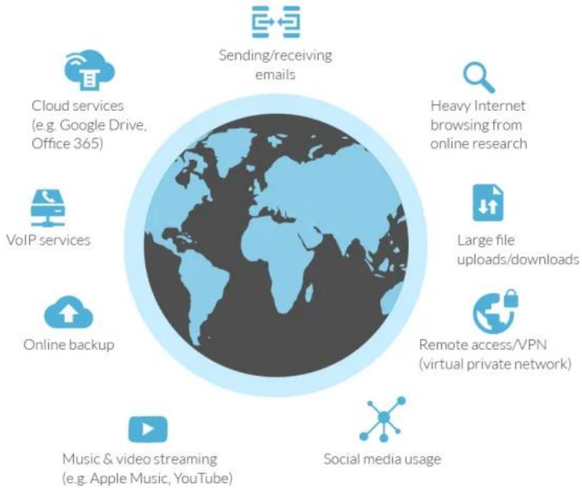

Sending/receiving emails
Cloud services (e.g. Google Drive, Office 365)
Heavy Internet browsing from online research
VoIP services
Large file uploads/downloads
Online backup
Remote access/VPN (virtual private network)
Music & video streaming (e.g. Apple Music, YouTube)
Social media usage

Fig 4: Network bottlenecks

The above analysis focuses on the inference step but one needs to also consider the time required for data preparation after data acquisition, as well as data transmission time from the camera node to the processing element. Cameras can output image sequences at a rate of multiple hundreds or even thousands of frames per second. The amount of data to be transmitted could therefore be large, especially in the case of hyperspectral cameras that generate a significant amount of data for a frame. Various specialized protocols attempting to reduce data transfer bandwidth have been proposed. There has also been work demonstrating a modified version of coaxial cable that has high bandwidth and low power, area, and latency. These developments collectively allow sensor design with greater efficiency, enabling low latency dense sensing of the environment. However, the choice of communication medium is always a trade-off. Latency must be also included in the design equations guiding platform choice between edge and cloud.

VII. INNOVATIVE SOLUTIONS FOR LATENCY REDUCTION

Achieving ultra-low latency in next-generation wireless networks demands exploring innovative solutions. For instance, for a given network architecture, assignment of wireless resources to user equipment (UE) in a non-orthogonal manner at the transmitter side, and joint decoding of the non-orthogonally assigned signals at the receiver side, can greatly reduce latency. In addition, leveraging multi-connectivity can be very useful in lowering latency. Using multiple air interface technologies simultaneously to create several links between a UE and the RAN is referred to as multi-connectivity. The links can either route the same service traffic over separate links in parallel, or separate the service's different flows and route them over separate links. In both cases of multi-connectivity, recovery from link failures, load balancing, and traffic optimization can occur at different time scales, including load balancing across links every transmission time interval.

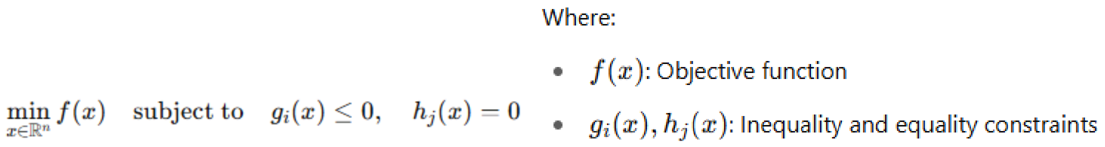
Also, as 5G is designed to connect people and things to the Internet, it has to support several protocols to access the services that are available on the Internet. Packet routing involves more than just its transmission time across the wireless link; the latency-conscious decisions can also involve the time spent in the processing of the packet and forward at the routing nodes, the queue, and the path it takes to traverse the routing nodes. Latency incurred at these points can unwittingly overshadow the naturally low latency of the radio access, and thereby warrants careful management and optimization. This fact necessitates the need for AI-driven optimization techniques for low latency in combined radio access and networking scenarios.

As multiple links may involve wired and wireless components, achieving ultra-low latency demands joint low-cost hardware acceleration techniques on these components. Be it wired or wireless, the RF front-end, routers, and switches have clock cycles that require fast analog-to-digital converters and fast digital-to-analog converters for signal quantization. In addition, the routers and switches have to provide fast packet storage since packet queuing delays can be non-vanishing and consume significant delay at larger network scales, while the temporal variations seen by separate packets routed through the same device can cause greater over-scaling times. Fast packet storage is essential to realize ultra-low-latency networks.

7.1. AI-Driven Optimization

AI applications, especially those operating in edge environments, often require ultra-low latency, which is only possible through consideration and optimization of all conflicting constraints at both the hardware and the software algorithms level. These constraints are many-fold, including the model compression and optimization requirements aimed at reducing the inference latency and increasing the energy efficiency/operational lifespan; the service level objectives mapped to quality-of-service requirements on the user experience; the requirement for a minimal number of connections with good quality at a point in time in order to ensure the revenue generation by mobile operators; the limited radio resource allocations that guarantee fair access to resources by users while maximizing the operators revenues; and the privacy and safety/security constraints ensuring that the edge AI applications do not over-leak data and do not incur denial-of-service at the edge.

Eqn 3: General Optimization Objective



$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{subject to} \quad g_i(x) \leq 0, \quad h_j(x) = 0$$

Where:

- $f(x)$: Objective function
- $g_i(x), h_j(x)$: Inequality and equality constraints

In general, edge AI applications offer conflicting objectives which are often NP-hard optimization problems requiring a graph partitioning on dynamically updating graphs, such as the case with mobile-enabled device SLAM. These and other problems with AI that dynamically evolve in time can be solved efficiently using exploratory phase-reduction. The specific versatility of the decision spaces of these AI objective problems illustrates the necessity of using specialized AI hardware acceleration chips that can be configured for optimal performance in solving these particular analytical yet solution-guiding problems, in order to become compute-completed and lowered their energy utilization while satisfying their SLOs.

