Edge-AI Market Analysis: Applications, Processors & Ecosystem Guide

April 2025
SHD102.a-25

Authors: Richard Wawrzyniak, Joseph Byrne

Table of Contents

List of Tables

List of Figures

I. Executive Summary

Report Methodology

In conducting the research for this report, The SHD Group surveyed and interviewed over 40 companies that were either SIP vendors, silicon device manufacturers, software companies EDA Tool companies, system developers or who were present in a combination of one or more of these areas. We also conducted primary and secondary research in the preparation of the market actuals and forecasts.

This report examines AI processing at the edge, focusing on embedded applications. As AI began to be commercialized in the 2010s, its first killer applications were voice and image recognition. A wave of chips integrating AI accelerators (NPUs), discrete offloads and licensable designs (SIP) emerged. Since then, transformer networks have led to large language models and other neural networks, and multimodal networks (e.g., those handling both language and image processing) have shown promise in edge designs.

Interview Findings

To understand this revolution, The SHD Group interviewed and surveyed technology suppliers, including SIP and chip companies. We found that convolutional neural networks (CNNs) are the most common models, owing to computer vision being a dominant AI application. Newer designs support transformers but principally out of concern that they'll be used during a product's lifetime, not because of immediate deployment.

Nonetheless, recurrent neural networks and other model types are important because many edge applications process audio and other sensor data that aren't well handled by CNNs. Spiking neural networks (SNNs) are a niche technology but have power and cost advantages for some embedded designs.

Our research found that neural networks don't map to vertical markets (e.g., consumer versus industrial) but only to sensor data. For example, disparate applications that process images may employ the Yolo CNN or a derivative. Some industries are ahead of others in adopting new models. For example, respondents generally indicated customer interest in transformers, but a few indicated the automotive industry was ahead of other segments in adopting vision transformers. Also, standard models, such as a specific Yolo version, may help customers evaluate a chip. However, for production, customers may deploy a custom model.

As with any processing technology, performance, power and cost are key differentiators among AI accelerators. Our research found, however, that there's no good performance metric, particularly for SIP, where other design factors constrain throughput. A further differentiator is software enablement, for which there's no practical metric.

Survey Findings

Our survey respondents mostly supplied NPUs and microprocessors / microcontrollers. They mostly targeted markets broadly instead of focusing on a single segment, which is consistent with neural networks mapping to sensor data instead of segments. Most commonly targeted data are for vision, 3D depth, and audio / speech. Less commonly targeted data includes text / language and machine condition.

Regarding AI frameworks, respondents predominantly support TensorFlow variants and PyTorch. Some cited ONNX Runtime support, but we suspect that they were confusing the ONNX Runtime with the ONNX file format. A typical development the deployment workflow entails computer-based model training, converting toe model to the ONNX format, and then using an edge-AI vendor's tools to convert it to a proprietary format for loading onto hardware.

As for specific models, MobileNet, ResNet, and Yolo are the most commonly supported for product evaluation. All CNNs, these models come in different versions, with Yolo varying the most from among versions. Significantly, despite touting transformer support, few respondents support Bert, a mature transformer model that MLCommons has used in its MLPerf benchmark.

II. Foreword

Over the past decade, significant technological advancements in semiconductors and software have fueled the rapid evolution of Artificial Intelligence (AI), particularly deep learning. Initially, AI workloads were predominantly processed in large cloud data centers, leveraging powerful GPUs and specialized AI accelerators to handle complex computations at scale. These AI systems have driven a wide range of applications, from natural language processing and recommendation engines to predictive analytics and autonomous systems. Various neural network architectures, including convolutional, recurrent, and more recently, transformer networks, have been optimized to extract meaningful insights from massive datasets, transforming industries worldwide.

Continuous evolution in the semiconductor industry has been a driving force behind AI's expansion. With more efficient and specialized semiconductor architectures, alongside increasingly optimized neural networks, AI inference has become not just possible but practical on Edge AI devices. This shift has been enabled by advances in low-power AI accelerators, such as NPUs and domain-specific processors, which allow complex models to run efficiently on embedded and mobile platforms.

Market and technological advancements are accelerating the shift of AI inference from large cloud data centers to SoC devices at the edge. This transition has driven rapid innovation in SoC architectures, with new designs optimized for low-power, high-efficiency AI processing. As a result, Edge-AI solutions are proliferating across diverse industries, unlocking new opportunities for startups and established companies while fueling growth in market revenues.

The recent introduction of Agentic AI (AI Agents) into the market heralds another boost in innovation. These agents hold the promise of offering personalized AI partners to everyday end users, increasing productivity and informed decision making. If properly sized with the right security guards in place, these agents will primarily be run on mobile platforms, offering new opportunities to semiconductor, software, SIP vendors and systems' developers alike.

Evolution in Nature is driven by genetic and environmental factors alike. In the Semiconductor Industry evolution is driven by technology, which is in turn driven by people and companies. Many of the companies we interviewed are at the forefront of this technological evolution and revolution.

The following table contains a list of the companies who contributed data and insights to the making of this report. We also want to acknowledge the interest and support of the Edge AI and Vision Alliance along with our report sponsors; Andes Technology, Arteris IP, Ceva, Inuitive, MACSO, MagikEye, OpenMV, Quadric and S2C. Without the support of these companies, this report would not have been possible to complete.

We also wish to thank the Edge AI and Vision Alliance for their support and to our sponsors for their support of our research:

Achronix Semiconductor Corporation	Faraday Technology Corporation	Nuclei System Technology
ALIF Semiconductor	Flexlogix Technologies, Inc.	NXP Semiconductors N.V.
Alphawave Semi	Gigantor Technologies	OpenMV, LLC.
Ambarella	Horizon Robotics	Piera Systems
Andes Technology Corporation	Huawei Technologies Co., Ltd.	Plumer.ai
Apple, Inc.	Imagination Technologies Group plc	Quadric, Inc.
Arm Holdings plc	Intel Corporation	Qualcomm Incorporated
Arteris IP	Inuitive Ltd.	S2C Inc.
Aspinity, Inc.	Kinara, Inc.	Samsung Electronics Co., Ltd.
Avassa Systems AB	Krispan Incorporated	Siemens EDA
Axelera AI	LIST Semiconductor	SiFive, Inc.
Baya Systems	Macso Technologies	Silicon Intervention
Brainchip Holdings Ltd.	MagikEye Inc.	Sondrel Ltd.
Breker Verification Systems	Matsusada Precision Inc.	Sophgo
Cadence Design Systems, Inc.	MediaTek Inc.	STMicroelectronics Inc.
Ceva, Inc.	MemryX Inc.	Synaptics Incorporated
Expedera, Inc.	Mythic, Inc.	Synopsys, Inc.
	Edge AI and Vision Alliance	TetraMem Inc.

Source: The SHD Group, March 2025

Key Highlights

Survey Results

- The companies we surveyed are accelerating CNNs, Transformers and RNNs today.
- These same companies are looking to accelerate Transformers, CNNs and GANs in the future.
- The companies are using Visible image Sensors, IR Image Sensors and Audio sensors today.

Edge-AI Market Growth:

- Edge-AI based SoC unit shipments are forecast to surge to 8.7B units, with revenues reaching $102.9B by 2030, boasting CAGRs of 25% and 21%, respectively.

Edge-AI SoC Market Penetration

- Edge-AI device revenues were 22.1% of the SoC market in 2024 and are projected to reach a market penetration of 55.4% by 2030.
- Edge-AI unit shipments were 13.6% of SoC unit shipments in 2024 and are forecasted to reach a market penetration of 31.3% by 2030.

3rd Party SIP Market:

- In 2024, the worldwide SIP market reached $9.2B, marking a 10.9% growth from 2023. Forecasts predict a 9.1% increase to $10.1B in 2025, projecting a potential $15.2B market by 2030, with a CAGR of 8.6%.
- The Central Processing Unit (CPU) SIP market soared by 26.3% in 2024 to $3.2B and is anticipated to hit $5.4B by 2030, demonstrating a robust 9.3% CAGR.

3rd Party AI SIP Market

- AI SIP revenues increased to $633.2M in 2024 and are forecast to reach $2.2B by 2030, a CAGR of 23.2%.
- The AI SIP revenues for CPU cores are the largest segment of the AI SIP market in 2024 at $295.0M and are projected to reach $980.0M by 2030, a CAGR of 22.2%.
- GPU AI SIP was the second largest segment of the AI SIP market in 2024 reaching $126.0M and is projected to grow to $310.2M by 2030, a CAGR of 16.2%.

Edge-AI Regional Revenues

- Total Edge-AI Regional revenues were $32.4B in 2024 and are forecast to reach $102.9B by 2030, a CAGR of 21.2%.
- China has the largest regional revenues in 2024 at $8.5B and is projected to grow to $27.8B by 2030, a CAGR of 21.9%

Edge-AI SoC Design Starts:

- Edge-AI SoC design starts are forecast to reach 853 designs by 2030, a 7.4% CAGR.
- Designs start for Consumer applications, are expected to show the largest number of designs by 2030, with Industrial and Automotive applications following closely behind.

III. Introduction

Although computer-based applications of artificial intelligence and machine learning incorporated into the Data Center dominate headlines, another revolution is occurring in processing at the edge, where data is generated and applied. New AI-based decision-making and analysis techniques are empowering consumer IoT devices, improved manufacturing process flows, various types of industrial electronics and other embedded systems.
Understanding the revolutions answering several questions:

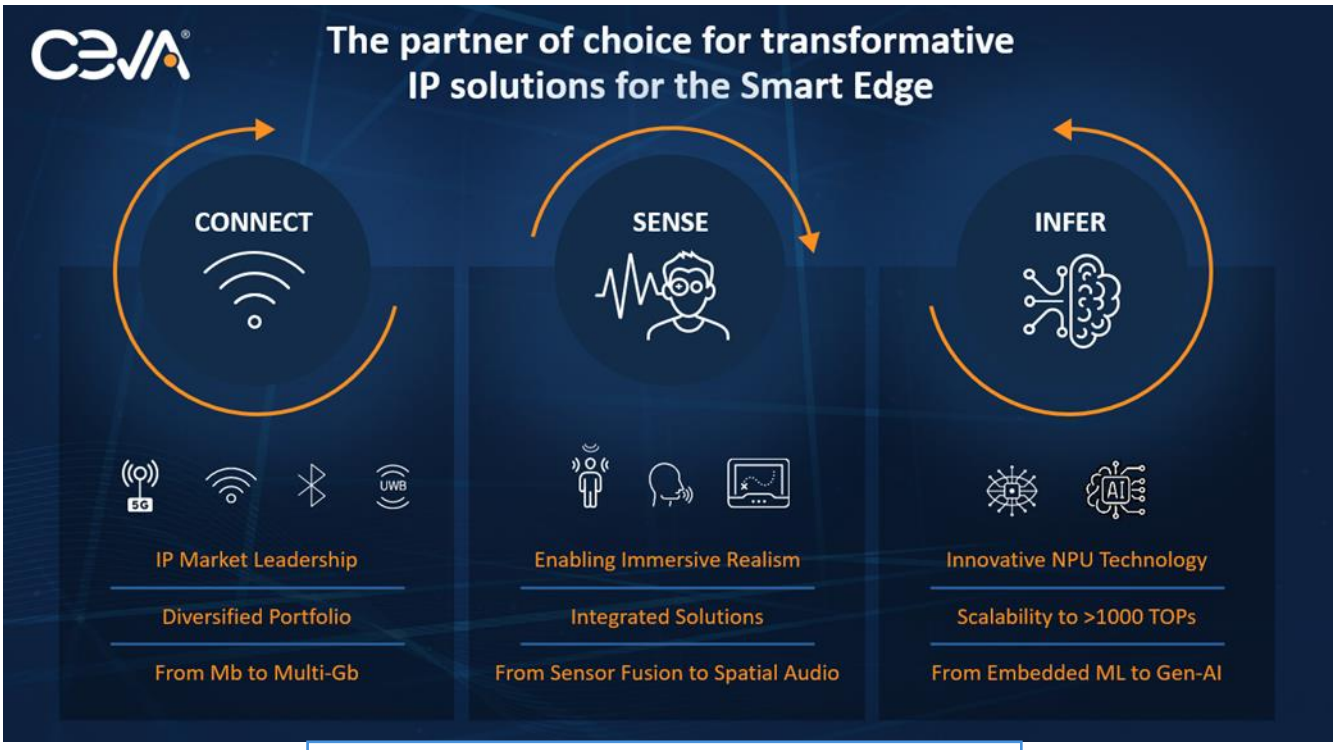
- What AI technology suppliers are available to developers?
- What neural network types do they support?
- Which specific off-the-shelf networks do they support?
- How do neural networks vary by industry?
- Are their supported models, software tools, and industry focus aligned with developers' practices?

