A community initiative, anchored by

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

25 IIIT Hyderabad

TechForward

DISPATCH

JUNE EDITION

AI ON THE EDGE

is making technology smarter, more responsive, and accessible. In academia, researchers are exploring cutting-edge ideas in AI that don't always require powerful computers or internet connections. In industry as well, AI on the edge is revolutionising many fields, helping companies create smart devices that work faster and more efficiently as they process information right at source. The first edition of TechForward Dispatch delves into the scope of AI on the edge and its use-cases in academia and industry.

IIITH's TechForward research seminar series is an academia-industry confluence around emerging technologies. The deep insights, directional talks and industry outlooks from accomplished thought leaders at the seminar are compiled monthly in the Tech Dispatch as a ready reckoner for technology directions.

From the Chair's Desk

Welcome to the Tech Forward Research Seminar series. This session is the first among many planned every month. The idea is to bring industry and academia together to have a conversation around cutting-edge technology and applications across various verticals. This is a great platform for industry and academia to know the latest on academic research as well as applications across industry.

The topic this month is closely related to the biggest momentum seen in the tech sector over the last one and half years. AI and more specifically Gen AI is becoming mainstream and will disrupt pretty much every sector imaginable. The focus over the last 1.5 years has been on large language models and training with billions of parameters on the server/cloud. The next disruption expected is on the edge – where end devices such as mobiles, laptops, connected cameras, routers, cars and other IoT devices will use AI not only for training but mostly for inferencing using already trained and optimized models.

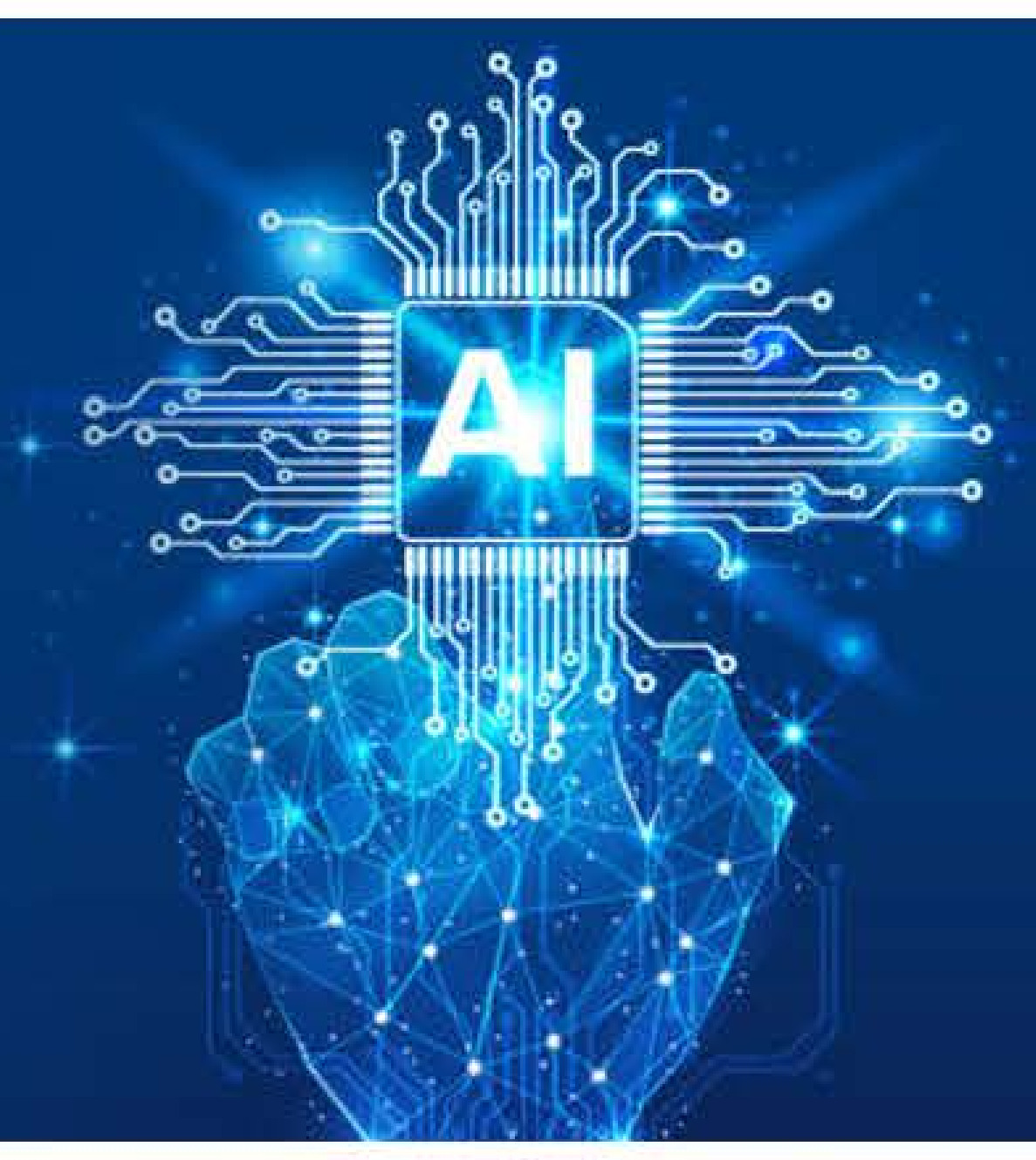
To cover more details on this topic of "AI on the Edge", we had two distinguished speakers - Rajesh Narayanan, VP of Engineering from Qualcomm and Prof Anoop Namboodiri from IIITH. Their talks have been summarized in the Dispatch along with insights from other thought leaders in the field.

RAJEEV KURUNDKAR
VP Engineering, Qualcomm India Pvt Ltd

CONTENTS OF THIS EDITION

The 1st Edition of TechForward was hosted at Qualcomm campus

How Computer Vision Is Driving Innovation In Edge Devices



***Prof. Anoop Namboodiri** provides perspectives of general research trends in Computer Vision and those taking place at IIITH in particular to explain how they are propelling advancements on the Edge.*

Computer Vision on the edge may sound like a new development but the truth is that it has been around for decades. Some applications that we take for granted include defect detection in factories, mobile phone unlocking using either fingerprint or face recognition, x-ray machines in airports, QR codes, and licence plate recognition at toll gates. Thanks to AI on edge, we now have far more powerful applications on the edge such as autonomous navigation and home automation devices like Amazon Echo having a built-in camera. With improved AI algorithms, newer and more powerful use cases are being enabled.

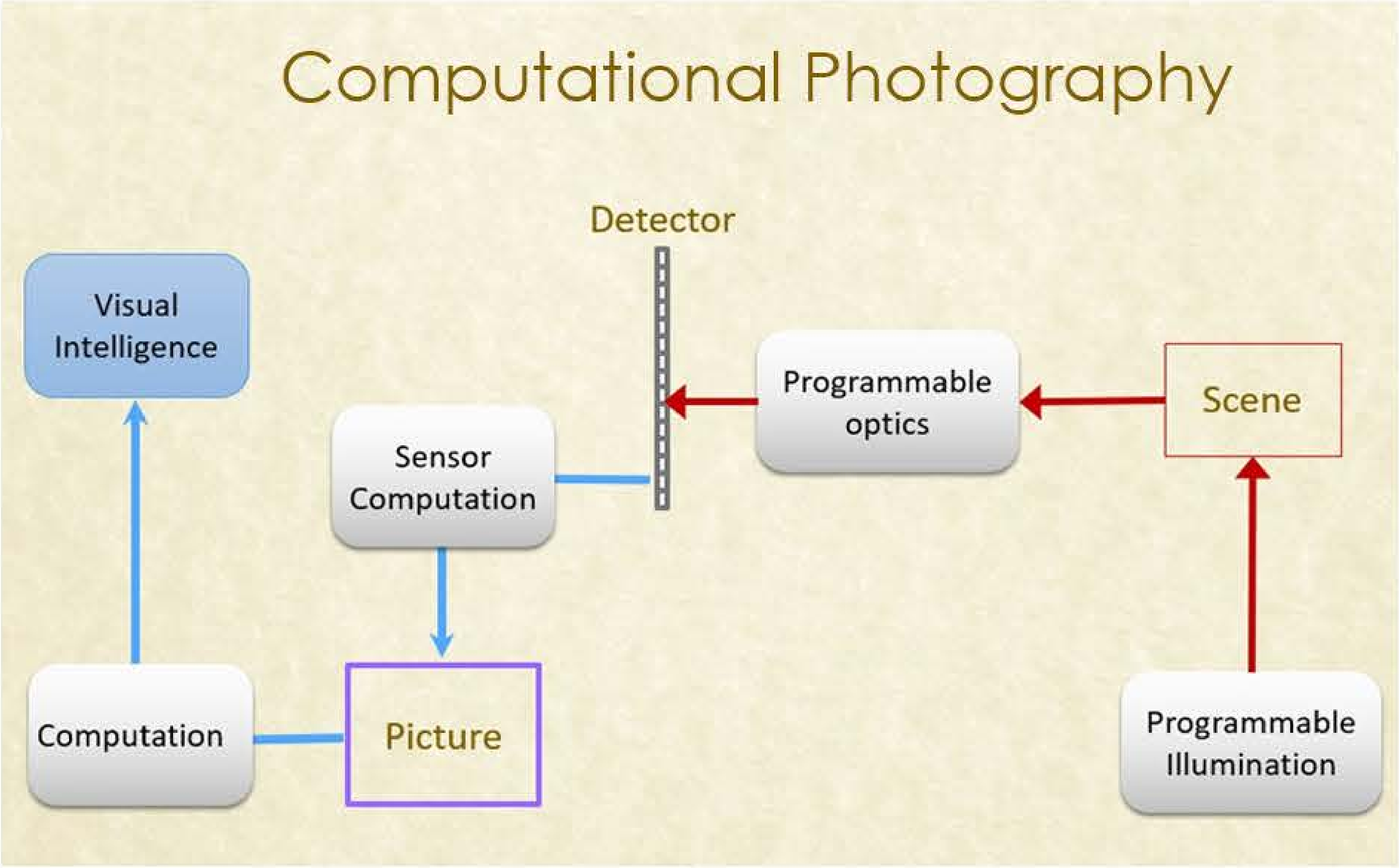
Computer Vision itself has played a pivotal role in pushing the cutting edge of AI. Examples of this include CNNs, Gen AI and multimodal foundational models. In future, when AGI systems become viable, we can be sure that computer vision and 3D understanding of the world will play an important role in it. Here, we focus on the recent trends that are pushing AI to edge devices.