7.2. Hardware Acceleration Techniques

While AI-driven optimization and learning techniques are very useful in mitigating or eliminating latency sources, they still introduce additional processing overhead. Hence, in many latency-sensitive applications, performance benefits may be limited. Moreover, the use case may not lend itself well to AI-driven optimizations. For this reason, many latency-sensitive application endpoints, especially those that are computationally intensive, rely on hardware acceleration techniques to achieve ultra-low processing latencies.

In general, hardware accelerators can be implemented in two broad categories: 1) flexible hardware that provides a good compromise between efficiency and flexibility, and 2) custom hardware that is the most efficient alternative, but at the cost of lack of programmability.

Functionally programmable hardware such as FPGAs and GPUs can be considered as flexible alternatives to custom hardware accelerators. Although such solutions achieve superior performance-per-watt efficiency compared to general-purpose CPUs, they lag far behind custom hardware accelerators in terms of efficiency. Despite their superior energy efficiency, custom hardware accelerators face the challenges of programmability, scalability, and latency associated with extra overheads, such as fixed latency-based data movements, and fixed interfaces for chip-to-chip or board-to-board interconnections. Although these issues certainly limit the application domains of custom hardware accelerators, they are more than offset by their superior cost, area, performance and power efficiency on such domains. Their ultra-low processing latencies and negligible processing overhead thus make them the best alternative for ultra-low latency Wireless Edge Applications.

VIII. CASE STUDIES OF EDGE AI APPLICATIONS

The Edge AI market is rapidly emerging with new applications deployed around us. In order to realize the promise of ultra-low latency and high resilience of edge platforms utilizing semiconductor innovation, we need to build these application scenarios in a thick cloud infrastructure integrated with networks to uplift the demand side. We will present some example AI applications that we can realize using edge intelligence. As technology rolls out in parallel, it will become easier to apply our semiconductor innovations at the edges.

The whole automation in the transportation, supply chain and logistics industry ranging from early-stage trucking automation to level 5 driverless techno-logistics will motivate the need for ultra-low latency compute capabilities. Consider a level 5 driverless taxi using a dense sensor network collecting information related to weather/cloud, building/road surface, any diverse physical supports etc. The data is collected, pre-processed and analyzed in real-time using networks over the edges. The amount of data that needs to be transmitted over the network and processed in the cloud should be minimal and only the results of the analysis (possible risks) should be sent to the cloud for further processing.

Smart cities are also rapidly emerging using networks and Edge AI. Optimize your cities to reduce air pollution, smart waste management, disaster management, smart buildings, maximize your tourism, transport and business with Edge AI. Embedded sensors will collect data and perform optimal AI based data processing and share information over the edges. Use sensors to detect motion and record all details related to irregular movements. Deploy cameras on roads to monitor for accidents or unusual object presence. Intelligent lighting is an integral part of cities. Smart streets will detect and update the status over the network. Smart parking system notify available parking spots and it should be automated for payment.

Industries which have mostly been convert to smart industry during a recent period need a smart edge-IoT controller that connects many field devices and has a significant level of intelligence for data collection, preprocessing, control and optimization of processes. Process and factory wide data collection by creating a mesh network of edge devices/controllers allow you to optimize multiple use-cases to prevent equipment downtime and ensure quality control.

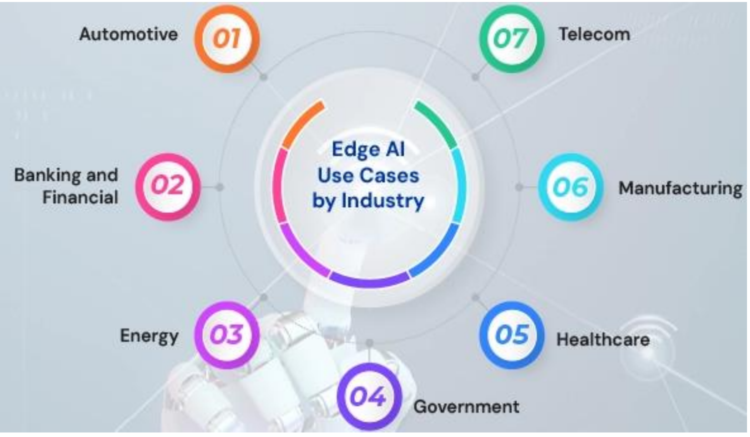


Fig 5: The Future of Technology

8.1. Autonomous Vehicles

The automotive industry is undergoing a digital transformation powered by AI. Just as the Internet ushered in a digital revolution for content, services, and economy, tools to capture and analyze data can reshape the automobile to become a smart, entwined system of transportation and communication. The next generation of automobiles will transition from being simply vehicles of transportation to being fully autonomous electric vehicles powered by AI algorithms, capable of sending and receiving vast amounts of data, and integrated into smart logistics. The ultimate objective of an AV is to be capable of operating in any environment under any conditions and with any traffic regulations without human intervention. This demands a wide range of technology innovations with respect to sensing, usage of map data, hardware development and decisions fused from the many participating subsystems.