To address these questions, The SHD Group interviewed chip and SIP suppliers and surveyed them, in addition to software companies, and service providers. This document reports our findings, provides an ecosystem guide, and—for sponsors—provides detailed forecasts for the industry. The first section defines terms, the second summarizes our interviews, and the third reports results from our survey. Where possible, the latter section compares our findings with those from the Edge-AI and Vision Alliance's developer survey. We're indebted to the alliance for supporting our work. The forecasts are in an optional addendum containing our analysis of Edge-AI acceleration silicon by application.

Scope

This report covers processors and the surrounding ecosystem for artificial intelligence and machine learning at the edge, focusing on embedded systems ranging from TinyML to those capable of hundreds of TOPS.
The SHD Group surveyed technology suppliers about AI. The survey, our interviews, and this report use various terms, which we define in the next section.

Figure 1: Ceva, Inc., Diversified IP Portfolio



Source: Courtesy Ceva, Inc., March 2025

Ceva's transformative semiconductor SIP and embedded software offerings are used by the world's top semiconductor and electronics companies to develop extraordinary and differentiated products that **connect, sense, and infer** - the three critical pillars of the rapidly evolving era of AI-enabled Smart Edge. Ceva's SIP solutions for SoC integration enable a new generation of connected and distributed intelligence to make our lives safer, enjoyable and more efficient.

IV. Industry Segments and Edge-AI Definitions

Industry Segments

The term covers various embedded systems sold to industries. Table 1 (below) lists the segments used in our Edge-AI survey along with example applications that could employ AI. Our survey additionally asked respondents targeting the industrial and consumer categories about the following subcategories:

- Industrial—smart farm, manufacturing, smart grid, smart-city infrastructure, building automation, medical / healthcare, retail analytics, test and measurement, distribution / logistics / warehousing, military / aerospace, security / surveillance, and transportation.
- Consumer— entertainment, home security, white goods, and mobile / wearable devices.
- Computer – Desktop PCs, Laptops, Tablets, Solid State Drives (SSDs) and Industrial PCs.
- Networking & Communications – High, Mid and Low-end Routers, Cable and DSL Modems and 4G/ LTE Picocell and Femtocell Base Stations.
- Automotive – High, mid and Low-end Passenger Cars and Commercial Vehicles.
- Other – applications which are not covered in the five categories mentioned above.

A further breakdown of the specific categories by industry segment is listed in Table 1 below.

Table 1: Selected Systems by Category and Segment

Category	Segment	Systems	Systems	Segment	Category
Consumer	Smart Home	Security Cameras	Industrial Robotics	Industrial	Industrial
		Robotic Home Appliances	Smart Grid		
		White Goods	IIoT (Factory Floor)		
			AgriTech -Farm Equip		
	Retail	POS Terminals	AgriTech-Farm Animals		
		Handheld POS Terminals	Drones & Controllers		
			Industrial PC		
	Personable Wearables	AR / VR			
		Smart Watches	Land	Military	
		Smart Clothing	Sea		
		Headsets / Earbuds	Air		
			Space		
	Mobile	Handheld Game Consoles			Computer
			Desktop PC	Computer	
	Home Entertainment	UHDTV	Laptop PC		
		Streaming Media Devices	Edge Computer		
		Smart Speakers	Tablet		
		Set Top Boxes	Solid State Drives		
Automotive					Networking
	Transportation	Commercial Vehicles	High-end Routers	Communications	
		Low-end Passenger Cars	Mid-range Routers		
		Mid-range Passenger Cars	Low-end Routers		
		High-end Passenger Cars	Cable Modems		
Other			DSL Modems		
	Other	Other	4G/LTE Picocell Base Stations		
			4G/LTE Femtocell Base Stations		

Source: The SHD Group, March 2025

V. Interview Findings

The SHD Group interviewed technology suppliers to discuss the development and deployment of AI models at the edge. These interviews add qualitative insight to our survey findings.

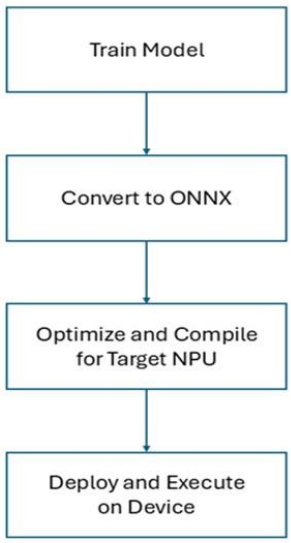
Deployment

Suppliers of AI chips and SIP focus on inference, expecting customers such as OEMs to train models independently using computing resources in the cloud or elsewhere. They mostly assume that developers will train models using PyTorch or TensorFlow. Most provide development tools that can ingest neural networks in the ONNX format. Some can accept models in native PyTorch, TensorFlow, or TensorFlow Lite formats. A few support standard runtimes, such as the TensorFlow Lite engine, but all support—and generally favor—using their proprietary runtime. A standard runtime doesn't buy them anything because the effort required to build the backend that maps it to their hardware is similar to developing a proprietary runtime, and a standard one offers few user advantages.

As Figure 2 shows, a developer's typical flow, therefore, is:

1. Develop and train a model using PyTorch or TensorFlow, likely using a popular open-source model.
2. Convert the trained model to ONNX format.
3. Use the vendor's tools to optimize the model, such as by quantizing it, and then compile it to run on the vendor's accelerator.
4. Execute the compiled model on the hardware, which may entail the host-based application calling a driver, the runtime engine, and running functions on the accelerator.

Figure 2: Typical Edge-AI model development and development flow



Source: The SHD Group, February 2025

Companies offering SNN hardware may follow the typical flow by converting an ONNX-formatted CNN to an SNN running on their NPU. They also offer a proprietary training flow to directly produce an SNN model, which can better take advantage of SNN's advantages.

In addition to developing hardware capable of the various activation functions and operators that models require, vendors must also implement them in their compilers. They start with a core set required by popular models and add operators or activations as customers require or new models gain currency.

Vendors offer model zoos, models already converted and optimized for their hardware. Customers can use them directly or apply the optimizations to their equivalent models. An indicator that a vendor is struggling to keep up with adding new activations and operators is if they focus on an older version of a popular model but not a newer version, such as Yolo v5 but not Yolo v9.

Training Expectations

Exceptions to the practice of training models on computers instead of edge devices exist in industrial and consumer applications. In the industrial area, condition monitoring involves processing data from audio and other sensors to detect if a machine is showing signs of impending failure. If a machine's bearing goes bad, it could make an unusual noise. Every machine sounds different, and the sound may change over time without indicating a problem. Condition-monitoring systems, therefore, may retrain themselves on startup and periodically thereafter to adapt to the new normal. A consumer example is a smart speaker that learns the specific voices of people in a home to personalize responses.

Although it isn't training per se, retrieval-augmented generation (Rag) is akin to training in that it adapts a large language model (LLM) to provide better responses. Moreover, it alleviates the need to periodically retrain a model to incorporate new data. Rag feeds a data set alongside a user's query to an LLM, and the LLM formulates its response using this data set. A couple of vendors we interviewed gave an example of a car manual. The user asks a chatbot built into a car a question about the vehicle, and the bot responds with content from the manual.

Neural Network Types

In our interviews, companies consistently responded that CNNs are the most common neural networks, with transformers becoming increasingly important. Object detection, object classification, and similar vision models are mostly CNN based, and vision applications are common at the edge. Thus, CNNs were among the first network types that edge-AI silicon accelerated.

Transformers are at the heart of LLMs and are finding applications elsewhere. Vision transformers apply the technique to image classification and other computer-vision tasks such as image segmentation (e.g., dividing a scene into people, cars, roadways, buildings, etc.), finding not just boundaries but identifying what's in the scene. Transformers have also been applied to 3D vision perception, using multiple cameras to replace lidar.

There's also interest in LLMs at the edge for natural-language interfaces and answering questions (e.g., the above car manual example). Although state-of-the-art LLMs have 400 billion parameters or more, work is still being done to improve the quality of LLMs with fewer than 10 billion parameters, The TinyLlama project seeks to produce a usable model with only 1 billion. Researchers have also begun implementing LLMs with much smaller state-space models and mapping them to SNNs. Other "tiny" projects seek to similarly scale down large neural networks to execute on edge systems.

The NPUs developed before transformers' emergence, however, didn't work generally well on the new network type. Most were optimized for CNNs. Although both depend on multiply-accumulate operations, CNNs have more data locality. For example, in image processing, a CNN looks at only nearby pixels, whereas a transformer model seeks to relate pixels across the screen. Vendors, therefore, have redesigned their NPUs to improve transformer performance.

Although customers have expressed interest in transformers, some interviewees reported their customers have no specific plans to employ them but are only interested in transformer-capable hardware to ensure their products can employ the new network type should the opportunity arise.

Interviewees also expressed interest in multimodal transformers. We expect multimodality to have little or no effect on hardware design. In the worst case, software manages collecting constituent networks' results and feeding them to a subsequent model. Thus, claiming multimodality support costs nothing.

Other Neural Networks Still Have Roles

Most audio processing employs LSTMs and other RNNs. In this case, CPUs or DSPs with vector units are typically sufficient to achieve the required performance, and the matrix-math accelerators needed for CNNs, and transformers are unneeded. However, CNNs find use in some audio cases, either replacing or complementing RNNs.

A few Edge-AI companies, however, such as BrainChip and Innatera, focus on SNNs. Their key advantage is millisecond latency and milliwatt power for the relatively small, event-driven models they excel at. Like RNNs, they're suitable for processing time-series data, including speech samples and video streams.

Classic machine-language techniques that don't employ neural networks remain relevant, particularly for low-cost applications, such as those served by microcontrollers. Small neural networks, however, are encroaching, particularly as MCUs as AI accelerators.

VI. Looking Forward

Transformers straddle the line between current and future technology. As noted above, customer curiosity about transformers exceeds real demand for the technology. Nonetheless, many vendors have enhanced their products to support the new neural network type. Building on this interest, companies expressed interest in multimodal transformer models. Further, in a freeform response to our survey, they mentioned multimodal models, such as LLaVa, as neural networks they're considering addressing. Interviewees expect that as transformers find use in more applications and the barriers to executing them come down, they'll become more common at the edge.

A possible transformer alternative, the state-space model, was cited by a few interviewees as an area they're investigating. Separately, many survey respondents are considering generative AI techniques, such as GAN and diffusion models. Vendors targeting smartphones and PCs (both outside our scope) have employed these models for eye-catching demos. Diffusion models can also be employed for adapting training data, such as synthesizing a winter scene from a summer video for training an ADAS system. However, this adaptation and training takes place in the data center, even if the model ultimately executes at the edge. Thus, the practical application of GAN and diffusion at the edge remains unclear.

Specific Neural Networks

Among these model types are hundreds of predefined neural networks, and system designers also create their own. Customers can employ off-the-shelf models for production, and such networks are also commonly used for evaluating AI accelerators. Performance on these models should correlate with performance on similar proprietary networks used in production.

Most suppliers specifically support the ResNet-50 CNN model, and it's the de facto standard for characterizing NPU performance. Edge-AI companies also widely support MobileNet, designed to be more efficient and less computationally complex than ResNet—and thus more relevant to edge applications. The Yolo family of models is also widely used as a benchmark. Many suppliers accelerate CNNs and support all three for customers seeking to evaluate their products.

Consistent with many customers being curious about transformers but lacking a specific need, interviewees did not cite a specific transformer model as a neural network commonly used for product evaluation. Bert would be a natural candidate for comparing transformer performance at the edge. Having more than 300 million parameters in its largest incarnation, it's much smaller than recent LLMs but still often too large for embedded use. However, other versions and derivatives are smaller. Although some may find Bert outmoded, it still stands as a benchmark for accelerators' transformer performance and has long been a part of various MLPerf benchmark suites. Companies that say their accelerators handle transformer networks but don't support Bert or another transformer model for evaluation may have immature transformer capabilities.

Particularly past the initial screening phase, customers bring their own models to vendors and request performance and power data according to interviewees. This is unavoidable for companies with unusual approaches, such as SNN acceleration, or offering silicon customization and design services. Large OEMs are also more likely to have custom models and require software customization.

Segments/Categories and Required Neural Networks

Our research found no mapping between neural networks and industrial segments. Networks instead map to functions. Disparate segments and system categories performing the same functions will employ the same or similar neural networks. For example, a smart industrial security camera, consumer video doorbell, and automotive driver-monitoring system all perform person detection and employ a CNN for this function.

Some industries are ahead of others in adopting new models. For example, respondents generally indicated customer interest in transformers, but a few indicated the automotive industry was ahead of other segments in adopting vision transformers. Compared with companies seeking to develop the next consumer sensation, automakers have the benefit of consumer demand, government mandates, and industry objectives to improve safety. When it comes to a safer car, if they build it, the consumer will come, provided the price is right. Thus, they have greater motivation and resources to deploy advanced technologies.