Computer Vision, AI and Edge: Computer Vision involves processing of large amounts of visual data that could be in the form of images and/or videos. In order to make sense of all this data, it is imperative that we are able to do a significant amount of processing on the edge. Hence computer vision is an area that can significantly benefit from processing on the edge. There are 3 ways in which we can look at the interplay between Vision, AI and the edge:

- What innovations in Vision can allow us to put AI on the edge?
- What advancements in AI can help us push Computer Vision to the edge?
- What improvements in edge devices will allow us to utilise AI and Vision?

Innovations in Vision: Computer vision enables a machine to look at the world the way we look at it, i.e., perceive the world. But to understand where we are today in the field, it is useful to see what we started off with. The very first ancestor of the photographic camera was the pinhole camera, which was accidentally discovered a long time ago and can be found in many ancient monuments. It is essentially a lensless 'camera' or a dark room with a small hole in one wall. It was Leonardo Da Vinci who first compared the human eye to this 'camera obscura' and it was used as a model to explain human vision for centuries. From this stage, cameras developed by incorporating lenses and mirrors, recording the images on a silver plate, development of color and chrome films and finally computer vision became possible with digitization of photographs.

With digital cameras becoming popular at the turn of the century, they allowed us to process the data from the sensors before storing as an image. This also opened up a significant change in the imaging pipeline as the data from the sensors need not be close to the final image as long as it can be processed to obtain one. This field is referred to as **Computational Photography**. Simple examples of this include the projection of structured light patterns on an object to capture its 3D shape, which can then be used in fine grained recognition (say Hand Geometry-based person authentication).


Computational Photography

Stereo Vision

A more complex problem is that of capturing a panoramic stereo imaging. i.e., to capture the left and right eye views when we turn our head around from a point. Rotating a stereo pair will work only on static scenes and using a set of cameras in a circle will cause significant occlusions. One of the solutions proposed by Google and FB was to use a set of cameras in a larger circle facing radially outwards and use parts of these images to stitch together the left and right eye views. However, the processing required was of the order of minutes on a compute cluster for generating a single stereo frame.

Innovations at IIITH

At IIITH, we worked on a solution that uses catadioptrics (mixture of lenses and mirrors) to capture the relevant light rays from a location as close as possible to the rotating stereo camera pair. This innovation in computer vision allows us to reduce the processing significantly and we can now move the stereo image composition to the edge and do so at 30FPS. The optics-based solution also allows us to avoid any blind spots that are inherent to multi-camera solutions.

The Journey from Lab to Land

- Mirror Design, Moulding, Coating
- Enclosure, Material, Coating
- Camera Design
- Sensor, Compute, Interface
- Manufacturing, Testing
- Intelligence, Testing
- Software, Optimization

3D Reconstruction

With advances in the capabilities of edge processors, we can now run compact deep-learning models on the edge. This would allow us to add additional functionalities like depth estimation, semantic segmentation, and navigation on the edge. One can also integrate edge processing, which is highly data intensive, along with server-based processing of tasks that are compute intensive. An example of such processing in this case would be 3D reconstruction of the world from multiple images.

Improving Deep Learning Models

The second factor that has contributed to the migration of vision to the edge is the improvement in efficient deep-learning models. One possible way to improve the efficiency is to represent all (or most) parameters of the network with fewer bits (quantization). Some of the work that was done in our lab in this direction include improvements in binary quantization that takes into account the distribution of parameter values, and ternary quantization, which integrates binarization and pruning into a single framework and that optimises the whole network.

Using Expander Graphs

Most strategies for pruning involve training the full-sized network to identify the weights that can be pruned. After removing smaller weights, the pruned network is trained further. This process is repeated until we reach the desired pruning level or until the error rate reaches its maximum allowed limit. A drawback of such training is that it is very compute-intensive and can be done only on dedicated high-end servers. We tried to pre-prune the network and then train it so that the training process can become more efficient. For this we had to bring in the concept of expander graphs from graph theory. Essentially combining knowledge from theoretical computer science with AI, which in turn is improving computer vision that can be deployed on the edge.

We have also done some work on modelling the performance of deep-learning models on the edge. This allows us to predict the maximum number of parameters of a particular model, given a performance target. This allows one to determine the best models that can be run on a given edge hardware, train them, and compare the resulting accuracies.

Securing Biometrics

Another interesting aspect of computer vision on edge-computing is the fact that these devices can now detect and recognize humans in their vicinity. This poses both privacy and security challenges. However, the availability of computational capability on the edge also allows us to deploy secure multiparty computational techniques to improve the security and privacy of biometric algorithms. Other capabilities that can be deployed on the edge in this context include biometric spoof detection.

Analysis of Edge Device Liability

We also created an autoencoder network that models the physical image transformations at the edge. We then use this model to analyse the vulnerabilities of the edge devices, specifically for the purpose of presentation attacks (PA). Essentially we showed that it is possible to bypass the PA detection. With the help of such a model, one can do hill climbing or gradient attack and fool most of the presentation attack detection systems with over 80% success rate.

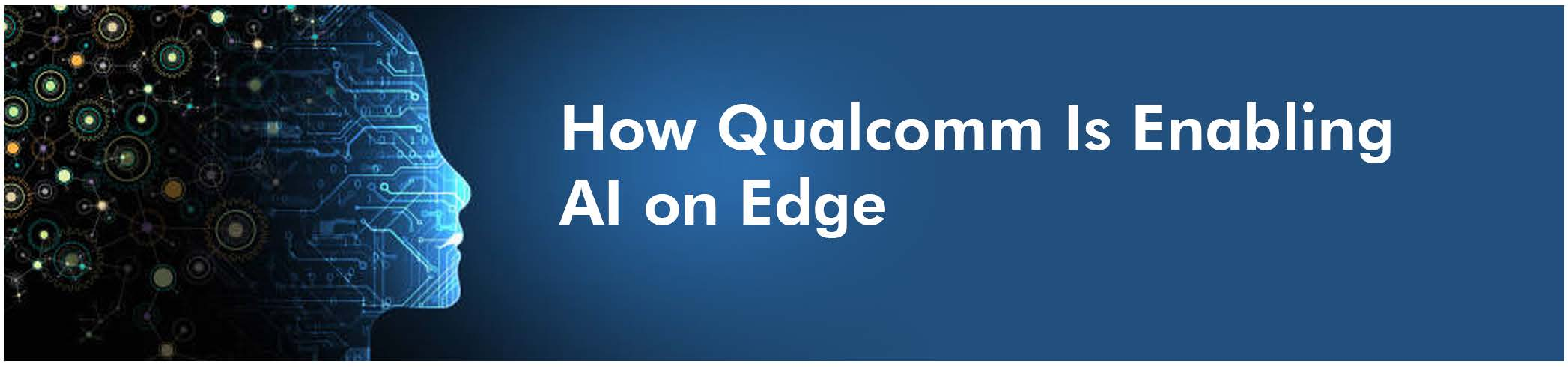
In summary

We can see that the fields of computer vision, AI and edge computing are highly interlinked and improvements in one field can affect the others in a mutually beneficial manner. This symbiotic growth is also enabling a variety of applications on the edge. The resulting model of distributed computing and learning will become extremely important in the near future.

PROF. ANOOP NAMBOODIRI

IIITH

is a faculty member at the Centre for Visual Information Technology (CVIT) in IIIT Hyderabad. He works in the areas of Computer Vision, Machine Learning and Biometrics. His work on Computational Photography has resulted in the first-of-its-kind solution for capturing stereo panoramic videos with a single sensor. The solution is now productised by DreamVu Inc., for whom he serves as the Chief Science Officer. Anoop has also worked closely with the Aadhaar project and developed several biometric solutions for them.



How Qualcomm Is Enabling AI on Edge

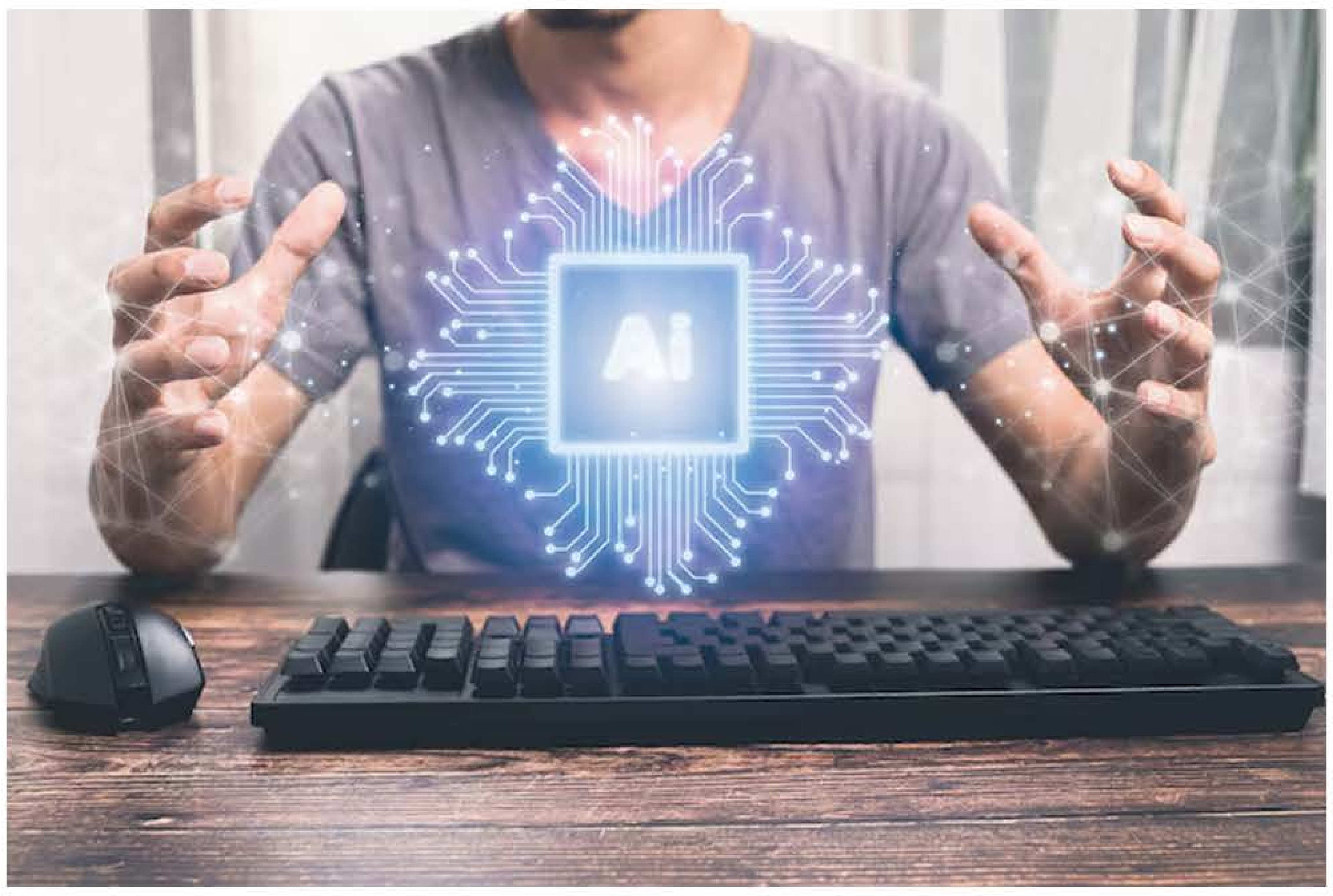
Rajesh Narayanan enumerates all of Qualcomm's research efforts to enhance AI user experiences especially with regard to on-device AI.