Edge AI technology that enables real-time 3D object detection, classification, segmentation and tracking is created in road testing of these AVs. Data from camera sensors is captured en route, labeled and annotated to create transparent 3D objects using 3D Moving Object Detection. Resilient AI engines fused with map data are made more efficient and accurate through joint training that optimizes resources including latency. 3D information from the edge is merged with information from radar and localization data to produce highly accurate 3D maps at the spatial resolution desired. Use cases for the data from AVs include transportation, logistics and delivery, urban and highway applications, security and defense, entertainment, mobile offices, charging, mobility, data service, and new sensors. An AV's drive is accomplished by many servers doing thousands of tasks; therefore there are thousands of use cases for every drive of the AVs.

8.2. Smart Cities

Wireless communication connects both autonomous vehicles and smart city sensors to remote repositories for persistent data storage and big-data analytics. Such configurations transfer raw sensor data over slow backhaul exactly to achieve the massive scale of the Internet of Things, but incur roundtrip overheads and latency for any use that requires updating vehicle operations in plans selected by onboard agents. Here, we focus on applications of Edge AI where speed is paramount, and hence we do the opposite, distributing wireless sensor data processing to the edge, yielding a regional, lower-complexity solution. Localized detections and labels for the vehicle surroundings are expected to reduce latency more than centralized cloud solutions and increase safety and security through situational awareness. Localized models are inherently robust to variable wireless condition and local edge sensor calibration, and localized predictions more easily allow templated, zero-shot learning of rare classes relevant only to the temporospatial deployment. Low-power, interedge-device wireless connections could enable peer-edge observations that could enhance vehicle operating situational awareness even for failure modes linked to sparse observations. High-speed edge computing is needed to enable localized predictions utilizing local edge observations, and a new multi-model, configurable architecture enabling enhanced situational awareness through the integration of observations and predictions from multiple neighboring devices is introduced.

8.3. Industrial IoT

Managing mission-critical operations in manufacturing, logistics, or quality control uncompromisingly requires extreme systematic and responsive information process flow in real time. For instance, in a semiconductor manufacturing factory, the re-entrant nature of the entire wafer production cycle mandates flip-flop information updates on submit, start, and finish of processing at every single process tool. These tool feedbacks drive multi-sensor multi-agent-based optimizations of wafer routing and staging in transport, in rack waiting, in reticle, and in equipment-processing resource allocation and scheduling in wafer front-end assembly, etch, implant, photo, polish, and test. Automatic dispatch algorithms-based automatic guided vehicles, robots, and/or lifts work in an assembly-line fashion under seamless communication with a centralized digital twin infrastructure on-site instruction and performance feedbacks. The manufacturing enterprise is then coordinated in a decentralized manner, with a cloud-based digital twin-controlled upstream and skyline tiering and third-party logistics. Beyond manufacturing floors, the above real-time computing is equally crucial or even more important in managing facilities networks comprising warehousing, temperature control, or HVAC at any of the in-house, outsourced, or two- to six- to seven-tier supply networks. Those facilities may contain racks and hooks for WIP and finished goods, or inventory in containers either dry, liquid, or gaseous.

Management of unpacking, commissioning, inspection, maintenance, degradation, fault tolerance, replacement, calibration, etc., of high-cost semiconductor production tools in fabs requires ultra-reliable low-latency communication as well. When qualified and validated, such AI-enabled computational engines, trained off data from historical operations or from the digital twin environment, would be used repeatedly, thrice or four-times during their lifetime for supervisory and operational control, optimizing cost and time.

IX. FUTURE TRENDS IN SEMICONDUCTOR DESIGN

As Moore's law slows, and as we approach the physical limits of conventional semiconductor scaling, innovative architectures operating on novel principles of computation represented by quantum, neuromorphic, and even topological computing appear poised to play an increasingly important role. These emerging computing paradigms not only present a solution for the exploding demand for performance and energy efficiency in areas such as machine learning, they also present an opportunity for the semiconductor industry itself, by enabling new devices and processes, a return to rapid innovation, and as a byproduct, revival of growth and investment as novel semiconductor technology nodes become economically necessary and thus valuable.

At the heart of quantum computing are specialized quantum memory devices called qubits, which exist in quantum states of superposition, allowing them to store and compute on enormous amounts of probabilistic data simultaneously. For quantum computers to achieve the full exponential speedup over what classical computers can achieve only quadratically for general search and optimization problems, are applications in high demand such as integer factorization for secure communication, high-dimensional sampling for chemical simulations, or optimization and training of Restricted Boltzmann Machines for deep neural networks — fault-tolerant quantum error-corrected circuits will be necessary. Practical implementation of larger, fault-tolerant quantum circuits requires special gates called logical gates, which serve to create unusual quantum states, perform quantum teleportation, and transform the states while contaminated by noise — and this in the presence of ambient heat that can break qubit coherence by tunneling to another energy state, as well as radiation that can ionize qubits, causing information loss. Quantum error correction of this kind relies upon redundant representation of information in a small set of qubits called a logical qubit, and we will describe the design implications this quantum fault tolerance creates for our traditional von Neumann-based transistor technology.