Other industries aren't sophisticated AI users but could be led by suppliers offering technology promising lower power and cost. For example, a hearing-aid maker could be a leading adopter of spiking neural networks for noise reduction because an NPU supplier targeting SNNs could enable a product with longer battery life than a supplier focused on RNNs.

Communicating NPU Differentiations

Interviewees report difficulty communicating their NPU's benefits. The most common metric is TOPS. TOPS, however, is simply a product of an NPU's number of multiply-accumulate (MAC) units and clock speed. However, two NPUs with the same TOPS rating may perform differently owing to memory-bandwidth bottlenecks, activation-function performance, and other factors. Further, memory bandwidth can differ for an IP block in isolation compared with a SoC where multiple units share memory. Moreover, for NPUs accelerating SNNs, the TOPS metric doesn't apply, and it's not relevant to NPUs targeting RNNs or classic techniques. Even a

conventional digital NPU with no performance gotchas could be a poor choice if it doesn't support enough memory to hold a particular model. Moreover, in all cases, throughput can't come at the expense of accuracy.

Throughput can also depend on how good a vendor's compiler is at mapping a neural network to its hardware. Related to this and activation-function performance is operator support. Few vendors will support every operator defined in TensorFlow, but an NPU that otherwise looks good could be inappropriate for a design if its vendor has yet to implement a customer-required operator.

Interviewees mentioned batch size as an issue when comparing performance. Batching multiple independent inferences can improve throughput compared with running them sequentially (batch size of one). Particularly at the edge, a system may have only a single inference task or have latency requirements. An NPU that performs well when batching may be unacceptably slow when running a single inference.

Additional Differentiators

Energy efficiency is another important metric. Chips and licensable designs (IP blocks) typically specify power, which doesn't comprehend how much time the power must be applied to accomplish a task. Thus, energy per inference is the relevant metric at the edge. In the above example, the NPU, with fewer bottlenecks, can complete an inference faster and could be, therefore, more energy efficient. Moreover, an analog NPU that can accept analog sensor input (e.g., from a passive microphone) can be more efficient (and result in a less costly system) than a digital NPU requiring signal conversion or an active audio sensor. An efficient NPU can also always be on, an important feature for many systems (e.g., smart speakers). In short, system-design considerations have as much or more influence on energy efficiency than throughput.

For SIP, area differentiates designs as much as performance and energy/power. If an NPU requires an additional CPU or DSP, the customer must also factor this in. This additional core will also increase power. Similarly, a standalone NPU may require a bigger host processor than a customer has planned for.

Software enablement also sets suppliers apart. An interviewee with a tool to help the customer evaluate different hardware configurations, batch sizes, and other parameters cites this as a differentiator. Another interviewee that offers hardware that doesn't require fine tuning (simplified retraining) to accurately execute a quantized model cites this as an advantage.

In summary, communicating AI accelerators' qualities is difficult for vendors and a minefield for customers. Often cited, TOPS is acknowledged by interviewees as a poor metric. Throughput on a specific model (e.g., MobileNet) is much better but still incomplete.

VII. Survey Findings

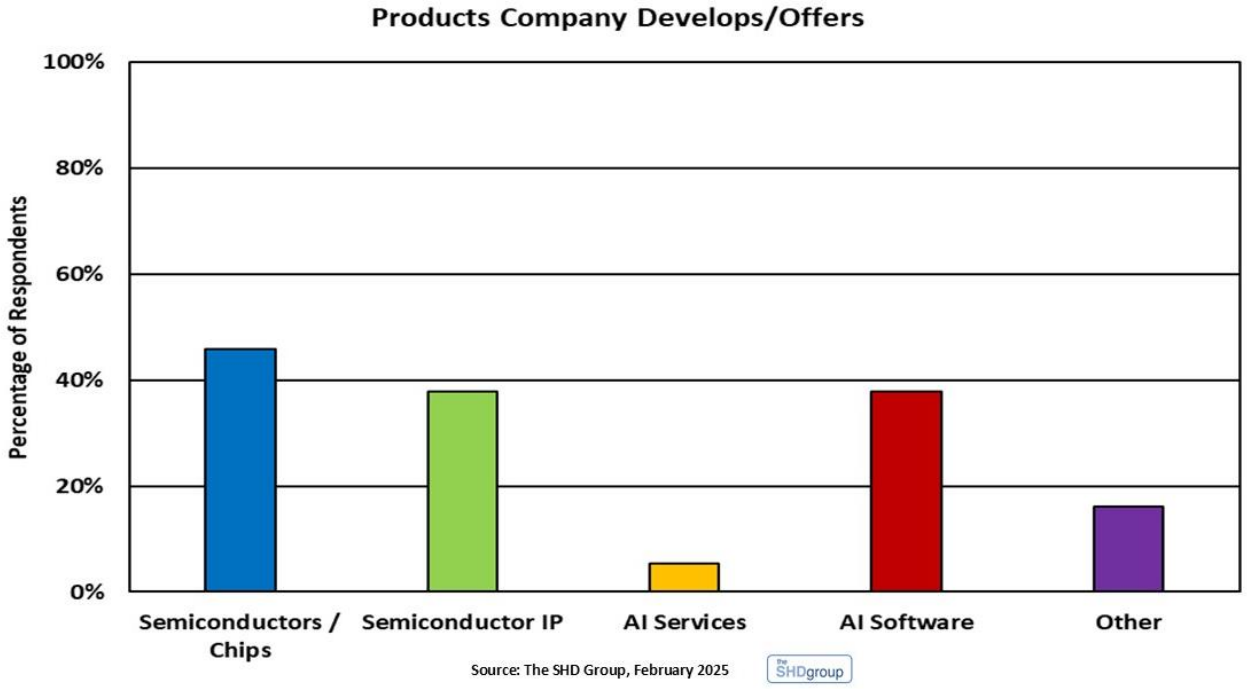
During 2024, the SHD Group surveyed IP suppliers, chip companies, independent software vendors, and others about neural networks and other AI technologies. The survey is available at Questions for Edge-AI Market Report Interviews | The SHD Group and in this report's appendix. It complements the Annual Computer Vision and Perceptual AI Developer Survey Now Open - Edge-AI and Vision Alliance (hereafter referred to as the developer survey), asking similar questions and using similar terms. That survey's respondents are mostly developers, whereas our survey's respondents are mostly technology suppliers. Comparing the two, therefore, reveals any potential mismatches between supply and demand.

Respondent Characteristics

Question 1 of our survey asks about the products a respondent's company produces. Companies could select multiple responses. Most produce semiconductors, SIP, or software, as Figure 3 shows.

While SoC designs remained monolithic, they now implemented many CPU cores, whereas before, there had been only one core to perform the work. In the SoC market, CPU architectures from Arm Holdings and the Intel x86 reigned supreme in most markets and applications. There were other competitive CPU architectures available from Synopsys (ARC), MIPS, Cadence (Tensilica), Andes Technology and others. However, they never quite had the support of the market and were not major factors in the SoC design landscape.

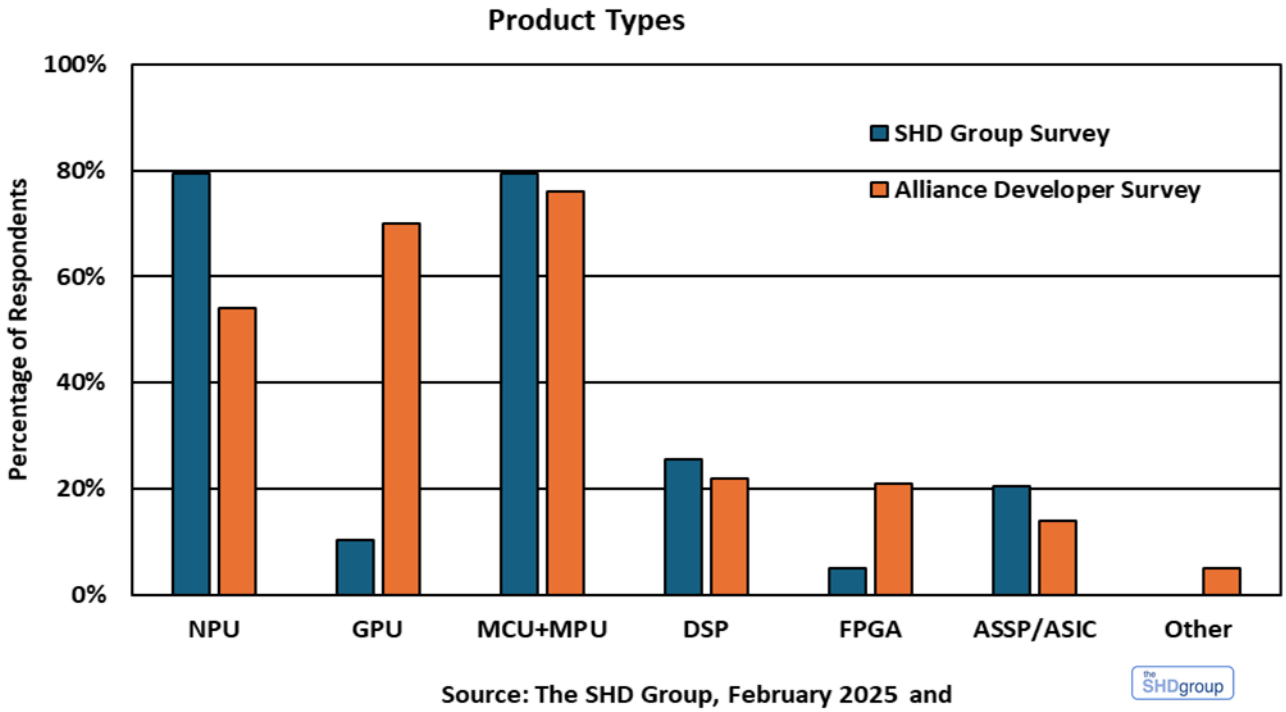
Figure 3: What products does your company develop/offer?



In Question 2, we asked the chip and IP companies to describe their AI hardware, again allowing for multiple selections. For example, Arm supplies CPU IP for MCU and MPU applications along with NPU and GPU IPs.

Figure 4 combines the responses for both groups and plots them next to responses from the developer survey. One should not expect the two surveys to be perfectly correlated. Indeed, the two diverge in the GPU portion. We expect many developers to target Nvidia's GPU-based products, at least for prototyping, if not also for production. Semiconductor companies targeting SIP, however, mostly offer an NPU—technology purpose-built for AI.

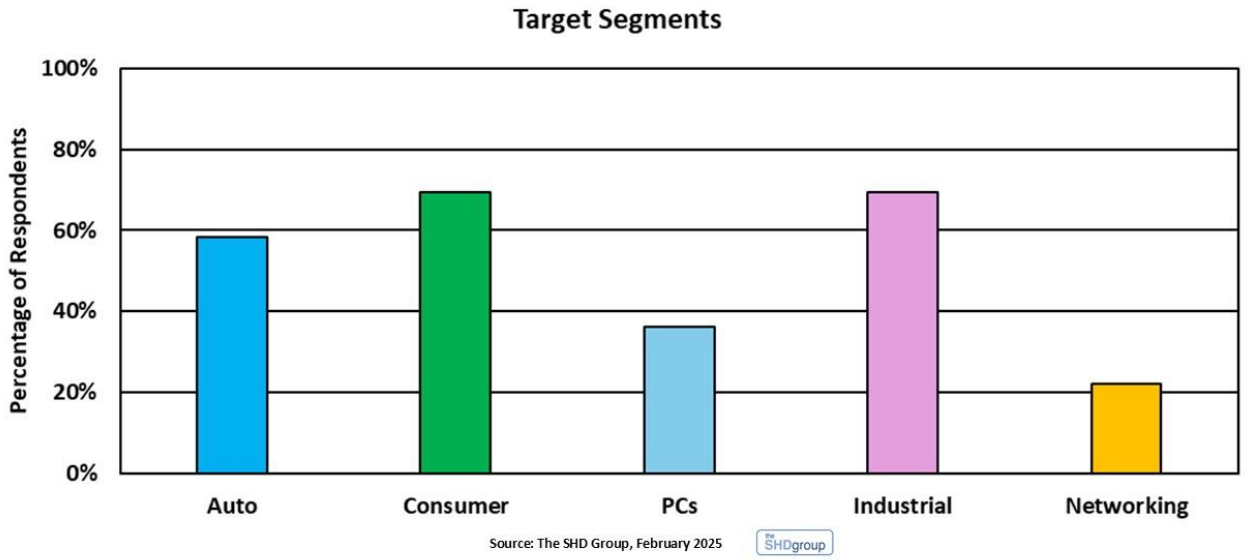
Figure 4: What terms best describe your AI hardware in the AI hardware you support?