We envision a world where devices, machines, automobiles, and things are much more intelligent, simplifying and enriching our daily lives. They will be able to perceive, reason, and take intelligent actions based on awareness of the situation, improving just about any experience, and even solving problems considered unsolvable. Artificial intelligence (AI) is the technology driving this revolution.

If you look at the last 10 years or so, AI has gradually evolved from what was very focused on image processing with the original Convolutional Neural Network (CNN). Over the last four to five years, AI started to make a difference in terms of language, speech, text processing etc. AI is enhancing our lives and is being used all around us in many everyday tasks. It is

- providing entertainment (smartphones, sports, TVs, video games)
- enhancing collaboration (video conferencing, extended reality applications)
- transforming industries (autonomous vehicles, smart factories, smart medicine, smart inspection and so on).

The need for intelligent, personalized experiences powered by AI is ever-growing. To make this a reality, we need to bring human-like understanding and behaviors to devices and other things through AI.



On-device AI is not new for us. For more than a decade, Qualcomm Technologies has been researching and working with customers, including original equipment manufacturers and application developers, to enhance the user experience through AI. On-device AI support to generative AI through optimized and/or specialized neural network models can further enhance the user experience through increased privacy and security, performance, and personalization while lowering the required costs and energy consumption.

What is Qualcomm doing to enable on-device AI
The opportunity for on-device intelligence is clear. There's a very large amount of research that we've been doing in terms of AI algorithms and all the other models that are being developed. After brainstorming on the best way for us to take all of that, curate it and make sure that it runs in the most effective manner in an end-consumer device, we're focusing on high performance hardware, software and optimized network design to make on-device intelligence pervasive.

Power and thermal efficiency are essential for on-device AI, but it is a big challenge since the AI workloads are compute-intensive and must work in a constrained mobile environment.
We are looking at three key hurdles for deployment of on-device AI:

- Low power devices which leverage heterogeneous compute to process the workload on the right hardware accelerator.
- Algorithmic advancement focused on embedded, low power operation to map state of the art deep neural networks onto these devices.
- Software runtimes and tools for execution and optimization of networks on embedded devices which reduce the burden of porting networks to these devices and unlock built-in optimizations.

Heterogeneous Compute

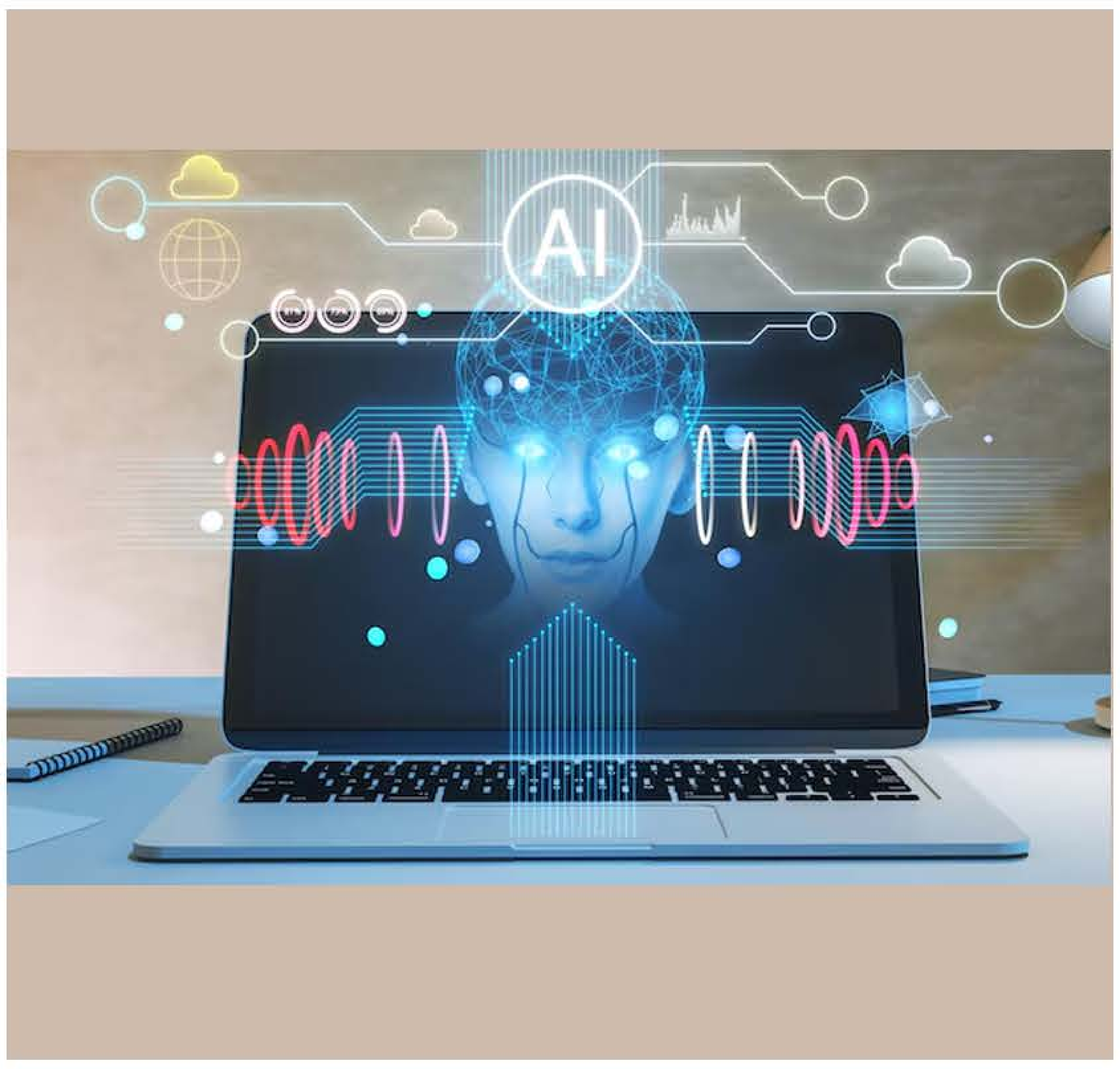
The truth is that it's hard to build a single processing engine that excels at everything. Qualcomm has our own portfolio of solutions with a complete notion of the neural processing unit (NPU), central processing unit (CPU) and graphics processing unit (GPU) all running in conjunction, with simultaneous usage of these processors. Having diverse processing engines gives us more opportunity that at least one of them will be truly excellent in running a particular workload.

For example,

- The CPU is excellent at complex sequential control flow such as in game artificial intelligence,
- The GPU is ideal for throughput processing and parallel number crunching, such as in image processing.
- The NPU combines features like scalar, vector, and tensor instructions with control flow agility to do sustained and high AI peak performance at low power for LLM and LVM processing.

So, specialization and diversity are key for edge AI devices. The Qualcomm AI Engine includes our latest and greatest CPU, Adreno GPU, Hexagon NPU, Sensing Hub and system memory. Traditionally, the CPU was responsible for executing AI algorithms. As the demands for processing performance skyrocketed, dedicated NPUs emerged as a specialized solution for handling software and applications leveraging AI. These processors are designed to efficiently handle the complex mathematical computations required for AI tasks, offering unparalleled efficiency, performance and power savings.

At the heart of Hexagon NPU, lies some foundational building blocks to support Scalar, Vector, and Tensor based AI operations that are nicely mapped to a variety of AI workloads. Additionally, the architecture also enables dedicated local memory module to limit data transfer to DRAM and thereby conserving precious power.



Algorithmic Advancements

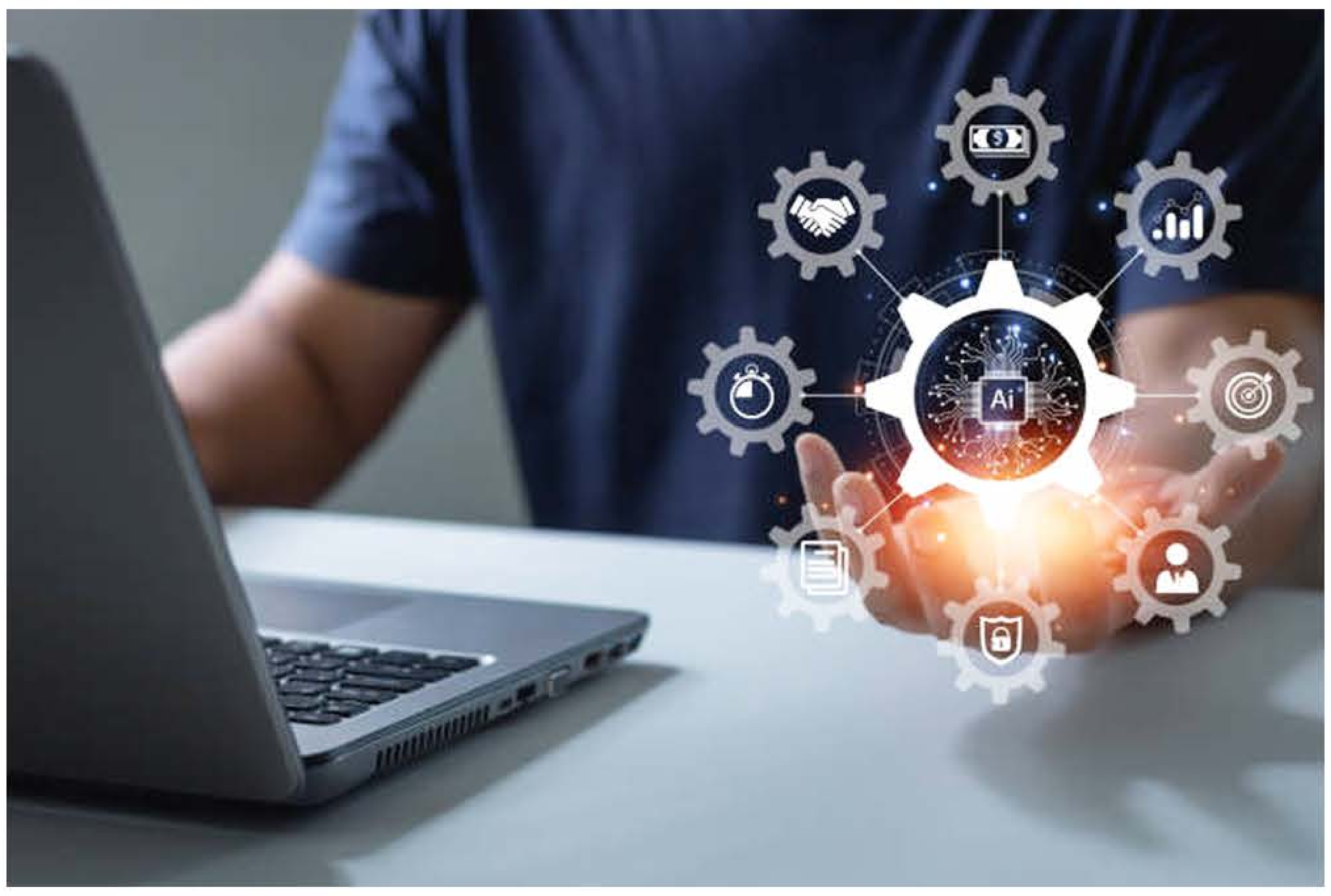
At Qualcomm, we are strong believers that having a vision and investing early in R&D are essential to leading the ecosystem forward. It often involves taking a holistic system approach, where we set out to solve big problems. When we talk about full-stack AI research, we mean taking theoretical research and proving it in the real world on real devices. Making research work outside of the lab often requires optimizations across the neural network model, the software, and the hardware, as well as working across disciplines within the company and sometimes externally. There are many layers of hardware and software that can be tuned and optimized to squeeze out every bit of performance at the lowest energy uses.