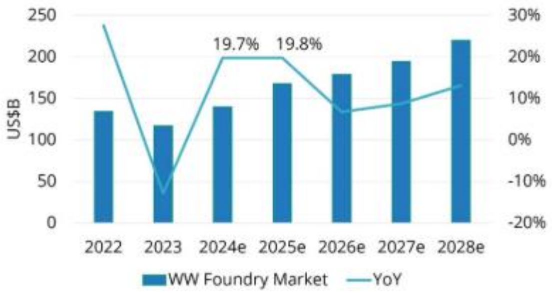


Fig: Global Semiconductor Market

9.1. Quantum Computing Implications

The next big thing in computing is quantum computing. We provide a brief overview of why physicists and engineers working in this area are so excited at the research direction they are pursuing. Consider that the classical computing model we have relied on for 80+ years to successfully develop more general-purpose applications is based on transistors working as switches, which are essentially logic gates implementing Boolean algebra. For semiconductor technology companies, these companies have focused on finding solutions to issues of scaling transistors to lower and lower dimensions while improving the theoretical upper bounds of speed and power; the majority of substantial improvements in speed and power efficiency has been observed with technology nodes every 2-3 years. The quantum computing model is based upon the utilization of qubits that rely on intrinsic quantum mechanical properties of superposition and entanglement to deliver massive parallelization of linear pathways and an exponential increase in the size of the search domain for quantum algorithms. Essentially, our classical computer only 'sees' one path of execution through a calculation at a time, while a quantum computer 'sees' all of them simultaneously. Not only have experimental physicists demonstrated physical implementations for small-sized quantum computers that exhibit quantum mechanical properties, the existing and future large-scale commercial applications are being developed now by software engineers with collaborations among the engineering and physics communities.

9.2. Neuromorphic Computing

Recent advances in AI have been dominated by deep neural networks for image classification, speech recognition, etc., applying increasing compute resources at centralized datacenter. However, the tremendous success of DNN training and inference has led to long-lasting criticisms on the increasing computation and energy footprint with limited user privacy and security. The lack of energy-efficient algorithms and hardware architectures with the ability to enable continual learning, associative memory, fast online learning, or active learning for intelligent edge devices for AI at the edge with user privacy during inference have limited the generalizability of the state-of-the-art DNNs for new tasks. New approaches inspired by neuroscience and brain models offer promise of the next generation of AI with ultralow compute and energy footprints close to that of the mammalian brain. These neuromorphic algorithms go beyond the current supervised DNN algorithms with enhanced memory capabilities and learning efficiency, incorporating both supervised and self-supervised learning, along with online and continual learning. Additionally, existing VLSI hardware accelerators for DNN inference do not efficiently map onto the data representation to efficiently reduce the model size and number of matrix-vector multiply operations, leading to further increase in compute and energy cost for DNN inference with high redundancy when deployed on edge devices. Large memory footprints of high-performance AI computation using DNNs, and potentially large data and energy requirements prevent deployment in many edge applications. Furthermore, training of large DNNs for general-purpose intelligent decision-making also imposes constraints.

X. REGULATORY AND ETHICAL CONSIDERATIONS

The potential of edge AI, along with the business models and revenue generation opportunities it affords, is immense in both private and public domains. Private companies will build products and services that will enable and help accelerate the adoption of edge AI systems, but they will also seek to monetize them. There are ethical considerations around how these companies might monetize the use of edge AI systems, in addition to the context and use cases powering their development. Working together, industry and government must seek to mitigate the risk of negative unintended consequences while supporting innovation for positive social impact. Although edge AI for wireless networks is ripe for development, there are a variety of regulatory and ethical detours that could slow down or halt technology product and service deployment. The area of regulatory and compliance challenges borders on overwhelming. We cannot explore each in detail but will highlight two key issues – data privacy and regulatory compliance – that are particularly relevant as we think about the deployment of edge AI networks. As intelligent devices are increasingly deployed to process video and sensor data for applications in defense, national security, and public safety, we must think not only about their operation but also about the data that underpins them. The inherent need to collect data from the public raises important concerns regarding data privacy that the private companies developing these new capabilities, along with government users, must carefully consider.

10.1. Data Privacy Issues

With the adoption and proliferation of real-time AI, the call for furthering data privacy has become all too relevant. Edge AI promises to collect and process sensitive information at the edge; how this data is used, and how much insight it allows for access controls, monitoring, and inferences without explicit consent are critical to both technology adoption and privacy safeguarding. Data storage at the edge does not alleviate privacy concerns, since malicious actors may hijack edge infrastructure and access stored, private information. Lifelogging and perceptual cameras are generating massive amounts of biometric data since streamlining the information processing workflow have made automatic analysis accessible. AI with emerging consumer uses and face biometric support has been accused of "moral hazard" or the adaptation outweighed by the loss of privacy; such considerations are ubiquitous in a world supporting predictive surveillance, operation oversight with face biometric support, or authoritarian and reactionary uses of AI technology. Through the constant cooperation workflow, technology designers have a social contract to enable words as a user-friendly opt-in and opt-out system for life and bodily monitoring in prevention of revolution, economic uncertainty, and perturbations to teenagers and youth.

Questions of ethical, moral, social, and economic impacts are the focal research area in many technology and political forums, regardless of means of surveillance verification. An open question is how careful should edge AI be to not inconvenience individuals both who do and do not opt-in to perceptual systems. Where consent is implicit, the border between usability and loss of privacy becomes important – if not decisive. Such considerations will undoubtedly dictate the paths chosen for semantic analysis and operating passes, and by what criteria incongruences are preventable.

10.2. Regulatory Compliance

As previously mentioned, radio hardware occupies only a tiny corner of the edge AI world. Building a feasible and commercially available Ultra-Low Latency Wireless Edge AI solution will need to not only capture technical advances in key areas such as new optimizations of radio design and channel coding, low-noise RF front ends, AI-driven beamforming and MIMO, computing and radio integration and optics-small form factor packaging, fault tolerating AI processing, and AI-assisted channel agnostic ultra-low latency signal processing, but also the necessary hardware certification and regulatory compliance in a timely manner. It is critical to efficiently bridge between an experimental demonstration of the technology and its normative constraints. At a system level, multi-Gbps real time wireless links over 15 km distance, for multi-Gbps throughput optical fiber-class links, has been demonstrated via optical wireless links. But in the millimeter and terahertz bands, implementation is subject to regulatory compliance for Spectrum Access Systems for E-band and V-band, and all-optical TX and RX devices.