Targeted Segments

Question 4 asks about the industries that respondents target. Unsurprisingly, most respondents are open to targeting multiple segments. We interpret the responses as reflecting where they see opportunities. The PC market is generally served by only a few logic-IC suppliers, such as AMD, Intel, and Nvidia, limiting its appeal to other companies. While the networking market is more open, opportunities for AI are limited. Thus, companies see their best chances in the automotive, consumer, and industrial segments, as Figure 5 shows. Our research focuses mostly on the two lattermost segments, leaving automotive, computing, and smartphone applications for later exploration.

Figure 5: What industry segments are you participating in?



For the industrial and consumer areas, we asked respondents to further break out their targets, as Figure 6 and Figure 7 show. The biggest industrial areas are manufacturing (including robotics), security (including surveillance), and smart cities. These responses align with computer vision being a major AI driver. Likewise, the biggest consumer areas are also vision-related: home security and mobile/wearable, which include AR/VR glasses.

Figure 6: Which industrial segments do you participate in?

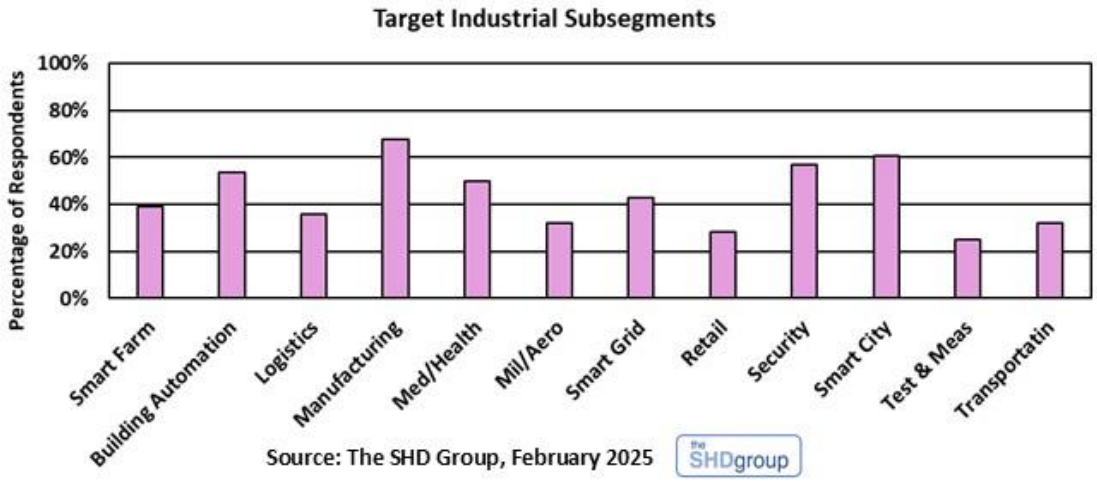
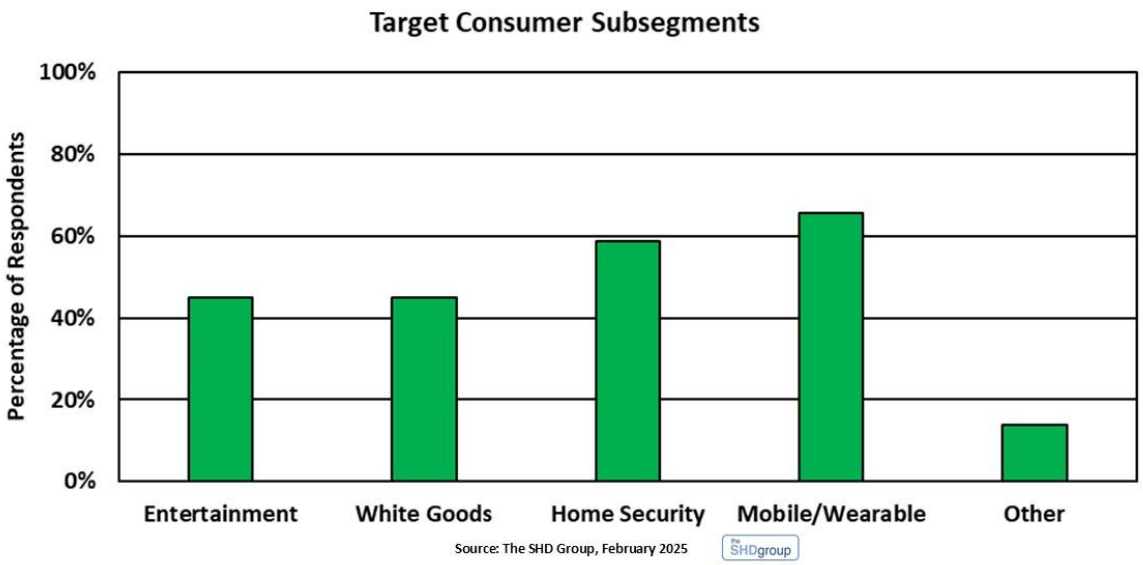
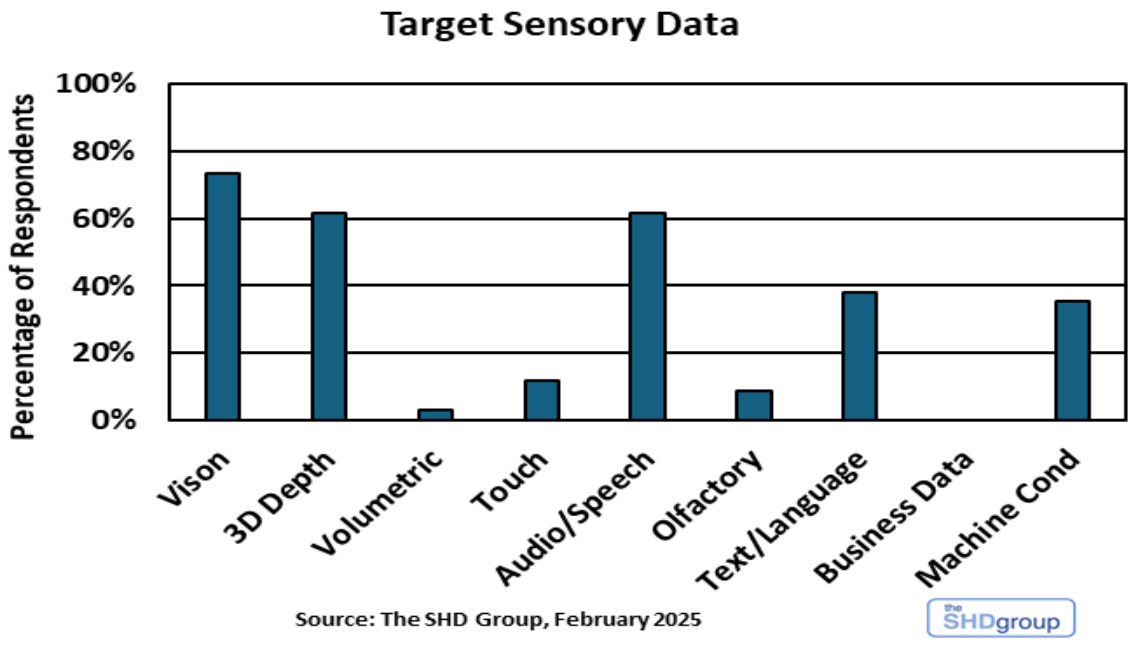


Figure 7: Which consumer segments do you participate in?



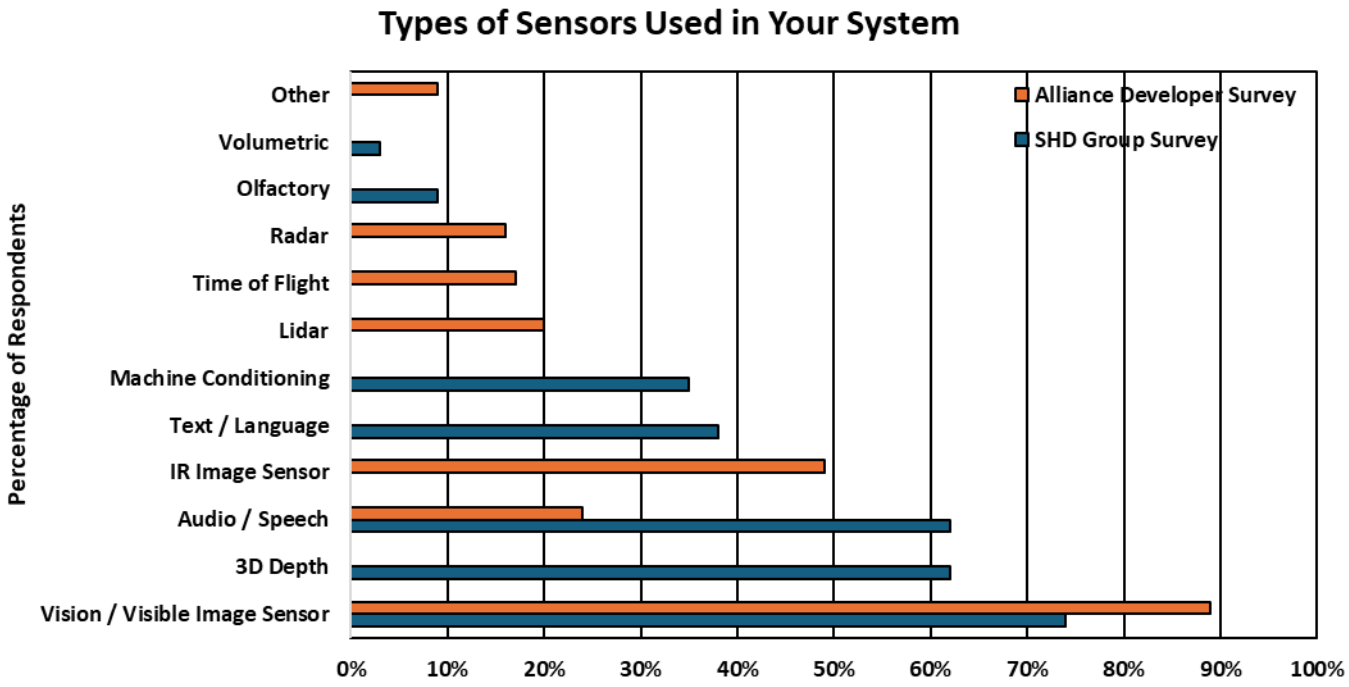
Question 8 asks about the sensory data that companies' products operate on. As expected, vision data is the most popular and is rivaled by 3D and depth data, audio data, text and language, and machine conditions, as Figure 8 shows. Volumetric, touch, and olfactory data rated low—unsurprising given how few edge applications involve these. Business data is in the computing and information-system realm and thus doesn't rate at all given our focus on embedded/edge applications.

Figure 8: What types of sensory data do you operate on?



The developer survey asks a similar question but with different choices. Thus, we plot its findings separately in Figure 9. Its findings align with ours, with vision rating the highest and audio being relatively common. As above, we don't expect a strict correlation between the surveys. Our survey reveals that audio may be overrepresented on the assumption that CNNs will increasingly process audio, whereas developers today may largely employ classic signal-processing techniques or DSP-based RNN models.

Figure 9: What types of sensors does your system use?



Source: The SHD Group, February 2025 and
Edge AI and Vision Alliance, Computer Vision and Perceptual AI Developer Survey 2024,
https://edge-ai-vision.com/survey

Supported NNs

A key objective of our research was to understand the neural networks companies are supporting. We can compare our findings with the developer survey. We found that network types employed by developers are indeed supported by our respondents today, as Figure 10 shows. In particular, CNNs dominate both groups. The tech suppliers we surveyed are ahead of developers with generative AI, including transformer networks, GAN, and diffusion networks. These network types also featured more prominently in responses to our question about future support, as Figure 11 shows.

Figure 10: Which types of neural networks are you accelerating today?