Once we have proven the technology, we can enable the ecosystem toward rapid commercialization at scale.

We have invested efforts into AI model optimization for power efficiency and performance because we believe this is what is going to allow AI to truly scale and become ubiquitous. In order to squeeze every ounce of efficiency out of AI models that have already been optimized by the industry, we begin with optimized models like MobileNet or optimized models output from frameworks like TensorFlow Lite. We are applying AI, such as reinforcement learning, across multiple techniques, specifically compression, quantization, and compilation, to shrink models and improve HW performance. Automation is key since hand-tuning is not feasible for many of these deep AI models and our AIMET (AI Model efficiency Tool Kit), enables this. These techniques are being made hardware-aware to further squeeze out efficiency.

Software runtime and tools

We created a unified AI software stack offering that combines all our AI software capabilities and tools into an integrated solution. Qualcomm® AI Stack is a unified AI software portfolio for our mobile, automotive, XR, compute, IoT and cloud platforms. It supports popular frameworks all the way down to accessing and accelerating the "metal". Our "Develop once, deploy anywhere" approach ensures our users can develop one feature or application, then move the same model across different products and tiers.

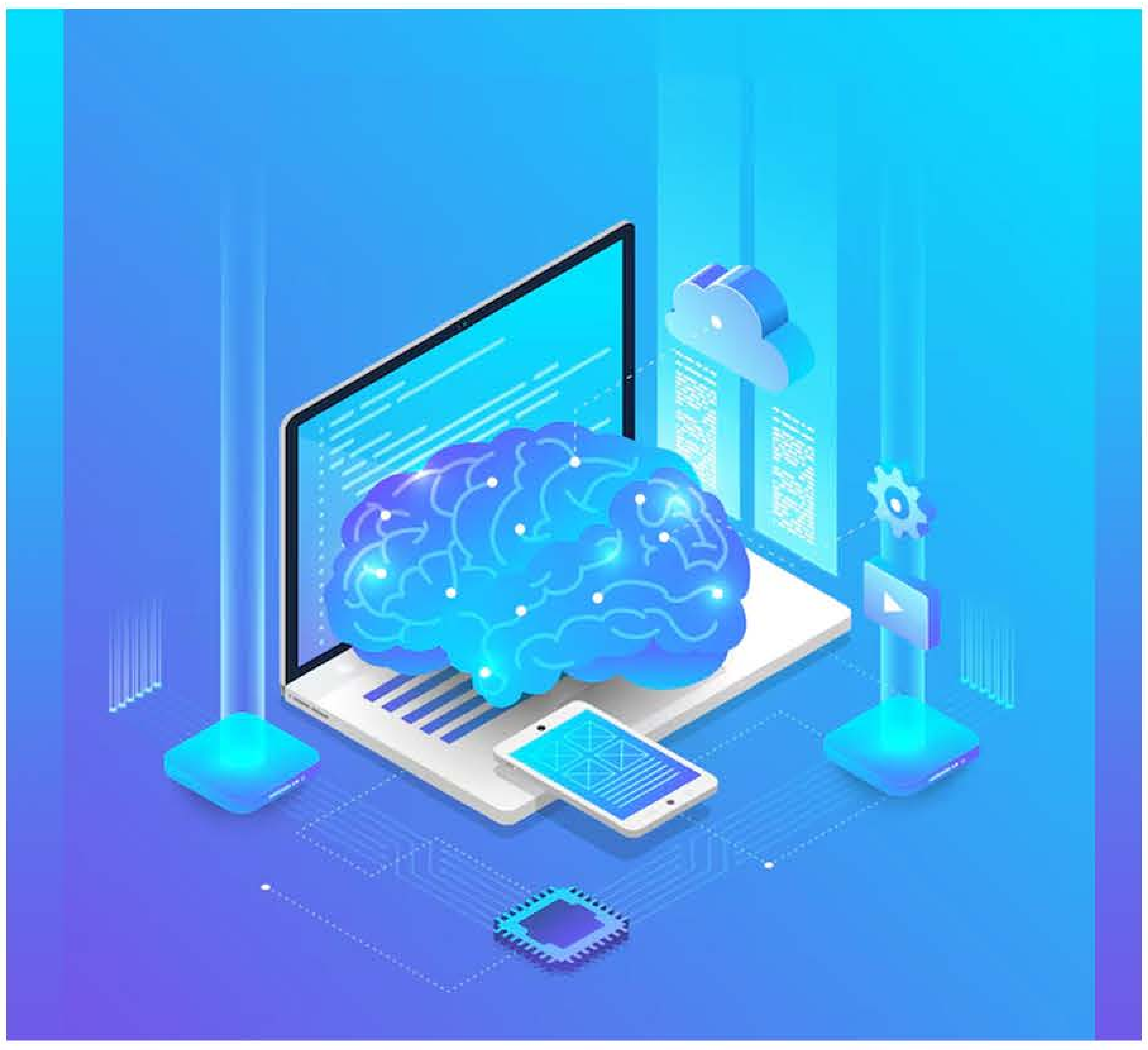


Gen AI Trends and enablement on the edge

A lot of us are using GenAI based solutions on our mobile and edge devices but aren't aware that the brain behind it is Gen AI. Like, text-based processing, text to image, text to voice, and so on. Also, productivity use cases on your laptop such as Teams calls after meeting being able to generate meeting minutes are based on GenAI. With an installed base of billions of AI-capable phones, PCs, and other devices in users' hands today, the potential to tap into a large amount of on-device AI processing for generative AI is already significant – and poised to grow steadily in the coming years. A key question becomes which generative AI models can run on device with appropriate performance and accuracy. The great news is that very capable generative AI models are getting smaller while on-device processing capabilities continue to improve. There's a broad number of generative AI capabilities that can run on device using models that range from 1 to 10 billion parameters. We aim to optimize generative AI models and efficiently run them on hardware through techniques such as quantization, microtile inferencing and heterogeneous computing.

Qualcomm AI Hub: Developer's Gateway

The Qualcomm AI Hub is a developer-first online platform that simplifies the process and streamlines on-device AI development for Snapdragon and Qualcomm platforms. It is a central location for AI developers to access resources aimed at Snapdragon platforms. This developer's gateway to superior on-device AI performance contains a library of pre-optimized AI models for seamless deployment on devices powered by Snapdragon and Qualcomm platforms. The library provides developers with more than 75 popular AI and generative AI models, such as Whisper, ControlNet, Stable Diffusion, and Baichuan 7B, which are optimized for superior on-device AI performance, lower memory utilization, and better power efficiency, across different form factors and packaged in various runtimes. Each model is optimized to take advantage of hardware acceleration across all cores within the Qualcomm® AI Engine (NPU, CPU, and GPU) resulting in faster inferencing times. The AI model library automatically handles model translation from source framework to popular runtimes and works directly with the Qualcomm® AI Engine direct SDK, then applies hardware-aware optimizations. Developers can seamlessly integrate these models into their applications, reducing time-to-market, and unlocking the benefits of on-device AI implementations such as performance, reliability, privacy, personalization, and cost savings.



Conclusion

It is an exciting and dynamic time as the adoption of on-device generative artificial intelligence grows, driving the democratization of AI on consumer devices to further enhance user experiences. Qualcomm is advancing AI research to make on-device AI ubiquitous and to make AI power efficient, we are creating AI platform innovations that are fundamental to scaling AI across the industry.

RAJESH NARAYANAN

QUALCOMM INDIA PVT LTD

is VP, Engineering at Qualcomm India Pvt Ltd., leading the Application Processor(AP) SW Technologies group covering Multimedia-SW, Multimedia-Systems, AI-SW and Linux. He and his team handle AP SW design, development and commercialization for Qualcomm SoCs catering to different business segments including Smartphone/Mobile, XR, Compute, IOT etc.

Optimization Techniques For Edge AI

In a Q&A about the emergence of edge computing, Dr. Suresh Purini explains its scope, the challenges confronting Edge ML, and briefly describes the active research on edge that is underway at IIITH.

Why has ML on the Edge emerged?
Many applications from domains such as agriculture, health care, retail stores, Industry 4.0, and so on require intelligence on the edge. Real-time response, privacy, intermittent connectivity are some important factors among others which have necessitated this move from the cloud to edge. Further, by moving computation closer to data, rather than vice-versa, we can save on the available network bandwidth and the power consumption from the communication modules. Also, such distributed computations reduce the compute capacity requirements of centralised data centers thus saving on both capital and operational expenses such as a real estate requirement, power consumption, etc. However, deploying modern complex machine learning (ML) models on edge devices which are both compute and memory heavy is fraught with many challenges.

What are some of the key challenges of edge computing?
Edge devices often have limited computational power, memory, and storage compared to centralised servers. Due to this, the cloud-centric ML models need to be retargeted so that they fit in the available resource budget. Further, many edge devices run on batteries, making energy efficiency a critical consideration. The hardware diversity in edge devices ranging from microcontrollers to powerful edge servers, each with different capabilities and architectures requires different model refinement and retargeting strategies. As the field of machine learning evolves at a rapid pace, it becomes increasingly challenging for hardware accelerator designers and the associated software stack, including compilers and runtime systems, to keep up with and efficiently support the latest state-of-the-art models.