DOI: 10.17148/IJARCCE.2022.111258

SAS necessity stems from potential harmful interference to incumbent services and handling of multi-point access, as is the case with other wireless systems in other bands, except that E-band links won't extend their 1 km (urban traffic surveillance) and 1.5 km (terrestrial narrow beam downlink connection) km distances isotropically into free space. Terahertz band is only subject to Federal Communications Commission rules, which specify that it is exclusively for non-Federal Government use, until the Federal Government has been accorded priority with no interference. The millimeter and submm wave links require electrical-optical components and antennas with appropriate bandwidths. The documentation need not be onerous or burdensome, but it should address the new developments as applied to the link architecture.

XI. MARKET ANALYSIS AND ECONOMIC IMPACT

In this section we provide an analysis of the potential economic impact of low-latency networks with support for distributed machine learning and AI feature processing at the edge. Low-latency wireless systems are projected to grow to encompass the largest segment of global wireless communication by the year 2026. The key differentiator of these systems is the expected service support for low latency and high reliability use cases. Flexible and customizable ultra-low latency service support will be essential to enable new and/or radically improved services for industries such as automotive, healthcare, education, and manufacturing. We explore several of these use cases below and highlight some investment opportunities as well as potential returns. We also estimate the market sizes for several different low latency use cases and present a breakdown of these revenues by associated technology components.

Economic growth reflects creative disruption driven by innovation and the deployment of new technologies. The semiconductor industry is an important component of global GDP and is well positioned to drive some of this creative disruption in the coming years. The semiconductor industry is deeply linked to the development and deployment of more efficient and advanced wireless systems and other communication infrastructure. Novel technology in support of low latency edge distributed AI and machine learning could spur a new wave of growth for the semiconductor industry as well as generating revenue from other vertical markets. The new generation of ultra-low latency wireless communication networks as well as the enabling component technologies that constitute the next generation have attracted significant interest from industry and in governmental circles, with predictions of large potential economic benefit from successful deployment.

11.1. Industry Growth Projections

Combining massive data processing capabilities with widely dispersed connected devices will accelerate advances in the Internet of Things and cloud computing. By 2030, the world is projected to have about 50 billion connected devices, up from 7 billion in 2018. A number of industries will benefit from these advances, including energy, healthcare, manufacturing, and transportation. The application of edge AI technology to Internet of Things use cases will add an estimated $1.1 trillion to $1.8 trillion to the global economy by 2030. Convergence of wireless communication and artificial intelligence is expected to generate huge value for hardware, software, and services over the next several years. The artificial intelligence market is projected to grow from an estimated $27 billion in 2020 to almost $200 billion by 2023. Similarly, the wireless communication sector is projected to exceed $200 billion during the same time frame. With the emergence of innovative wireless edge AI services, the size of the global artificial intelligence market may reach about $14 trillion by 2030.

Furthermore, edge AI technology is anticipated to contribute significantly to economic growth in several developing nations. Most prominently, recently published estimates forecast that the Promotion of Manufacturing through Artificial Intelligence and Internet of Things in India strategy could help drive $500 billion in economic growth and help create about 10 million new jobs in the country by 2025. Across all geographic markets, increasing demand for ultra-low latency applications such as remote robotic control, augmented reality, factory automation, and smart cities could lead to economic acceleration and job creation. To support this demand, telecom carriers and hyperscale cloud service providers are investing hundreds of billions of dollars to deploy next-gen wireless networks and edge AI ecosystems.

11.2. Investment Opportunities

Advancing the technology and scaling manufacturing capabilities of semiconductor-on-insulator (SOI) microchips specialized for edge AI can establish commercialization pathways promoting various product classes and generate investment return opportunities. One such pathway approaches rapidly growing edge AI applications where microchip performance, form factor, and power consumption are primary design factors. In this instance, SOI microchips with AI accelerator and RF wireless perimeter-processing communications circuits might be integrated in a single chip, and SDKs created. Such microchips could be deployed to edge AI system manufacturers in a variety of product classes, realizing superior system performance, in parallel with access to an increased number of content-generation AI applications.

DOI: 10.17148/IJARCCE.2022.111258

Following successful implementation of high-volume producting deployments, various opportunities to decrease chip advantages begin to emerge. In particular, our work predicts that specialized, light-weight hyper-parameter settings for depth-first quantization of the general model will be created by a select number of chips with a dimensionality-capacity equation throughout the next 12 months. Content generation, either as standalone products or incorporated into various functionality enabling task automater systems, is expected to emerge as a rapidly-growing application area within AI infrastructure support. In this environment, edge-based SOI chips with instruction throughput and power density increase rates might be leveraged to deploy numerous voice and text generation microchips for the duration of the rapid content generation growth.

Subsequently, these and similar near-equivalent chips configured with models trained on proprietary or domain-specific micro-datasets become rapidly increasing demand components of the business-supporting conversational agent systems of large and mid-sized companies. Subsequently, companies expanding conversational agent service offerings for enterprise clients engage these microchips to deploy proprietary chatbots to corporate intranet systems. These chatbots support corporate personnel enabling dropped call services for third-party conversational agent service firms during peak load periods. Through this vending cycle, SOI chip manufacturers realize downstream revenue growth, via chip component supply agreements with chip neural network deployment initiators.