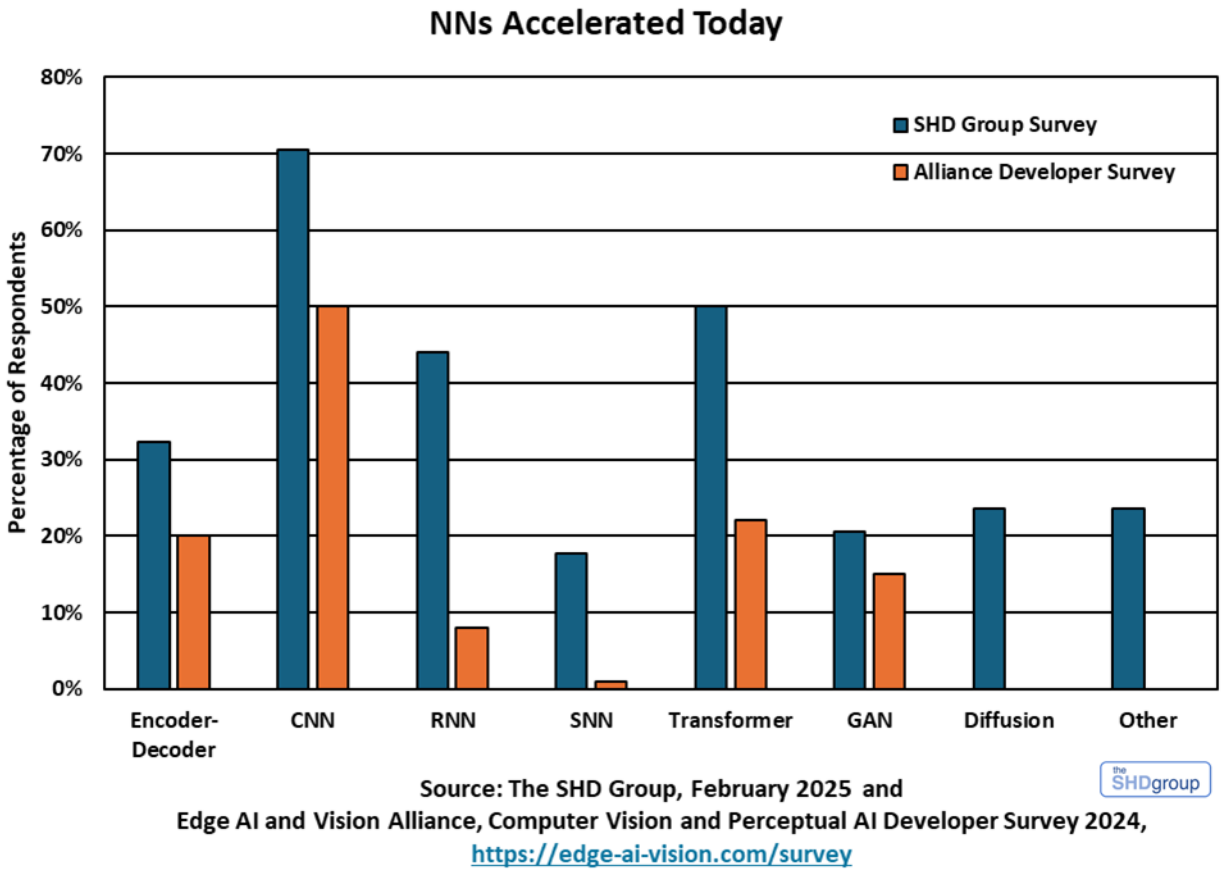
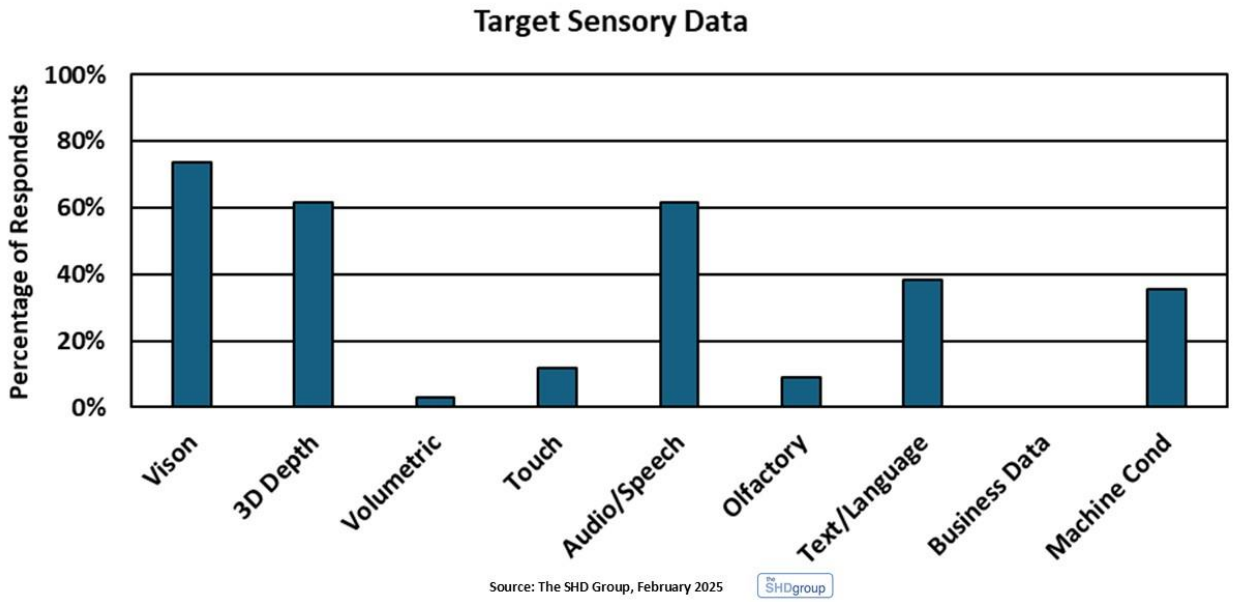


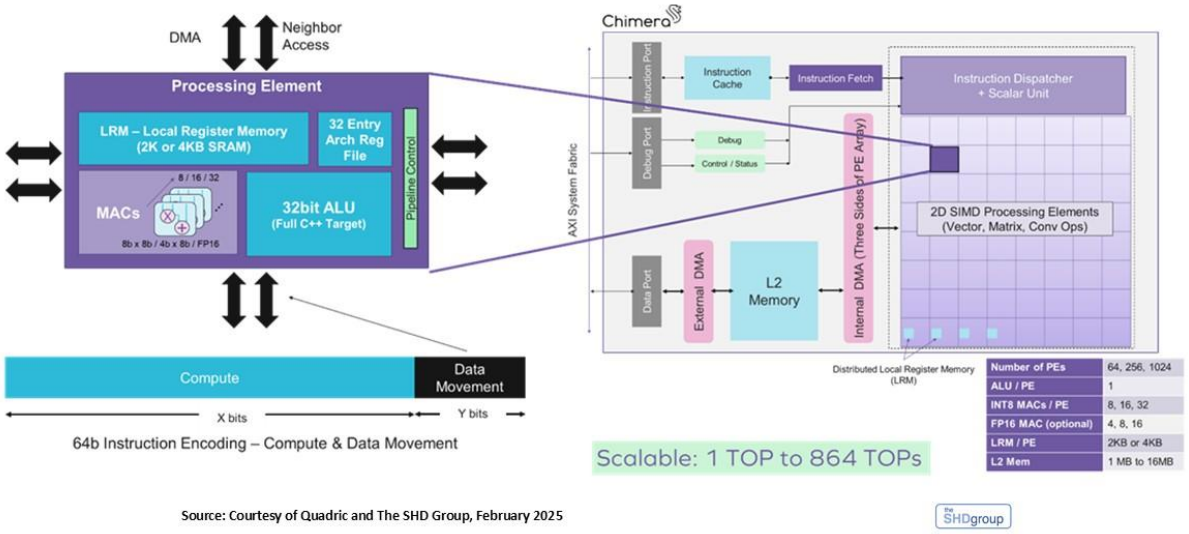
Figure 11: Which types of neural networks are you considering accelerating in the future



Although we didn't explicitly ask respondents to map network types to segments or to sensory data, we can join their responses to the separate questions and—with liberal interpretation—draw some conclusions.

Companies targeting auto universally support CNNs, reflecting the role of computer vision in automotive AI processing, as Figure 13 shows. These respondents are also slightly more likely to accelerate transformers, consistent with automotive companies being at the forefront of using this network type for vision applications, multimodal transformers having a role in driver assistance, and language models helping to computerize owners' manuals.

Figure 12: Quadric Chimera GPNPU Architecture

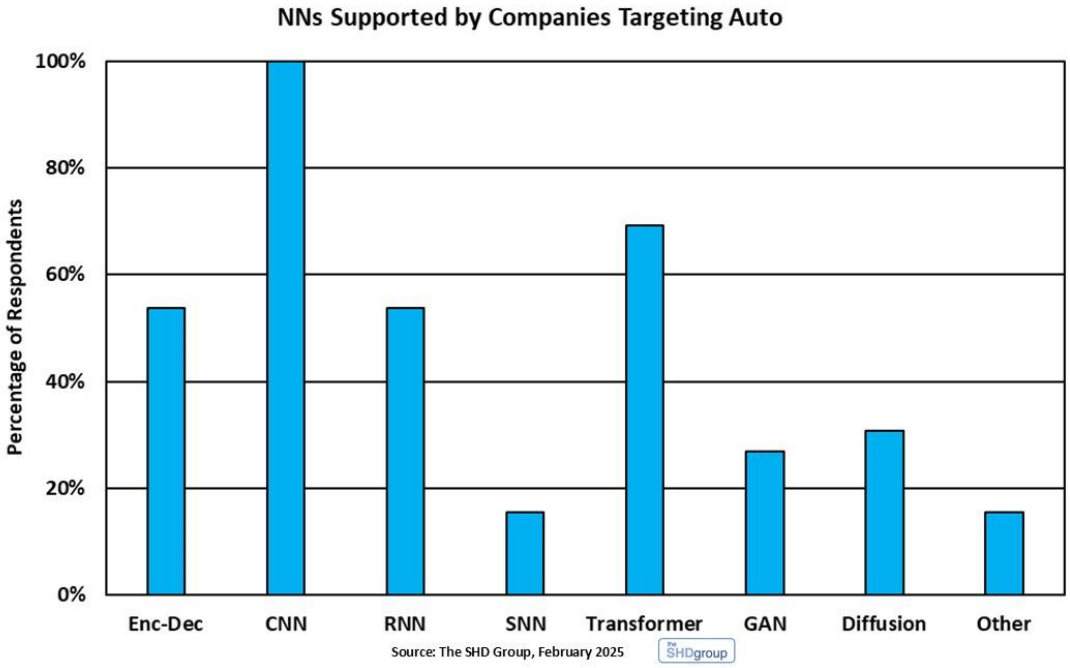


Source: Courtesy of Quadric and The SHD Group, February 2025

Quadric® is a semiconductor processor SIP licensing company delivering a unified hardware/software architecture optimized for on-device AI and DSP processing. The Chimera GPNPU is a revolutionary proprietary processor architecture designed for AI inference, combining the best of NPUs and DSPs into a single architecture that is fully programmable with C++/Python, delivering power-efficient on-device AI.

The Chimera GPNPU is a unified fine-grained processor architecture that merges the "DNA" of a systolic array with that of a conventional DSP. It eliminates the need for separate NPUs, DSPs, and real-time CPUs clustered into an AI subsystem. Quadric processors simplify system design and boost developer productivity. Chimera GPNPU Architecture: A hybrid between a CPU/DSP and a Systolic Array.

Figure 13: Which types of neural networks are you accelerating? (Automotive segment)



Taking the survey data we collected from companies targeting the Automotive market and applying it to the seven device types we analyzed, and the most popular Neural Networks being used, we have summarized it in the following table to show where each NN is most applicable in each device type for this market and how they perform in those devices.

Table 2 shows the applicability of the different NNs we surveyed for in the Automotive market and how they are applied in the seven devices we profiled.