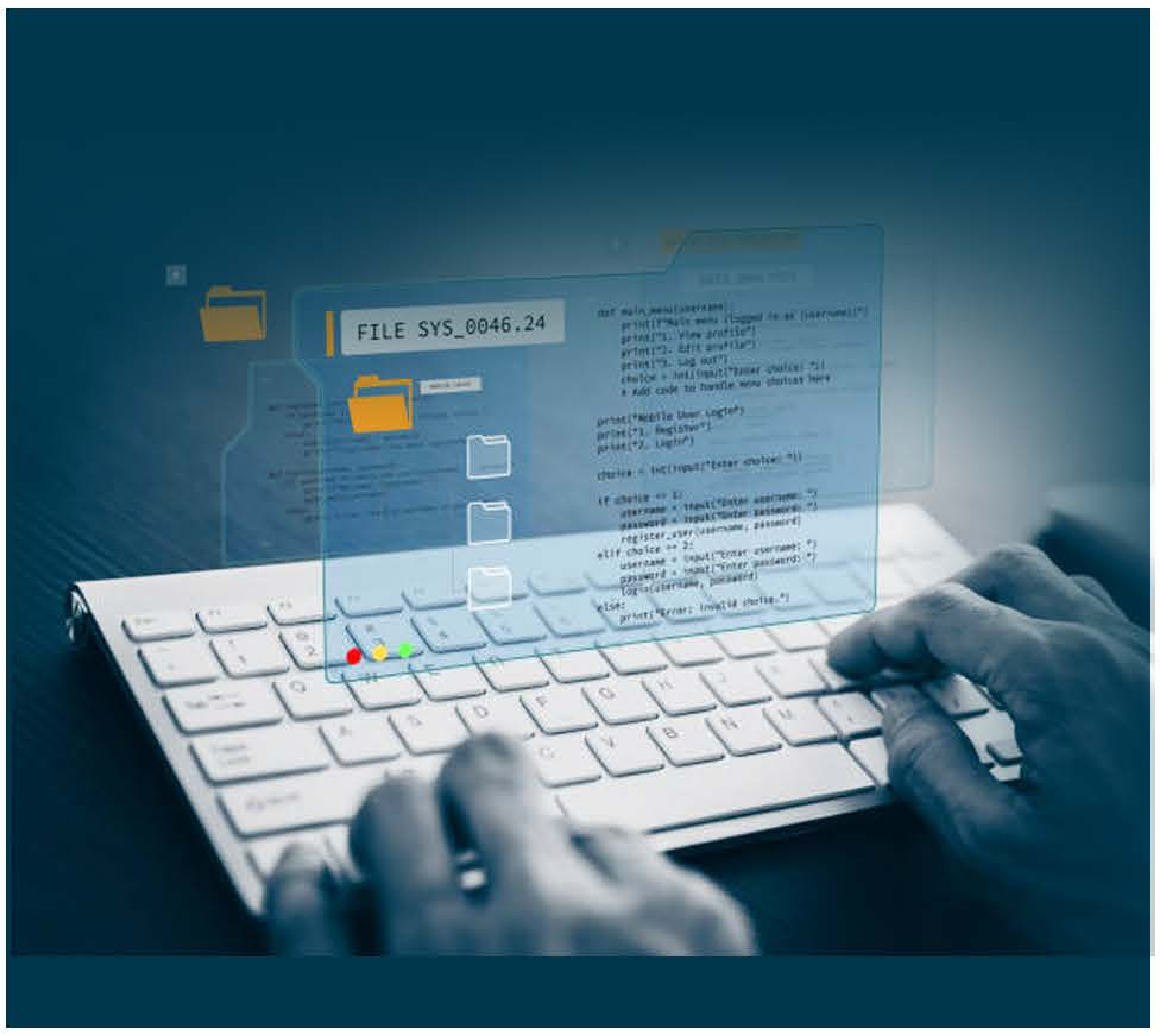
What are the typical strategies used to overcome the challenges?
Designing power-efficient hardware accelerators for ML inferencing tasks is an active area of research in both industry and academia. Some of the industry solutions include the NVIDIA Jetson Nano, Google Coral, Intel Movidius, Raspberry Pi, Qualcomm QCS605, and ARM Neoverse. However there are software tools and techniques too that are used to retarget machine learning models for high-compute and memory-constrained edge devices operating within a power envelope. While porting models to the edge, a typical trade-off arises between latency and accuracy. Higher accuracy requirements often lead to increased model complexity and, consequently greater latency, and the reverse is also true. Some of the well-known and emerging techniques that help in managing these accuracy-latency trade-offs are:

Model compression which involves transforming cloud-centric machine learning models to be edge-friendly using various techniques such as architecture simplification, use of lightweight layers, pruning, quantization, and knowledge distillation. For example, by employing these methods, the popular object detection model YOLOv5 has been optimized into edge-friendly variants like YOLOv5s (small) and YOLOv5n (nano).

Dynamic Model Selection: While techniques that are used for model compression focus on optimising single models, we can envision a model repository for a given task with varying accuracy-latency trade-offs under different conditions. Depending on the ambient conditions, we can dynamically switch between high and low complexity models as needed. For instance, under rainy conditions, we might switch to a more complex object detection model from a lighter one that performs well in clear conditions. A meta-model, performing cost-benefit analysis of model switching—possibly using reinforcement learning—can decide whether to switch and which replacement model to use. This type of model switching requires built-in support from the MLOps stack.

Distributed Inference: Many use cases involve the distributed deployment of numerous IoT or edge devices, such as CCTV cameras, working collaboratively towards specific objectives. These applications often have built-in redundancy, making them tolerant to failures, malfunctions, or less accurate inference results from a subset of edge devices. Algorithms can be employed to recover from missing, incorrect, or less accurate inputs by utilising the global information available. This approach allows for the combination of high and low accuracy models to optimise resource costs while maintaining the required global accuracy through the available redundancy.

Online Training and Refinement: Most of the discussion until now focuses on use cases where a model is trained at a centralised location and then retargeted for edge devices. However, some applications require that a deployed edge model be further refined based on ambient conditions using online training techniques. In such cases, an edge device can periodically send a sample of input data to a master model on the cloud and use the output to refine itself. This approach is akin to a distributed knowledge-distillation process.



Federated Learning: Large-scale deployment of distributed edge devices is inherently suited to federated learning, where a centralised master model in the cloud is trained using locally adapted models from the edge devices. By communicating model weights instead of raw data, power and communication bandwidth are conserved, and data privacy is preserved. However, this too faces numerous challenges, including device heterogeneity, intermittent network connectivity, bandwidth limitations, power constraints, privacy risks, and security concerns. Addressing these challenges requires substantial research in privacy-preserving methods, energy-efficient algorithms, and secure systems.

Can you briefly describe IIITH's strides in edge AI research?
Several research groups at IIITH are exploring various aspects of edge AI research, with some of these initiatives falling under the broader Smart City project.

Digitising Water Meters: Prof. Sachin Chaudhari and his team has developed an IoT-based economic retrofitting setup for digitising the analog water meters to make them smart. The setup contains a Raspberry-Pi microcontroller and a Pi-camera mounted on top of the analog water meter to take its images. The captured images are then preprocessed to estimate readings using ML/DL models.

Air Pollution Monitoring using Images: Prof. Sachin Chaudhari's efforts have led to an IoT-based real-time air quality index (AQI) estimation technique that uses images and weather sensors on Indian roads. A mixture of image features, i.e., traffic density, visibility, and sensor features, i.e., temperature and humidity, were used to predict the AQI. Object detection and localization-based Deep Learning (DL) method along with image processing techniques were used to extract image features while a Machine Learning (ML) model was trained on those features to estimate the AQI.

Model Balancers: Earlier, we discussed the concept of dynamic model selection to enhance resource efficiency. Similarly, a self-adaptive system that switches between different ML models is being developed, deployed, and successfully demonstrated on Qualcomm Edge Devices by Prof. Karthik Vaidhyanathan and his team.

Large Scale Distributed CCTV Camera Analytics: In this work, our team built a scalable distributed video analytics framework that can process thousands of video streams from sources such as CCTV cameras using semantic scene analysis. The main idea is to deploy deep learning pipelines on the fog nodes and generate semantic scene description records (SDRs) of video feeds from the associated CCTV cameras. These SDRs are transmitted to the cloud instead of video frames saving on network bandwidth. Using these SDRs stored on the cloud database, we can answer many complex queries and perform rich video analytics within extremely low latencies. There is no need to scan and process the video streams again on a per query basis. The software architecture on the fog nodes allows for integrating new deep learning pipelines dynamically into the existing system, thereby supporting novel analytics and queries.

Model Selection and Placement: In large-scale deployments of distributed edge devices, it is not always necessary to process data streams on all devices using highly complex and accurate models. We tried another approach where we can strategically deploy models with varying complexity and accuracy by exploiting the redundancy across data streams, thereby leveraging domain-specific meta information to achieve the same quality of analytics. This technique has significant practical applications and presents numerous research opportunities.

Overall, we are witnessing the advent of the edge AI/ML era, which is transforming the way we see and experience the world as intelligent edge devices bridge the human-machine continuum. However, realising this potential requires numerous innovations in hardware, system software, and algorithm optimization for machine learning, as well as scalable distributed systems and algorithms.

DR. SURESH PURINI
IIITH

is an Associate Professor of Computer Science at IIIT-Hyderabad. He leads the Computer Systems Group. He has wide research interests spanning compilers, architecture, parallel and distributed systems, and most recently Systems for AI/ML. He practises Heartfulness Meditation, and is a certified yoga and meditation trainer.

Self-adaptive EdgeML With Model Balancer

Dr. Karthik Vaidhyanathan explains a novel approach that allows a system on edge to self-adapt between different ML models, thus achieving better resource efficiency.

Computing and AI Through The Ages

Lately, leaps in technological progress have meant that AI is not only popular but is now democratised. Simplistically put, it means that applications like ChatGPT are available or easily accessible to all - even those without specialised tech skills. Historically if one were to look at the field of AI, it has existed since the 1940s and 50s. The obvious limitation then was the lack of sophisticated hardware to perform the computations. Over the last two decades, with the advancements in computing and infrastructure capabilities that grew with the emergence of cloud computing, the field of AI also has progressed rapidly. However along with an improvement in hardware capabilities on one end, on the other, computing started becoming more and more pervasive with a plethora of connected devices (mobile phones, IoT devices, smart watches, etc.). It's these concerns relating to data privacy, security, resource constraints and latency which demanded attention eventually leading to the emergence of edge AI.