XII. COLLABORATION BETWEEN ACADEMIA AND INDUSTRY

The collaboration between industry and academy, for example by exploring research partnerships or facilitating corporate research engagement programs, has been recognized as key for enabling faster transition of innovative research outcomes into products and services. The semiconductor industry was established on the collaborative foundation between university research centers and research oriented corporations. It is a partnership model that proved to be beneficial to both sides, in terms of quality of the school's education and cutting-edge research on one side, and visionary-innovation and fast product deployment on the other.

We have witnessed some promising partnership models emerging recently for quickly accelerating discoveries, driven by the fusion of previously distinct areas of research. Examples are a partnership model for machine-learning enabled nanoscale imaging, enabling new breakthroughs in extreme-ultraviolet lithography defect detection, or a novel partnership model introduced for the development of the next-generation audio processing chips, leveraging co-design concepts. Can we establish such a partnership model in the semiconductor industry? On one hand, there are the fastest and biggest technological drivers worldwide, AI and 6G, for which the semiconductor industry has a pivotal role. On the other hand, research challenges which require the most innovative breakthroughs, such as nanodevices and ultralow-power circuits and systems for extreme Edge AI, are being tackled by university researchers worldwide, and where the semiconductor industry has deep expertise. Can we establish a thriving partnership model in a global world?

12.1. Research Partnerships

The design and implementation of next generation hardware accelerators for executing AI at the edge will require collaboration between several groups of researchers: those from industry who are interested in deploying new solutions at scale; those from algorithmic AI fields who are developing the innovative techniques that push our understanding of AI; and those at the intersection of AI and computer engineering who are designing the new specialized dataflow requirements and bottlenecks so that the algorithmic researchers can consider these constraints when proposing new models. Each of these groups have expertise and proposed solutions to contribute, and a collaboration between them will help to ensure that the resulting next-generation hardware accelerators will be deployed to solve useful real-world problems.

Public research partnerships using government funds have traditionally provided the link between the industry and the academy. Research engineers develop proof-of-concept systems using the latest innovations in algorithm and hardware design. Then, these systems are used by the original designers to better understand the constraints of actual implementations. The research partnership program is an important mechanism to provide an early market for research technology since typical TTM are 5-10 years. Afterwards, the companies who are interested in deploying the solutions either implement the technology in a product or buy a startup. Typically, specialized products have a limited market and there are not enough resources to deploy a mature company. Furthermore, without some incentive, the corporate partner may not provide sufficient support for the scaleup while the academic partner is unlikely to be willing to work on the project alone in the absence of a funding contract or agreement.

12.2. Knowledge Transfer Initiatives

The complexity, diversity, and dynamic nature of the semiconductor technology cycle and the edge AI wireless network applications fueled by it cannot easily be captured by traditional textbook education. Connecting students with industry

DOI: 10.17148/IJARCCE.2022.111258

for learning is no longer an option; it is a necessity. However, it cannot be done by shorter summer internships or semester-long co-ops alone but rather by systematic partnership models that structure the learning experience at the interface of the university ecosystem and industry partners. The success of such initiatives requires careful design of the concept and realization of the partnered experiences. The merging of bachelor and master phases of engineering education over the past decades exacerbated the challenge. Students need to be brought into contact with their future employers long before graduation, at an early stage in their education.

A STEAM collaborative knowledge transfer project has been running for the past five years. Called "Learning by Designing", its goal is to realize actual technology products. Teams of undergraduate students from various majors within the university work closely with an interdisciplinary team of engineers towards the release of an actual product on a tight timeline. Students realize system and circuit designs for various wireless applications, such as Wi-Fi microwave Power Amplifiers, Ultra-wide Band PAs, Bluetooth, Digital Amplitude Modulation, and Frequency Shifting Keying, as well as RFID and CMOS Microelectromechanical Systems/CMOS systems prototypes. The project emphasizes knowledge transfer through industry-liaison during all phases involved in technology realization, from the actual wireless signal specification, system design, circuit realization, device performance modeling, circuit characterization, PA integration, and testing, to the final prototype presentation.

XIII. CONCLUSION

In the future network deployments are envisioned to be more intelligent and use resources more effectively by offloading and processing data at centralized locations within the network edge. Building Artificially Intelligent (AI) based wireless networks is very challenging because of the ultra-low latency, real-time and high reliability requirements. Wireless edge AI will require a massive amount of metadata to not only be transferred but also processed efficiently and in a timely manner to avoid bottlenecks that could lead to stale decisions. We discussed the architectural components and protocol overheads that add to delay because of the imposed need for collaborative processing and transfer of data before converted into actionable intel. We also explored the building blocks of the semiconductor innovation that will empower the next generation of edge intelligence and enable the new applications like intelligent beamforming-based AR/VR and Layer-0 data-driven intelligent computing.

Wireless edge AI is a mobile-driven, federated architecture that is collaborative in nature, supporting traffic and data transfers at a more decentralized nature. Edge AI systems will cover a wide range in terms of application areas, including data-driven intelligence at the edge, federated learning, model inference acceleration for mobile platforms, automated system design for intelligent health data sensing, 6G-networking with AI-native systems, optimized resource management systems using AI models. With the security issues involved in personal health data transferred to central clouds, keeping the data local for federated learning on personal devices, transfer some compressed data with privacy preservation to local edge servers for localized AI models, and offloading to the federated systems that yield the best prediction accuracy, represent unique challenges.

REFERENCES

[1] Kannan, S., Annapareddy, V. N., Gadi, A. L., Kommaragiri, V. B., & Koppolu, H. K. R. (2023). AI-Driven Optimization of Renewable Energy Systems: Enhancing Grid Efficiency and Smart Mobility Through 5G and 6G Network Integration. Available at SSRN 5205158.