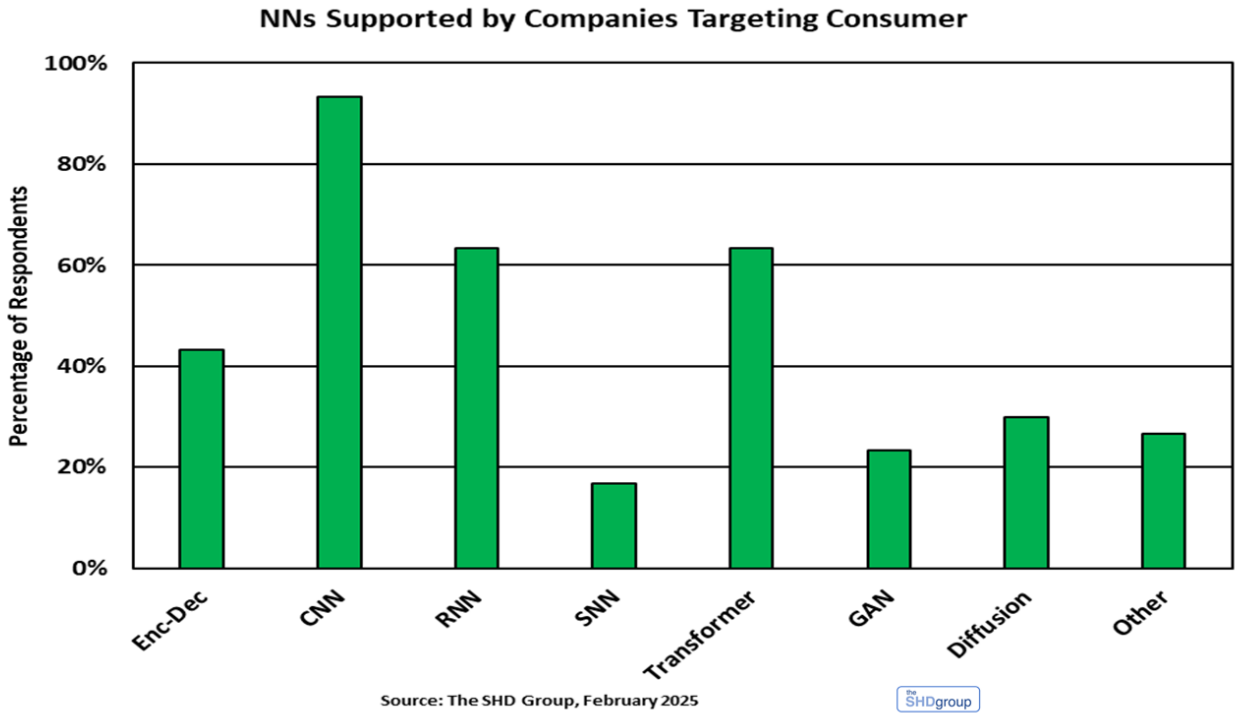
Table 2: Neural Network Usage by Device Type for The Automotive Market

Neural Network	CPUs	GPUs	NPUs	DSPs	FPGAs	MCUs	Custom ASICs / SoCs
Encoder-Decoder Architectures	Moderate (small models)	High	High	Moderate	High (small-medium models)	Low to Moderate	High
Convolutional Neural Networks (CNNs)	Moderate (lightweight models)	High (all scales)	High (optimized models)	Moderate (lightweight)	High (optimized)	Moderate (lightweight)	High
Recurrent Neural Networks (RNNs)	Moderate (simple RNNs/LSTMs)	Moderate (with parallelism)	High (sensor fusion/time-series)	Moderate (lightweight)	Moderate (optimized)	Moderate (lightweight)	High
Spiking Neural Networks (SNNs)	Low	Low	High (optimized)	Moderate	High	Moderate	High (with neuromorphic ASICs)
Transformers	Low (simplified models)	High	Moderate (optimized)	Low	Moderate (quantized)	Low	High (optimized)
Generative Adversarial Networks (GANs)	Low (inference only)	High	Moderate (inference only)	Low	Moderate (inference)	Low	Moderate to High
Diffusion Models	Low	High	Moderate (simplified)	Low	Low to Moderate	Very Low	Moderate to High

Source: The SHD Group, March 2025

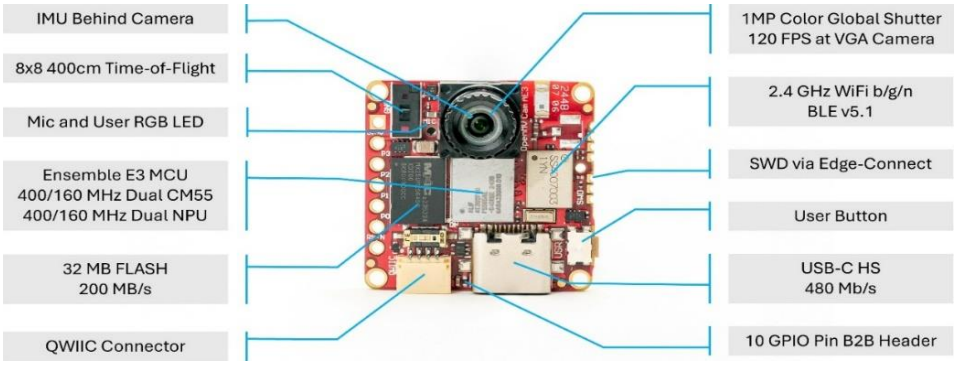
High: Neural Network optimized for use by this device type
Moderate: Small or moderately optimized versions of the NN for use by this device type
Low: low performance in this device type
Very Low: Poor performance in this device type

Figure 14: Which types of neural networks are you accelerating? (Consumer segment)



Taking the survey data we collected from companies targeting the Consumer market and applying it to the seven device types we analyzed, and the most popular Neural Networks being used, we have summarized it in the following table to show where each NN is most applicable in each device type for this market and how they perform in those devices.

Figure 15. OpenMV



OpenMV, a systems developer for Edge-AI applications, makes compact AI cameras powered by microcontrollers with neural processing unit accelerators onboard. The OpenMV AE3 is the smallest and lowest-power OpenMV Cam the company has made to date, featuring **dual-core processors** and **dual Ethos-U55 NPUs** that can be used simultaneously. The OpenMV AE3 can run object detection models like YOLO at **~30 FPS**, drawing only **60mA at 5V (0.25W)**, and in deep-sleep mode, it draws **<500uA at 5V (2.5mW)** for years of battery life while waking up on sound, motion, and a date/time.

The OpenMV AE3 is mighty; it can run object detection models like YOLO at **~30 FPS**, drawing only **60mA at 5V (0.25W)**, and in deep-sleep mode, it draws **<500uA at 5V (2.5mW)** for years of battery life while waking up on sound, motion, and a date/time.

Source: Courtesy of OpenMV and The SHD Group, March 2025

OpenMV specializes in Python-powered, machine vision modules that can track colors, detect faces and more. The OpenMV Cam is a general purpose MicroPython powered microcontroller platform with camera input and computer vision processing. In addition, OpenMV offers camera shields, vision sensors and camera lenses.

Table 3 shows the applicability of the different NNs we surveyed for in the Consumer market and how they are applied in the seven devices we profiled.

Incrementally more respondents supporting RNNs target Consumer, as Figure 14 shows, consistent with this area having more audio applications than other segments. Those targeting Computer, Industrial, and Networking reflect the overall distribution of supported network types, as Figure 16 and Figure 18 show.

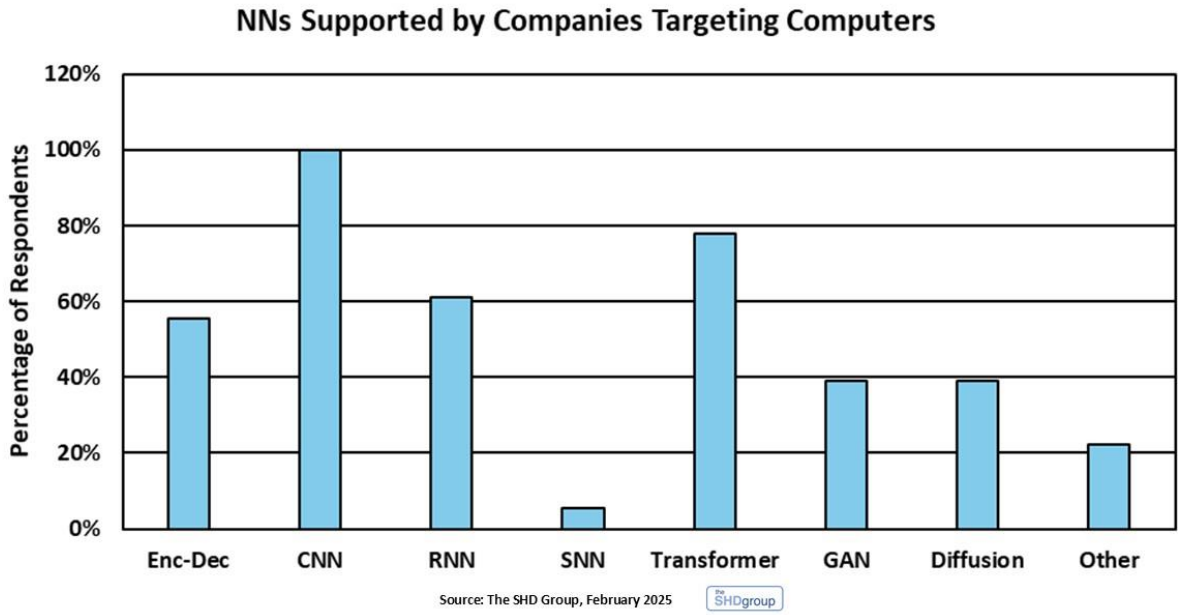
Table 3: Neural Network Usage by Device Type for the Consumer Market

Neural Network	CPUs	GPUs	NPUs	DSPs	FPGAs	MCUs	Custom ASICs / SoCs
Encoder-Decoder Architectures	Moderate (small models)	High	High	Moderate	High (small-medium models)	Low to Moderate	High
Convolutional Neural Networks (CNNs)	Moderate (lightweight models)	High (all scales)	High (optimized models)	Moderate (lightweight)	High (optimized)	Moderate (lightweight)	High
Recurrent Neural Networks (RNNs)	Moderate (simple RNNs/LSTMs)	Moderate (with parallelism)	High (sensor fusion/time-series)	Moderate (lightweight)	Moderate (optimized)	Moderate (lightweight)	High
Spiking Neural Networks (SNNs)	Low	Low	High (optimized)	Moderate	High	Moderate	High (with neuromorphic ASICs)
Transformers	Low (simplified models)	High	Moderate (optimized)	Low	Moderate (quantized)	Low	High (optimized)
Generative Adversarial Networks (GANs)	Low (inference only)	High	Moderate (inference only)	Low	Moderate (inference)	Low	Moderate to High
Diffusion Models	Low	High	Moderate (simplified)	Low	Low to Moderate	Very Low	Moderate to High

Source: The SHD Group, March 2025

High: Neural Network optimized for use by this device type
Moderate: Small or moderately optimized versions of the NN for use by this device type
Low: low performance in this device type
Very Low: Poor performance in this device type

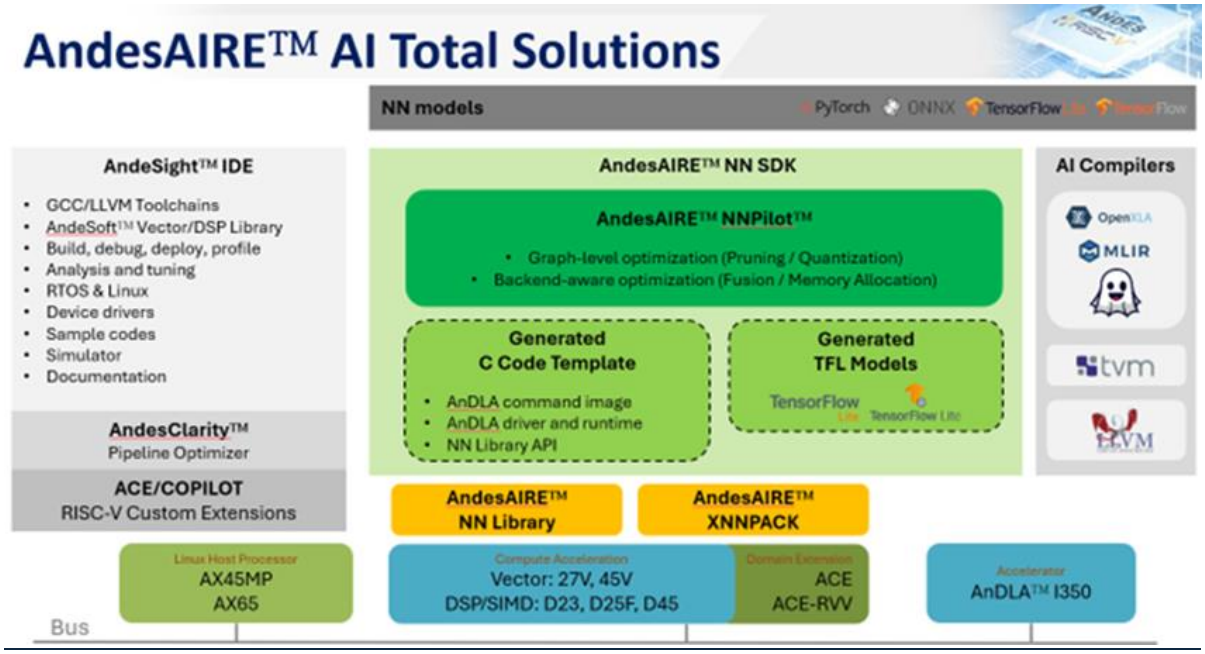
Figure 16: Which types of neural networks are you accelerating? (Computer segment)



Taking the survey data we collected from companies targeting the Computer market and applying it to the seven device types we analyzed, and the most popular Neural Networks being used, we have summarized it in the following table to show where each NN is most applicable in each device type for this market and how they perform in those devices.