An Edge Over Data Centers

The underlying thought behind the genesis of 'AI on the edge' was that a lot of the problems could be solved if one could perform computation at the location where data is gathered or at the source of data itself rather than sending it anywhere else. One problem that can be addressed is that of data privacy with data stored on your device itself. The second relates to latency, because now you can get immediate responses without waiting for long. The third aspect is that of sustainability. From an energy consumption angle, *as per studies conducted,* data centres consume roughly 1-2% of global electricity and account for about 2- 4% of global carbon emissions. Hence having ML models on the edge or on devices is the need of the hour. We see this on our phones with multiple AI models already in operation, like facial recognition or classification of photographs in the Gallery based on events and so on, which require ML models in place. As researchers working in the intersection of software architecture and ML, we are constantly trying to improve the efficiency and effectiveness of such ML systems. One way we attempted to do this is via a self-adaptive mechanism which selects the right model to process data based on user demand and resource constraints. This concept stems from our research on self-adaptive ML-enabled systems [1] which proposes the use of a model balancer that switches between machine learning models considering operational context and environment such as number of user requests, response time of models, accuracy, energy consumption [2] , etc.

[1] *Towards Self-Adaptive Machine Learning-Enabled Systems Through QoS-Aware Model Switching -https://arxiv.org/abs/2308.09960*

[2] *EcoMLS: A Self-Adaptation Approach for Architecting Green ML-Enabled Systems https://arxiv.org/abs/2404.11411*

Switching For Efficiency

Typically, there's a trade-off between latency and accuracy of predictions of an ML model; the more accurate the prediction, the longer the time taken for processing but there are scenarios when lighter models can offer higher accuracy with lower latency. In lieu of creating a single lighter model or a heavier one, our solution looks at the context and then decides whether to use a heavy ML model vis a vis a lighter model. To illustrate how this works, consider a phone camera that is being used to monitor crowds. If the number of people is very high at a given point in time, you would probably require a sophisticated and highly accurate model that is 'heavy'. But if there is no one around, you could probably use a smaller model that consumes less battery power, and processes data faster. The concept is that of a 'model balancer', that is, if you have a suite of ML models performing the same task, the system intelligently switches between models based on the input that is coming in. When this is done, it can improve accuracy, save cost, and improve response time, thus striking a balance between the critical parameters. We have developed a suite of algorithms that can achieve this including clustering-based ones, epsilon greedy, etc which allows us to have a trade-off between accuracy and response time/energy consumption. Based on this concept, that has been published in various top-tier international conferences, we have also developed an exemplary tool that researchers and practitioners can use. It is called Switch which aims to provide an ML system with the ability to autonomously adapt to different scenarios through model switching [3] .

Prototyping On the QIDK

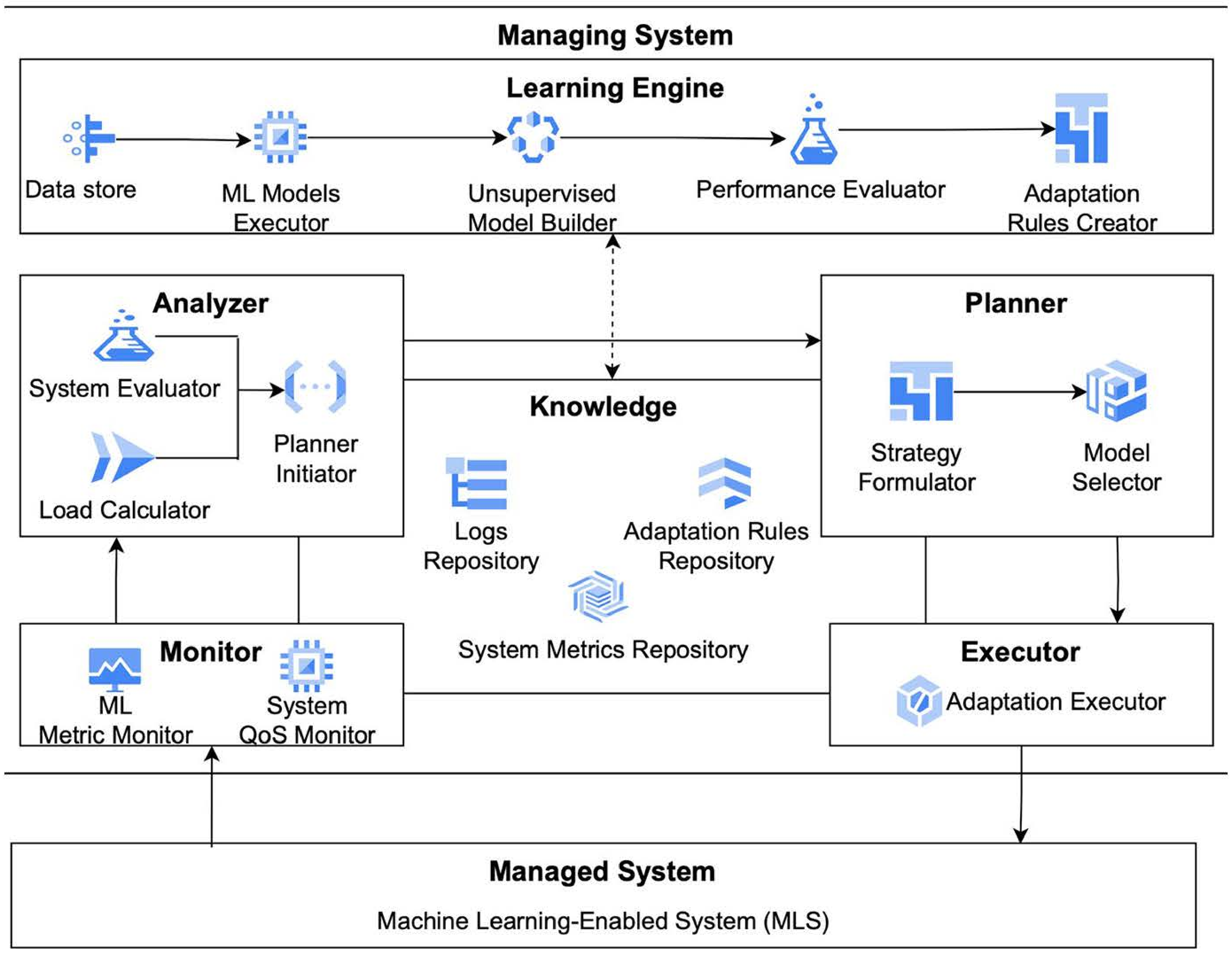
For developing the self-adaptive mechanism, we had picked computer vision as the application domain. This was partly due to the fact that we were working on developing a crowd monitoring solution in the Smart City Lab. We were able to successfully apply our model balancer concept to computer vision domain. This is now being extended to different natural language processing domains, or even regression domains which require numerical predictions like those used for predicting the weather. In a crowd monitoring situation with dense crowds, a mobile camera or drones capturing data is a more efficient way of analysing crowds. When we began exploring the model-switching mechanism on smartphones, coincidentally Qualcomm, which has been a pioneer in the edge domain, came up with the Qualcomm Innovators Development Kit (QIDK). The kit is a SnapDragon mobile platform enabling developers to build prototypes for a variety of applications. We developed an Android app based on our model balancer concept and deployed it in the QIDK obtaining promising results. The kit performed about 10x faster in terms of processing compared to some of the state-of-the art devices out in the field thanks to its dedicated hardware accelerator and GPU. It was thanks to Qualcomm that this work was also presented during the Qualcomm University Platform Symposium and was highlighted during the Qualcomm developer conference in April 2024.

[3] https://tool-switch.github.io/

Onward and Forward

Currently, one of the biggest challenges in software engineering is software sustainability. Edge computing can play a big role there but even that needs to be sustainable. One aspect to that is energy consumption of the device itself. Regardless of the device one is using, no one wants its battery to drain out soon. We also want to be mindful of the carbon emissions of the software on the edge *considering that at present software emissions equate to the carbon emissions of rail, air and shipping combined*. Another dimension is that of technical sustainability which is about how maintainable the system is. Over a longer period of time, we will have a large number of models of several different types in the system worked upon by different individuals and the system must continue to seamlessly switch between different models. In order to enable this, we need more work to be done on the EdgeMLOps side.

We also need to look into the economic angle of sustainability. Running on edge might be cheaper compared to cloud depending on a given operational scenario. With the support of Qualcomm Edge AI Lab, our group is working strongly towards the goal of sustainability from all these angles. In this direction we are also collaborating with Dr. Suresh Purini to integrate the notion of model balancer with the framework that was developed by his research group for large scale CCTV camera analytics. In addition to this, another compelling research angle that is being actively explored is this notion of edge cloud continuum where switching between the cloud and the edge itself can be performed based on the energy, performance and its subsequent trade-off. We may want to process data on our phones and that is highly effective but sometimes we may need the power of the cloud instead and this sort of self- adaptation by the edge system is perhaps one of the ways forward.



DR. KARTHIK VAIDHYANATHAN

IIITH

is an Assistant Professor at the Software Engineering Research Center, IIIT-Hyderabad, India where he is also associated with the leadership team of the Smart City Living Lab. His main research interests lie in the intersection of software architecture and machine learning with a specific focus on building sustainable software systems in the cloud and the edge. Karthik also possesses more than 5 years of industrial experience in building and deploying ML products/services. Karthik is also an editorial board member of IEEE Software. https://karthikvaidhyanathan.com

SLM & LLM system models for Edge AI solutions

***Janakiram MSV** elaborates how a hybrid system with a small language model (SLM) at the edge and a large language model (LLM) in the cloud can optimise data locality, enhance privacy and provide timely, context-aware insights.*

The Hybrid Approach to GenAI - SLMs and LLMs

The architecture of integrating generative AI into healthcare applications can be optimized by strategically deploying SLM at the edge and LLM in the cloud. This hybrid system leverages the strengths of both computational paradigms to deliver efficient, scalable, and privacy-compliant solutions.

The high processing and memory requirements of LLMs make them inappropriate for deployment at the edge. LLMs, such as OpenAI's GPT-4o or Google's Gemini, have billions of parameters, necessitating massive processing power and storage capabilities that far exceed the limitations of most edge devices. These models also require significant energy to operate efficiently, which poses a challenge for battery-powered or resource-constrained edge environments. Additionally, the latency associated with processing complex tasks on LLMs can be prohibitive for real-time applications, where immediate responses are crucial.

In contrast, SLMs are designed to be lightweight and efficient, making them ideal for edge AI applications. SLMs have a reduced number of parameters, which allows them to run on devices with limited computational resources, such as smartphones, IoT devices, and medical equipment. Recent advancements in GenAI research have made SLMs multimodal, allowing them to handle diverse data sources such as images, audio, and video.

SLMs can perform essential tasks such as data preprocessing, filtering, and anonymization swiftly and with lower energy consumption. This efficiency enables them to provide real-time responses and maintain continuous operation in edge environments without the need for constant connectivity to the cloud. By offloading the initial data processing to SLMs at the edge, systems can ensure data privacy, reduce latency, and optimize bandwidth usage, thereby creating a robust and scalable solution for edge AI deployments in healthcare and other critical sectors.