[2] Komaragiri, V. B. The Role of Generative AI in Proactive Community Engagement: Developing Scalable Models for Enhancing Social Responsibility through Technological Innovations.

[3] Paleti, S. (2023). Data-First Finance: Architecting Scalable Data Engineering Pipelines for AI-Powered Risk Intelligence in Banking. Available at SSRN 5221847.

[4] Rao Challa, S. (2023). Revolutionizing Wealth Management: The Role Of AI, Machine Learning, And Big Data In Personalized Financial Services. Educational Administration: Theory and Practice. https://doi.org/10.53555/kuey.v29i4.9966

[5] Yellanki, S. K. (2023). Enhancing Retail Operational Efficiency through Intelligent Inventory Planning and Customer Flow Optimization: A Data-Centric Approach. European Data Science Journal (EDSJ) p-ISSN 3050-9572 en e-ISSN 3050-9580, 1(1).

[6] Mashetty, S. (2023). A Comparative Analysis of Patented Technologies Supporting Mortgage and Housing Finance. Educational Administration: Theory and Practice. https://doi.org/10.53555/kuey.v29i4.9964

[7] Lakkarasu, P., Kaulwar, P. K., Dodda, A., Singireddy, S., & Burugulla, J. K. R. (2023). Innovative Computational Frameworks for Secure Financial Ecosystems: Integrating Intelligent Automation, Risk Analytics, and Digital Infrastructure. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 334-371.

[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] Suura, S. R., Chava, K., Recharla, M., & Chakilam, C. (2023). Evaluating Drug Efficacy and Patient Outcomes in Personalized Medicine: The Role of AI-Enhanced Neuroimaging and Digital Transformation in Biopharmaceutical Services. Journal for ReAttach Therapy and Developmental Diversities, 6, 1892-1904.

[10] Sai Teja Nuka (2023) A Novel Hybrid Algorithm Combining Neural Networks And Genetic Programming For Cloud Resource Management. Frontiers in HealthInforma 6953-6971

[11] Meda, R. (2023). Developing AI-Powered Virtual Color Consultation Tools for Retail and Professional Customers. Journal for ReAttach Therapy and Developmental Diversities. https://doi.org/10.53555/jrtdd.v6i10s(2).3577

[12] 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).

[13] Lakkarasu, P. (2023). Designing Cloud-Native AI Infrastructure: A Framework for High-Performance, Fault-Tolerant, and Compliant Machine Learning Pipelines. Journal for ReAttach Therapy and Developmental Diversities. https://doi.org/10.53555/jrtdd.v6i10s(2).3566

[14] Kaulwar, P. K., Pamisetty, A., Mashetty, S., Adusupalli, B., & Pandiri, L. (2023). Harnessing Intelligent Systems and Secure Digital Infrastructure for Optimizing Housing Finance, Risk Mitigation, and Enterprise Supply Networks. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 372-402.

[15] Malempati, M. (2023). A Data-Driven Framework For Real-Time Fraud Detection In Financial Transactions Using Machine Learning And Big Data Analytics. Available at SSRN 5230220.

[16] Recharla, M. (2023). Next-Generation Medicines for Neurological and Neurodegenerative Disorders: From Discovery to Commercialization. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v10i3.3564

[17] Lahari Pandiri. (2023). Specialty Insurance Analytics: AI Techniques for Niche Market Predictions. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 464-492.

[18] Challa, K. Dynamic Neural Network Architectures for Real-Time Fraud Detection in Digital Payment Systems Using Machine Learning and Generative AI.

[19] Chava, K. (2023). Integrating AI and Big Data in Healthcare: A Scalable Approach to Personalized Medicine. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v10i3.3576

[20] Kalisetty, S., & Singireddy, J. (2023). Optimizing Tax Preparation and Filing Services: A Comparative Study of Traditional Methods and AI Augmented Tax Compliance Frameworks. Available at SSRN 5206185.

[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. (2023). The Role Of Cloud Computing And Big Data In Real-Time Payment Processing And Financial Fraud Detection. Available at SSRN 5236657.

[23] Koppolu, H. K. R. Deep Learning and Agentic AI for Automated Payment Fraud Detection: Enhancing Merchant Services Through Predictive Intelligence.

[24] Sheelam, G. K. (2023). Adaptive AI Workflows for Edge-to-Cloud Processing in Decentralized Mobile Infrastructure. Journal for Reattach Therapy and Development Diversities. https://doi.org/10.53555/jrtdd.v6i10s(2).3570

[25] Kummari, D. N. (2023). AI-Powered Demand Forecasting for Automotive Components: A Multi-Supplier Data Fusion Approach. European Advanced Journal for Emerging Technologies (EAJET)-p-ISSN 3050-9734 en e-ISSN 3050-9742, 1(1).

[26] Suura, S. R., Chava, K., Recharla, M., & Chakilam, C. (2023). Evaluating Drug Efficacy and Patient Outcomes in Personalized Medicine: The Role of AI-Enhanced Neuroimaging and Digital Transformation in Biopharmaceutical Services. Journal for ReAttach Therapy and Developmental Diversities, 6, 1892-1904.

[27] Balaji Adusupalli. (2022). Secure Data Engineering Pipelines For Federated Insurance AI: Balancing Privacy, Speed, And Intelligence. Migration Letters, 19(S8), 1969–1986. Retrieved from https://migrationletters.com/index.php/ml/article/view/11850

[28] Pamisetty, A. (2023). AI Powered Predictive Analytics in Digital Banking and Finance: A Deep Dive into Risk Detection, Fraud Prevention, and Customer Experience Management. Fraud Prevention, and Customer Experience Management (December 11, 2023).