Table 4 shows the applicability of the different NNs we surveyed for in the computer market and how they are applied in the seven devices we profiled.

Figure 17. Andes Technology Total AI Solutions



Source: Courtesy of Andes Technology and The SHD Group, March 2025

Andes Technology is a leading provider of high-performance, low-power RISC-V processor IPs, empowering AI, IoT, automotive, and embedded applications. With a robust hardware-software ecosystem, Andes delivers scalable solution, including vector processing, deep learning accelerators, and security-enhanced cores.

AndesAIRE™: A Full-Stack AI Solution

AndesAIRE™ is an AI platform combining Andes' RISC-V semiconductor IPs with neural network (NN) software. Customers can license specific components.

- **Software Stack:** Includes an NN SDK and optimized libraries (NN Library, XNNPACK) for efficient NN deployment on Andes-based IP.
- **Hardware Targets (Semiconductor Soft IPs):**
  - AnDLA (Andes Deep Learning Accelerator): Scalable up to 8 TOPs per core @ 1GHz, handling MAC/Conv/GEMM computations.
  - Andes RISC-V Vector Processors: Provides up to 512 GOPs per core @ 1GHz, ensuring flexible, efficient computing for non-linear or future-proof operations.

Table 4: Neural Network Usage by Device Type for the Computer Market

Neural Network	CPUs	GPUs	NPUs	DSPs	FPGAs	MCUs	Custom ASICs/SoCs
Encoder-Decoder Architectures	Moderate (small models)	High (video editing, NLP)	High (speech, video)	Moderate (lightweight)	High (media processing)	Low to Moderate	High (optimized NLP and video)
Convolutional Neural Networks (CNNs)	Moderate (small models)	High (image processing, gaming)	High (optimized models)	Moderate (lightweight)	High (customized tasks)	Low	High (optimized for tasks)
Recurrent Neural Networks (RNNs)	Moderate (small RNNs)	Moderate (sequential data)	High (text/audio tasks)	Moderate (lightweight)	Moderate (optimized)	Low	High (real-time inference)
Spiking Neural Networks (SNNs)	Low	Low	High (neuromorphic tasks)	Moderate	High	Low	High (specialized tasks)
Transformers	Low (simplified models)	High (NLP, AI assistants)	High (large-scale NLP)	Low	Moderate (simplified)	Low	High (speech and fusion tasks)
Generative Adversarial Networks (GANs)	Low (inference only)	High (art/media generation)	Moderate (light inference)	Low	Moderate (specialized)	Low	High (image/video synthesis)
Diffusion Models	Low	High (creative tools)	Moderate (optimized)	Low	Moderate (simplified)	Very Low	High (synthetic data/image generation)

Source: The SHD Group, March 2025

High: Neural Network optimized for use by this device type
Moderate: Small or moderately optimized versions of the NN for use by this device type
Low: low performance in this device type
Very Low: Poor performance in this device type

Figure 18: Which types of neural networks are you accelerating (Industrial segment)

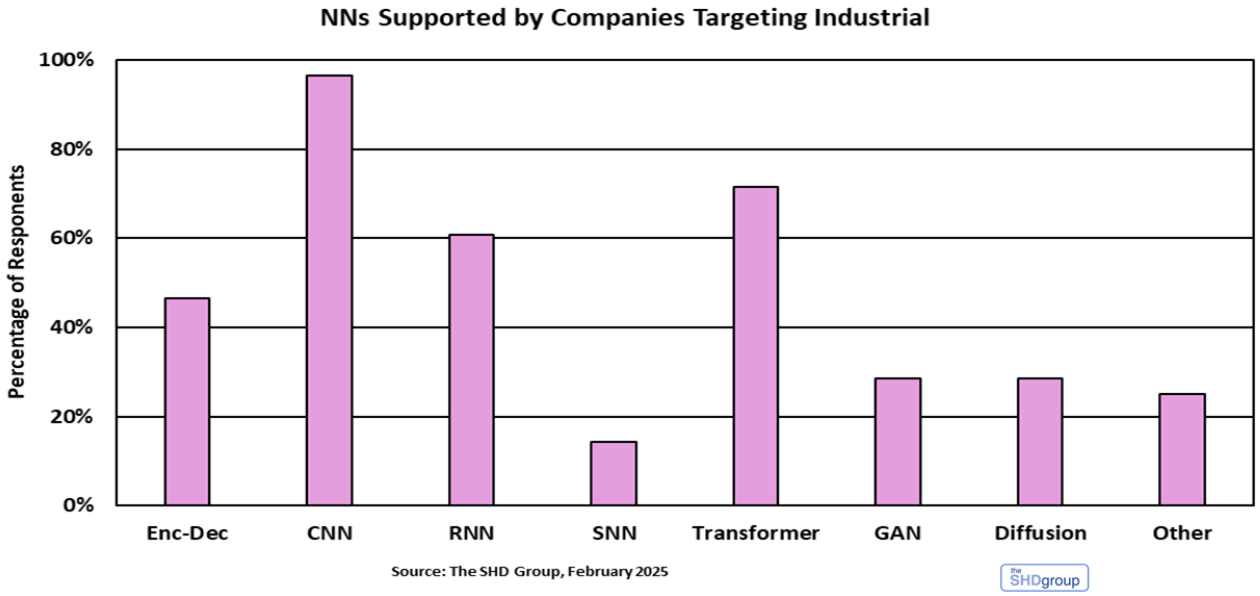


Figure 19: Inuitive 3D Computing Processor Chips

Source: Courtesy of Inuitive and The SHD Group, March 2025

Inuitive is a semiconductor and module vendor and has introduced the N4100, multicore SoC, supporting high-quality 3D imaging, deep learning, SLAM accelerators and computer vision processing for augmented and virtual reality, drones, robots and many other applications.

Table 5 shows the applicability of the different NNs we surveyed for in the Industrial market and how they are applied in the seven devices we profiled.

Table 5: Neural Network Usage by Device Type for the Industrial Market

Neural Network	CPUs	GPUs	NPUs	DSPs	FPGAs	MCUs	Custom ASICs/SoCs
Encoder-Decoder Architectures	Moderate (small models)	High (video editing, NLP)	High (speech, video)	Moderate (lightweight)	High (media processing)	Low to Moderate	High (optimized NLP and video)
Convolutional Neural Networks (CNNs)	Moderate (small models)	High (image processing, gaming)	High (optimized models)	Moderate (lightweight)	High (customized tasks)	Low	High (optimized for tasks)
Recurrent Neural Networks (RNNs)	Moderate (small RNNs)	Moderate (sequential data)	High (text/audio tasks)	Moderate (lightweight)	Moderate (optimized)	Low	High (real-time inference)
Spiking Neural Networks (SNNs)	Low	Low	High (neuromorphic tasks)	Moderate	High	Low	High (specialized tasks)
Transformers	Low (simplified models)	High (NLP, AI assistants)	High (large-scale NLP)	Low	Moderate (simplified)	Low	High (speech and fusion tasks)
Generative Adversarial Networks (GANs)	Low (inference only)	High (art/media generation)	Moderate (light inference)	Low	Moderate (specialized)	Low	High (image/video synthesis)
Diffusion Models	Low	High (creative tools)	Moderate (optimized)	Low	Moderate (simplified)	Very Low	High (synthetic data/image generation)

Source: The SHD Group, March 2025

High: Neural Network optimized for use by this device type
Moderate: Small or moderately optimized versions of the NN for use by this device type
Low: low performance in this device type
Very Low: Poor performance in this device type

Figure 20: Which types of neural networks are you accelerating? (Networking segment)

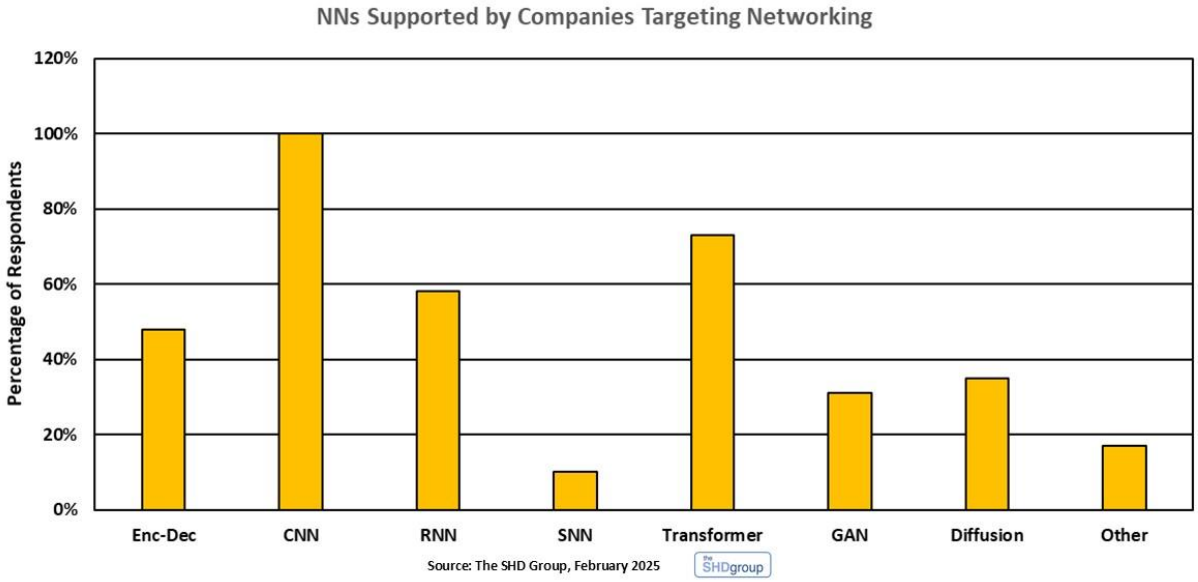


Table 6 shows the applicability of the different NNs we surveyed for in the Industrial market and how they are applied in the seven devices we profiled.

The neural network types accelerated also correlate with the sensory data targeted, as Figure 22 shows, although this is a liberal interpretation, as noted above.

AI is a core driver of Andes' growth, with 38% of its business coming from AI applications. Leading companies deploy Andes IPs for AI acceleration from the far edge to the datacenter.

Optimized AI Acceleration

Andes' high-performance RISC-V CPUs excel in AI workloads with:

- Advanced vector processing for AI inferencing.
- Automated Custom Extensions (ACE) Framework to:
  - Seamlessly integrate AI acceleration.
  - Enhance CPUs with custom instructions for optimized performance.
  - Accelerate non-linear operations and high-speed control.

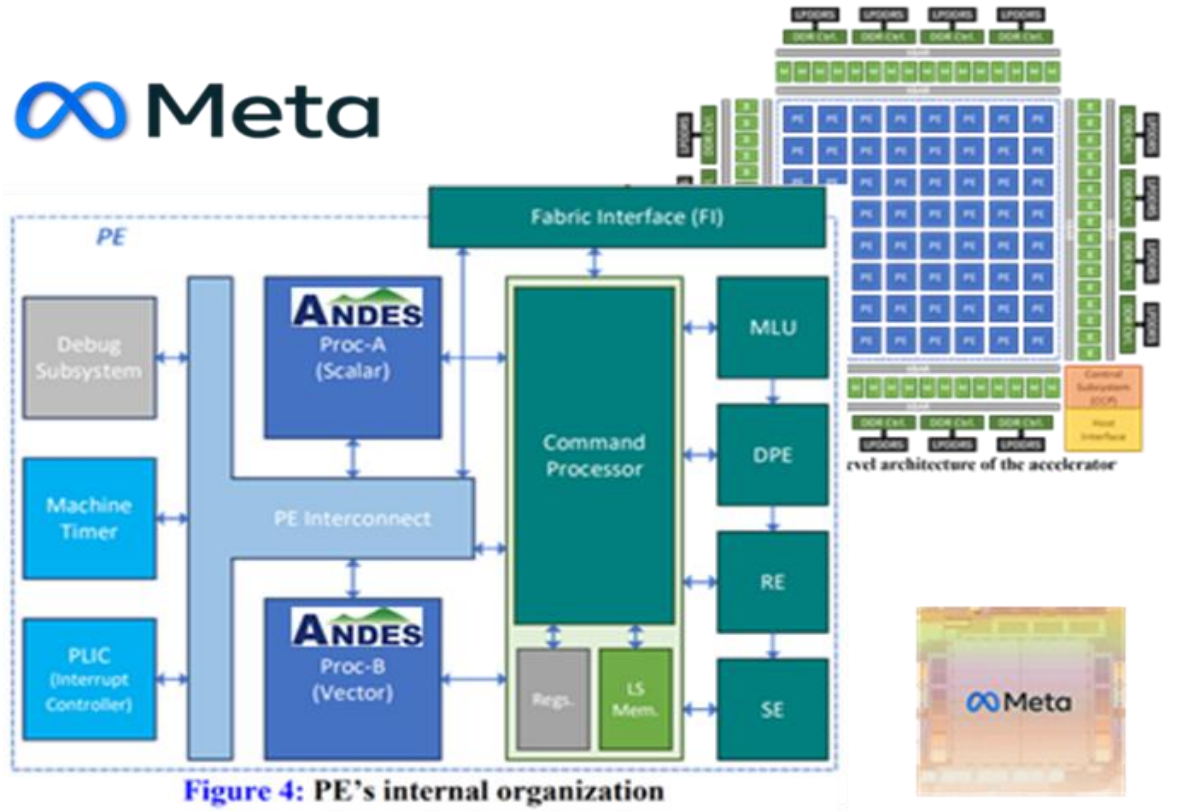
This approach ensures efficiency and adaptability, allowing AI models to evolve without hardware redesign.