The Workflow

Data Collection: Healthcare environments generate diverse data types, including text records, medical images, audio recordings, and sensor data from wearable devices. This multimodal data is essential for comprehensive patient monitoring and diagnosis.

Edge Processing: At the edge, the SLM preprocesses, filters, and anonymizes the collected data. This step ensures compliance with privacy regulations and reduces the volume of data transmitted to the cloud, addressing latency and bandwidth issues.

Data Fusion: The SLM integrates multimodal data into a structured and coherent dataset, providing a holistic view of the patient's health status. This fusion is crucial for accurate and real-time decision-making.

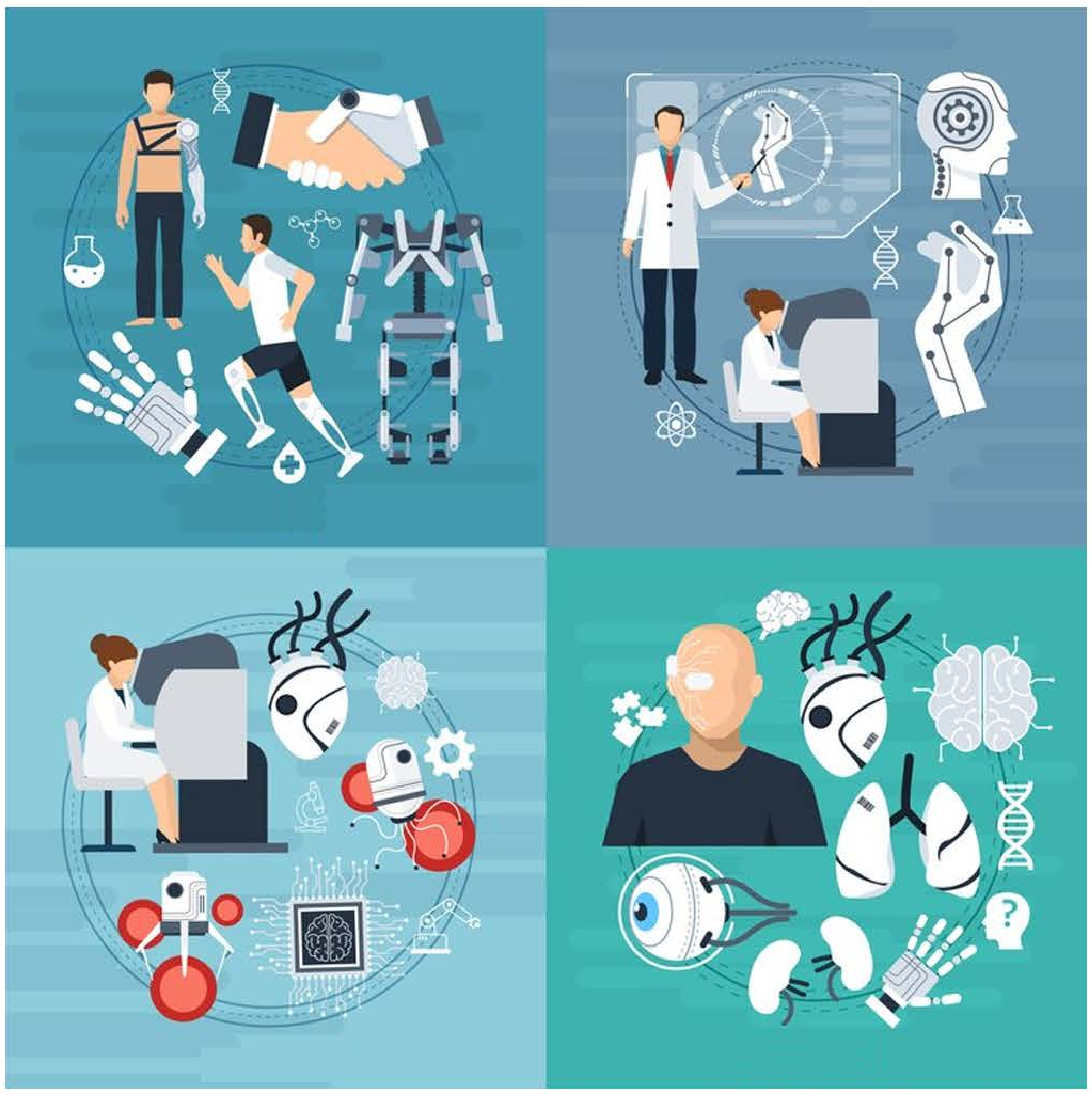
Selective Transfer: Relevant, anonymized data is securely transmitted to the cloud-based LLM. This selective transfer ensures that only necessary information is processed in the cloud, optimizing resource utilization.

Cloud Analysis: The LLM performs deep, multimodal analysis and generates comprehensive insights. These insights are then sent back to the edge for real-time decision support, enabling healthcare providers to make informed clinical decisions.

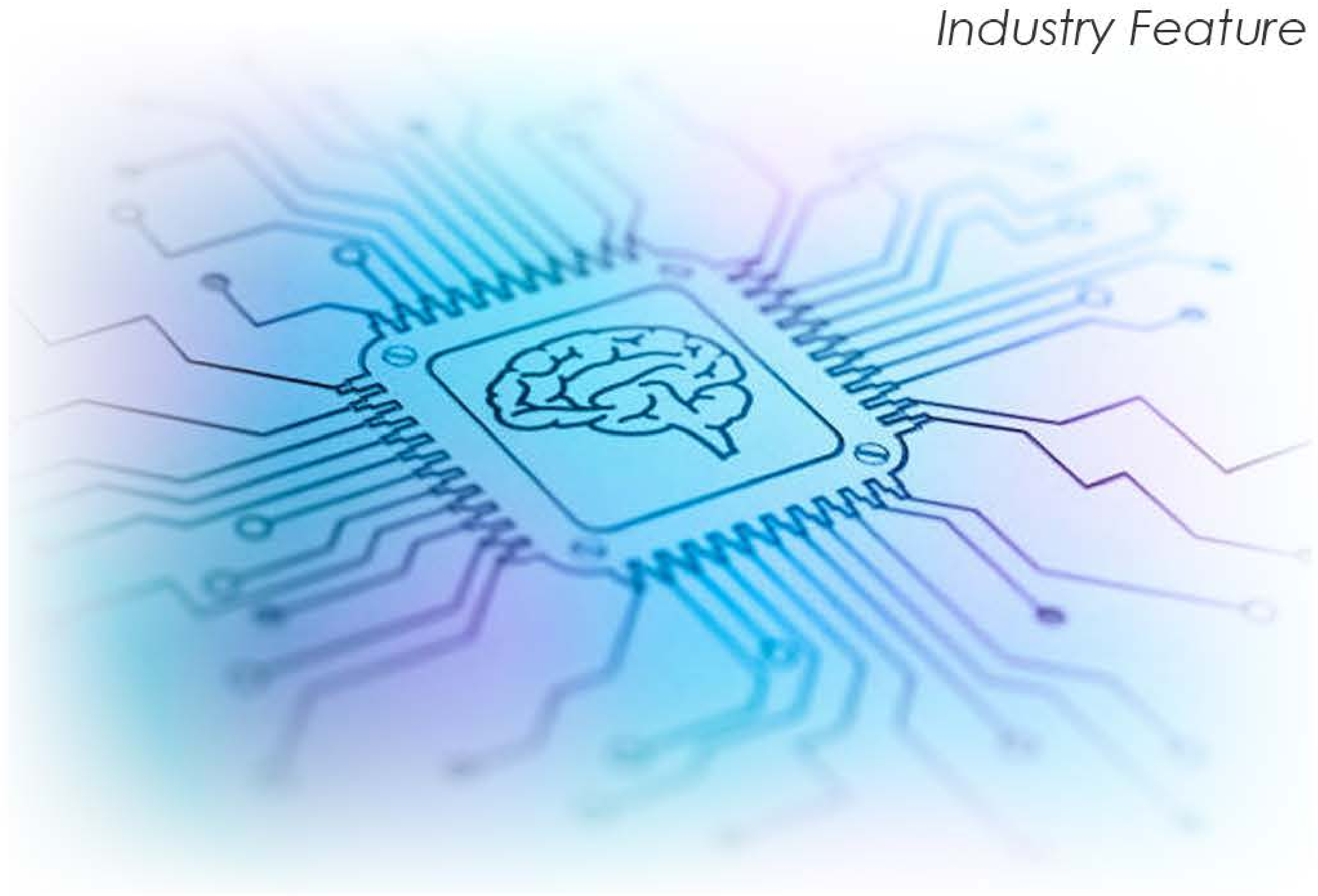
Case Study: Med-PaLM M

A notable case study illustrating this hybrid approach is Med-PaLM M, developed by Google Research and DeepMind. This multimodal AI system integrates clinical language, imaging, and genomics data, showcasing the potential of GenAI in healthcare. Med-PaLM M was benchmarked against radiologists in generating chest X-ray reports, with clinicians preferring its reports in approximately 40.5% of cases. This demonstrates the model's capability to enhance diagnostic accuracy and support clinical workflows.

Data Collection and Edge Processing: In Med-PaLM M, data collection involves capturing diverse types of medical information, including chest X-rays (images), clinical notes (text), and possibly patient history data (structured data). At the edge, an SLM preprocesses these images by anonymizing patient information and filtering out noise or irrelevant data. This initial processing ensures that sensitive data is not transmitted to the cloud, maintaining patient privacy and adhering to regulatory standards.



Data Fusion and Selective Transfer: The SLM at the edge integrates the processed images with relevant text data into a coherent dataset. For instance, an anonymized chest X-ray image could be paired with a summary of the patient's medical history or symptoms. This integrated dataset is then selectively transferred to the cloud for further analysis. Only the necessary and anonymized information is sent, which optimizes bandwidth and reduces latency.



Cloud Analysis: Once in the cloud, Med-PaLM M's LLM performs a deep analysis of the data. It interprets the chest X-ray images, cross-references them with the clinical notes, and integrates genomic data if available. This comprehensive analysis allows the model to generate detailed diagnostic reports and treatment recommendations. The cloud-based LLM leverages its extensive training on diverse datasets to provide accurate and context-aware insights.

Feedback to Edge for Decision Support: The insights generated by the LLM are then sent back to the edge, where they can be used by healthcare providers for real-time decision support. For example, the LLM might identify signs of pneumonia in a chest X-ray and recommend further tests or treatments. This information is immediately available to clinicians on edge devices, ensuring timely and informed clinical decisions.

Conclusion

Generative AI's applications in healthcare are vast, ranging from synthetic data generation to drug discovery. In clinical settings, GenAI can streamline administrative tasks, such as generating discharge summaries, synthesizing care coordination notes, and improving electronic health records (EHRs). By reducing the administrative burden on healthcare providers, GenAI helps prevent burnout and enhances the efficiency of healthcare delivery.