[29] 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.

DOI: 10.17148/IJARCCE.2022.111258

[30] Dodda, A. (2023). AI Governance and Security in Fintech: Ensuring Trust in Generative and Agentic AI Systems. American Advanced Journal for Emerging Disciplinaries (AAJED) ISSN: 3067-4190, 1(1).

[31] 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

[32] Pamisetty, A. Optimizing National Food Service Supply Chains through Big Data Engineering and Cloud-Native Infrastructure.

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

[34] Chakilam, C. (2022). Integrating Machine Learning and Big Data Analytics to Transform Patient Outcomes in Chronic Disease Management. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v9i3.3568

[35] 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

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

[37] 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).

[38] Malempati, M., Pandiri, L., Paleti, S., & Singireddy, J. (2023). Transforming Financial And Insurance Ecosystems Through Intelligent Automation, Secure Digital Infrastructure, And Advanced Risk Management Strategies. Jeevani, Transforming Financial And Insurance Ecosystems Through Intelligent Automation, Secure Digital Infrastructure, And Advanced Risk Management Strategies (December 03, 2023).

[39] 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

[40] Challa, K. (2023). Optimizing Financial Forecasting Using Cloud Based Machine Learning Models. Journal for ReAttach Therapy and Developmental Diversities. https://doi.org/10.53555/jrtdd.v6i10s(2).3565

[41] Pandiri, L., Paleti, S., Kaulwar, P. K., Malempati, M., & Singireddy, J. (2023). Transforming Financial And Insurance Ecosystems Through Intelligent Automation, Secure Digital Infrastructure, And Advanced Risk Management Strategies. Educational Administration: Theory and Practice, 29 (4), 4777–4793.

[42] Recharla, M., & Chitta, S. AI-Enhanced Neuroimaging and Deep Learning-Based Early Diagnosis of Multiple Sclerosis and Alzheimer's.

[43] 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).

[44] 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.

[45] Lakkarasu, P. (2023). Generative AI in Financial Intelligence: Unraveling its Potential in Risk Assessment and Compliance. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 241-273.

[46] 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.

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

[48] 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.

[49] 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

[50] Kannan, S. The Convergence of AI, Machine Learning, and Neural Networks in Precision Agriculture: Generative AI as a Catalyst for Future Food Systems.

[51] 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

[52] Singireddy, S. (2023). AI-Driven Fraud Detection in Homeowners and Renters Insurance Claims. Journal for Reattach Therapy and Development Diversities. https://doi.org/10.53555/jrtdd.v6i10s(2).3569

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

[54] Rao Challa, S. (2023). Artificial Intelligence and Big Data in Finance: Enhancing Investment Strategies and Client Insights in Wealth Management. International Journal of Science and Research (IJSR), 12(12), 2230–2246. https://doi.org/10.21275/sr231215165201

[55] Paleti, S. (2023). Trust Layers: AI-Augmented Multi-Layer Risk Compliance Engines for Next-Gen Banking Infrastructure. Available at SSRN 5221895.

[56] 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).

[57] Komaragiri, V. B. (2023). Leveraging Artificial Intelligence to Improve Quality of Service in Next-Generation Broadband Networks. Journal for ReAttach Therapy and Developmental Diversities. https://doi.org/10.53555/jrtdd.v6i10s(2).3571

[58] 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.

[59] 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.

[60] 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

[61] Vamsee Pamisetty. (2020). Optimizing Tax Compliance and Fraud Prevention through Intelligent Systems: The Role of Technology in Public Finance Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 111–127. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11582

[62] Paleti, S. (2023). AI-Driven Innovations in Banking: Enhancing Risk Compliance through Advanced Data Engineering. Available at SSRN 5244840.

[63] 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

[64] 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

[65] 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

[66] Singireddy, S. (2023). Reinforcement Learning Approaches for Pricing Condo Insurance Policies. American Journal of Analytics and Artificial Intelligence (ajaai) with ISSN 3067-283X, 1(1).

[67] 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

[68] 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.

[69] 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

[70] 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.

[71] 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

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

[73] 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).

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

[75] 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

[76] Chava, K. (2020). Machine Learning in Modern Healthcare: Leveraging Big Data for Early Disease Detection and Patient Monitoring. International Journal of Science and Research (IJSR), 9(12), 1899–1910. https://doi.org/10.21275/sr201212164722

[77] 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

[78] Dwaraka Nath Kummari,. (2022). Machine Learning Approaches to Real-Time Quality Control in Automotive Assembly Lines. Mathematical Statistician and Engineering Applications, 71(4), 16801–16820. Retrieved from https://philstat.org/index.php/MSEA/article/view/2972

[79] 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

[80] Adusupalli, B. (2023). DevOps-Enabled Tax Intelligence: A Scalable Architecture for Real-Time Compliance in Insurance Advisory. Journal for Reattach Therapy and Development Diversities. Green Publication. https://doi. org/10.53555/jrtdd. v6i10s (2), 358.

[81] Pamisetty, A. (2023). Cloud-Driven Transformation Of Banking Supply Chain Analytics Using Big Data Frameworks. Available at SSRN 5237927.

[82] 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.

[83] 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

[84] 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

[85] 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.

[86] 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

[87] 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

[88] 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

[89] Kannan, S., & Saradhi, K. S. Generative AI in Technical Support Systems: Enhancing Problem Resolution Efficiency Through AIDriven Learning and Adaptation Models.

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

[91] 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://www.philstat.org/index.php/MSEA/article/view/2977

[92] 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

[93] 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.