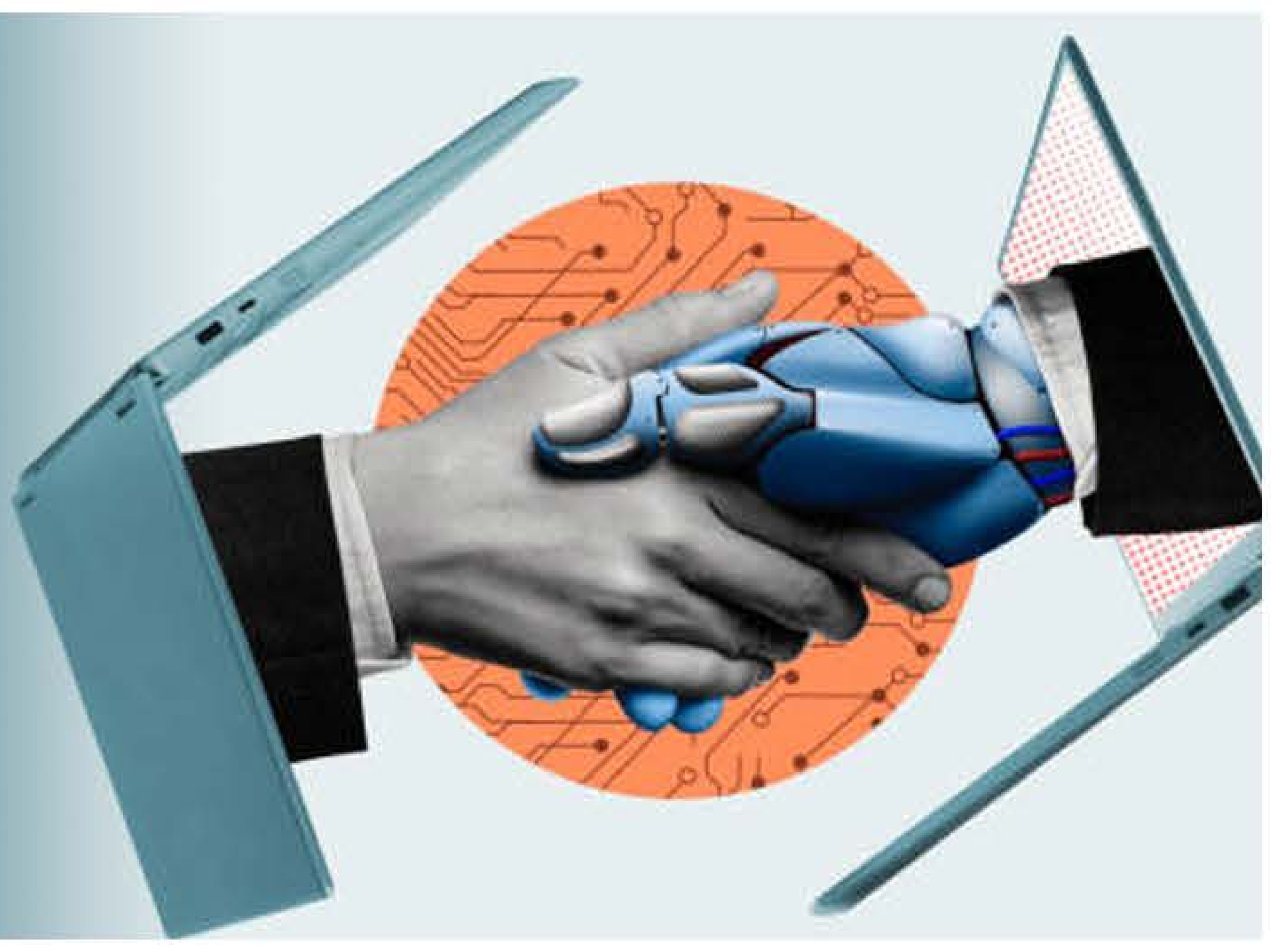
The integration of GenAI at the edge with advanced multimodal processing capabilities holds immense potential for revolutionizing healthcare. By combining the strengths of edge and cloud-based models, this approach ensures data privacy, optimizes resource use, and delivers timely, context-aware insights. As demonstrated by systems like Med-PaLM M, the future of healthcare lies in leveraging AI to provide more accurate, efficient, and personalized patient care.

JANAKIRAM MSV

ANALYST | ADVISOR | ARCHITECT

is a practicing architect, research analyst, and advisor to Silicon Valley startups. He focuses on the convergence of modern infrastructure powered by Kubernetes and machine intelligence driven AI. Before becoming an entrepreneur, he spent over a decade as a product manager and technology evangelist at Microsoft Corporation and Amazon Web Services. Janakiram regularly writes for Forbes, InfoWorld, and The New Stack, covering the latest from the technology industry. He is a keynote speaker for internal sales conferences, product launches, and user conferences hosted by technology companies of all sizes.

The Future of AI: A Global View



With the global landscape abuzz about breakthroughs such as OpenAI's GPT-4, Google Deep Mind's AlphaFold and Tesla's autonomous systems, Prof. Manish Gangwar explores diverse opinions from renowned venture capitalists, academics and entrepreneurs about the future of AI from a global perspective.

As a prelude, here's a vibrant mosaic of the AI landscape with different regions leading in specific areas:

United States: Dominates in research and development, attracting top talent and funding. Key players include Google, Microsoft, Anthropic and OpenAI.

China: Aggressively investing in AI, focusing on applications like facial recognition and autonomous vehicles. Baidu, Alibaba, and Tencent are prominent players.
Europe: Emphasizes ethical AI development, prioritizing privacy and data security. Companies like DeepMind (owned by Google) and the European Union's Human Brain Project are leading the charge.

Other Emerging Regions: India, Canada, and South Korea are actively contributing to the AI revolution, focusing on specific industries like healthcare and education.

The Future of AI

There have been many opinions in the media regarding the future of AI from multiple academicians, authors, entrepreneurs and venture capitalists. They can largely be divided by the tone of their thoughts into optimistic and cautious.

Optimistic Voices:
Elon Musk, CEO, Tesla, SpaceX, and Founder, The Boring Company, "Generative AI is the most powerful tool for creativity that has ever been created. It has the potential to unleash a new era of human innovation." - Musk, a visionary thinker, believes that generative AI is the most powerful tool for creativity ever invented. He suggests that this technology will usher in a new era of human innovation, pushing the boundaries of creativity beyond what was previously possible.

Ian Goodfellow, Research Scientist at DeepMind: "Generative models are a key enabler of machine creativity, allowing machines to go beyond what they've seen before and create something new." Goodfellow's ideas mark a significant change in how we think about machine intelligence. Instead of data replication, generative models allow machines to create new things. This shift towards machine creativity is a key part of the evolving understanding of what AI can do.

"AI is the new electricity." Andrew Ng, Founder of Landing AI envisions AI becoming a ubiquitous technology, powering countless applications and improving lives in profound ways. Jensen Huang, CEO at NVIDIA thinks, “AI will be the most transformative technology of the 21st century. It will affect every industry and aspect of our lives.” "As a technologist, I see how AI and the fourth industrial revolution will impact every aspect of people's lives". - Fei-Fei Li, Co-Director, Stanford Human-Centered AI Institute.

"Artificial intelligence is not just about efficiency gains, it’s about opening new possibilities, unlocking human potential and solving some of society’s biggest challenges." -Yoshua Bengio, Computer scientist and Turing Award laureate, 2021. “Generative AI has the potential to change the world in ways that we can’t even imagine. It has the power to create new ideas, products, and services that will make our lives easier, more productive, and more creative. It also has the potential to solve some of the world’s biggest problems, such as climate change, poverty, and disease.” - Bill Gates, Microsoft Co-Founder. Gates envisions a future where generative AI plays a key role in not only enhancing creativity and productivity but also in tackling major global issues like climate change, poverty, and disease.

Cautious Voices:
Yuval Noah Harari, Historian and Author in his book “Homo Deus: A Brief History of Tomorrow”: "AI could create a new kind of inequality, with a small elite controlling a vast majority of wealth and power." (2017). Harari highlights the potential for AI-driven automation to displace jobs and exacerbate social disparities. Demis Hassabis, CEO of DeepMind in 2023 cautions on the fast pace of companies improving LLMs and advocated not moving fast and breaking things.
This next generation of “AI will reshape every software category and every business, including our own. Although this new era promises great opportunity, it demands even greater responsibility from companies like ours.” Satya Nadella, CEO at Microsoft, 2023. “Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.” – Stephen Hawking

Key Trends Shaping the Future of AI: Amid the diversity of opinion, AI progress is relentless. Genie is already out of the bottle, and we see the following trends.

AI and Human Collaboration: AI is automating tasks across various sectors, from customer service to manufacturing, leading to increased efficiency and productivity but also raising concerns about job displacement. The future of AI is not about replacing humans but rather enhancing their capabilities, creating a symbiotic relationship between humans and AI systems in the education and healthcare domain.

AI Ethics and Governance: Growing awareness of potential ethical challenges associated with AI is driving the development of frameworks and guidelines for responsible AI development and deployment.

The Rise of AI-powered Edge Devices: AI is moving beyond the cloud to edge devices, enabling real-time decision-making and personalization at the point of interaction.

Democratization of AI: Open-source frameworks and cloud-based platforms make AI accessible to a broader range of developers and organizations, fostering innovation and applications across industries.

Conclusion

The future of AI is a complex and evolving landscape shaped by technological advancements, societal expectations, and ethical considerations. While the potential for AI to solve global challenges and improve lives is immense, it also presents significant risks that require careful attention and responsible development. The future of AI is not predetermined but rather a collective responsibility. By embracing a global vision of responsible AI development; we can tackle existing challenges and explore new opportunities to create a future where AI empowers humanity and advances our collective well-being. Policymakers, researchers, and industry leaders must collaborate and engage in a global dialogue about the future of AI. By fostering responsible innovation, promoting ethical guidelines, and ensuring equitable access to AI, we can harness its power for the betterment of humanity.

PROF MANISH GANGWAR

ISB

is a distinguished faculty member and Executive Director of the Institute of Data Science and Business Analytics at the Indian School of Business (ISB). He is a renowned pricing and business analytics researcher and has served as the Associate Dean of Research and RCI management at ISB. Professor Gangwar has a Ph.D. in Management Science from the University of Texas at Dallas, an MS from the University of Kentucky, and a BE from the Indian Institute of Technology, Roorkee, along with years of industry experience. Professor Gangwar is recognized as one of India's leading data science academicians and a prominent global academic data leader.

A community initiative, anchored by

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

IIIT 25 Hyderabad

TechForward
DISPATCH