Proven AI Deployments

Andes AI technology powers leading AI accelerators, including:

- RAIN.AI – Showcasing Andes-powered AI acceleration.

Figure 21: Andes Technology Custom Extensions and Meta Engagement
(https://www.youtube.com/watch?v=TYmhLZmtlhc)



Source: Curtesy of Andes Technology and The SHD Group, March 2025

- ISCA 2023 paper, "MTIA: First Generation Silicon Targeting Meta's Recommendation Systems

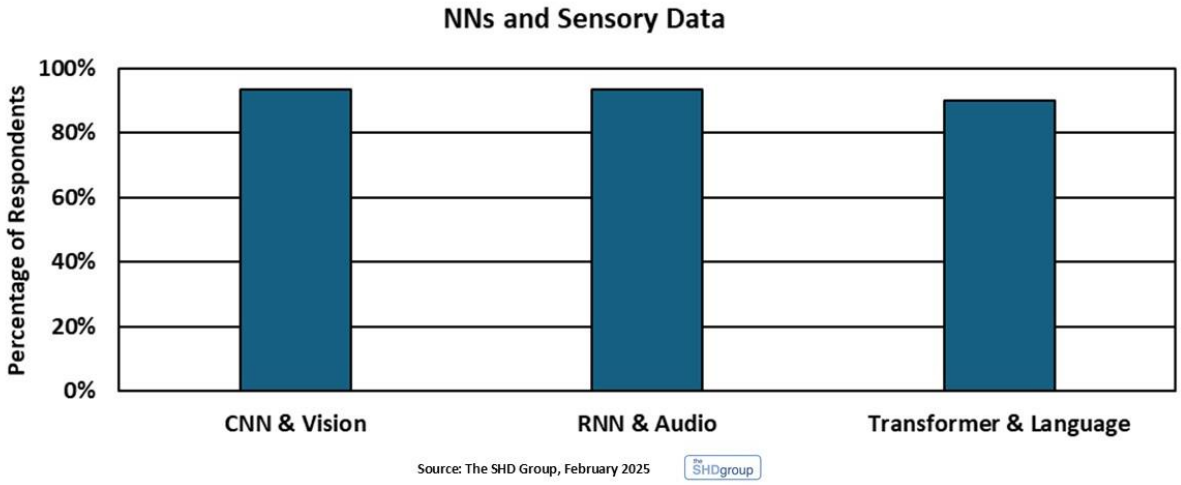
Table 6: Neural Network Usage by Device Type for the Networking Market

Neural Network	CPUs	GPUs	NPUs	DSPs	FPGAs	MCUs	Custom ASICs/SoCs
Encoder-Decoder Architectures	Moderate (small models)	High (NLP for chatbots)	High (traffic translation, parsing)	Moderate	High (signal processing)	Low to Moderate	High (optimized routing protocols)
Convolutional Neural Networks (CNNs)	Moderate (lightweight models)	High (traffic monitoring, image-based tasks)	High (optimized for vision tasks)	Moderate (lightweight)	High (optimized for vision)	Moderate (basic tasks)	High (optimized for routing/security)
Recurrent Neural Networks (RNNs)	Moderate (simple RNNs/LSTMs)	Moderate (sequential data)	High (network traffic prediction)	Moderate (lightweight models)	High (optimized for sequential tasks)	Moderate (lightweight)	High (real-time data processing)
Spiking Neural Networks (SNNs)	Low	Low	High (low-power IoT devices)	Moderate (low-power tasks)	High (event-driven systems)	Low	High (energy-efficient tasks)
Transformers	Low	High (NLP for chatbots, anomaly detection)	High (complex data fusion)	Low	Moderate (simplified implementations)	Low	High (real-time network optimization)
Generative Adversarial Networks (GANs)	Low (inference only)	Moderate (synthetic data)	Moderate (light inference)	Low	Moderate (anomaly detection, synthesis)	Low	High (security and traffic simulation)
Diffusion Models	Low	High (advanced traffic simulation)	Moderate (simplified)	Low	Moderate (synthetic data)	Very Low	High (synthetic traffic patterns)

Source: The SHD Group, March 2025

High: Neural Network optimized for use by this device type
Moderate: Small or moderately optimized versions of the NN for use by this device type
Low: low performance in this device type
Very Low: Poor performance in this device type

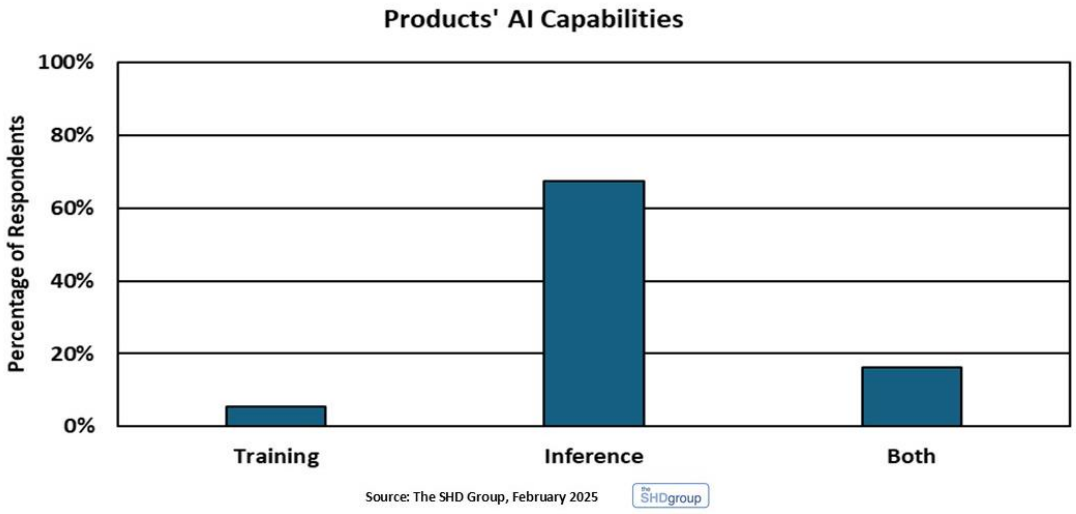
Figure 22: Which types of neural networks are you accelerating? (Vision, audio, and language data)



Frameworks and Benchmarks

At the edge, our interview findings supported the thesis that most companies seek to accelerate inference. Our survey findings bear this out as well, although a surprising number support training or both training and inference, as Figure 23 shows. Reviewing our survey results, we believe at least one respondent erroneously answered Training when it should have answered Inference. Other Training respondents offer services and software, not chips or SIP.

Figure 23: Which AI capabilities are your products geared toward?



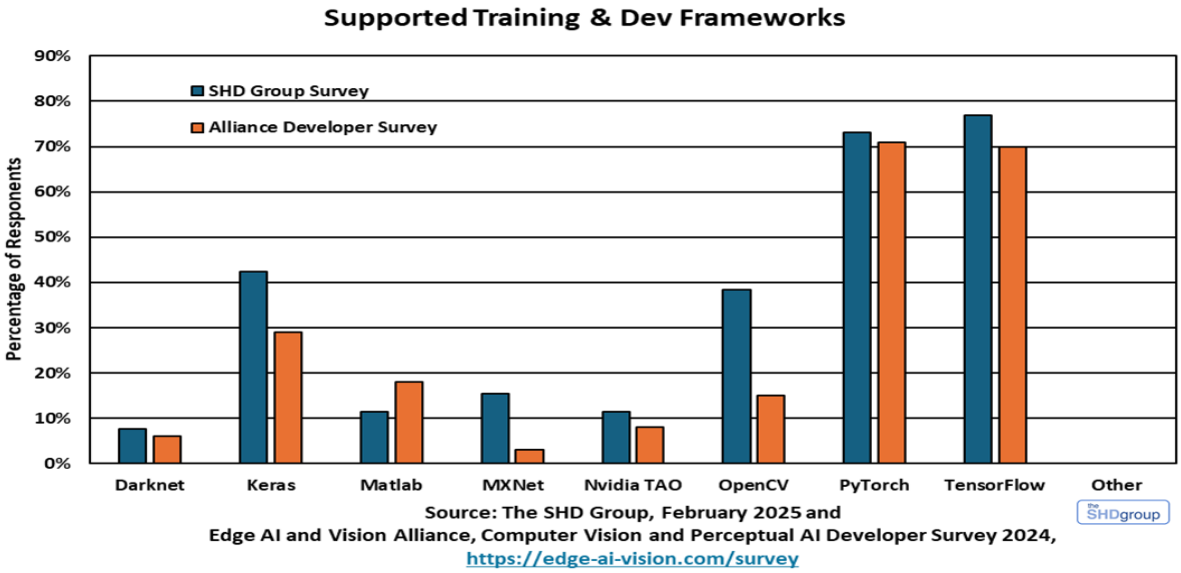
Nonetheless, products focused on inference must ingest trained models. Question 9 asked about supported training and development frameworks, libraries, and tools. The developer survey has a similar question, and the survey's findings are similar. The developer survey includes classic computer vision, which may account for its respondents' greater propensity to use Matlab.

Our survey revealed disproportionate support for Keras and OpenCV. The latter supports both classic computer vision and neural networks but received little attention in our interviews. In light of the developer survey's

findings, respondents to our survey should consider deemphasizing OpenCV. Keras aligns with Python establishing itself as the dominant language for AI-software development, at least in the exploration phase.

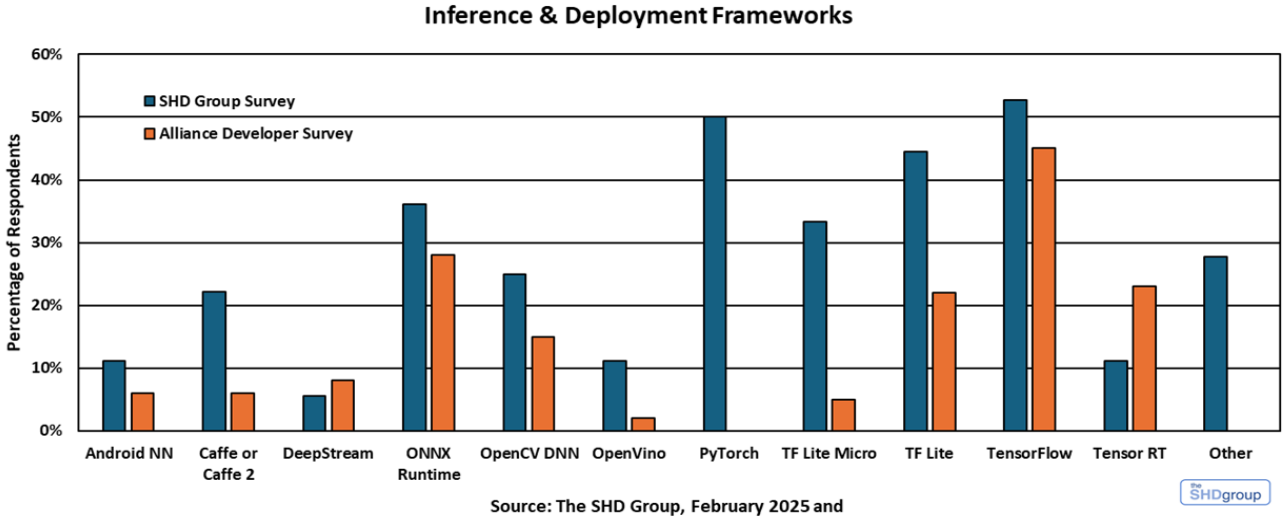
Reflecting our interview findings, both surveys show widespread PyTorch and TensorFlow backing, as Figure 24 shows.

Figure 24: Which training and development frameworks/libraries/tools do you support?



As we found in our interviews, most chip and SIP companies enable inference independent of a standard framework. A typical workflow they expect their customers to follow is to train their networks with arbitrary tools (typically PyTorch, TensorFlow, or TensorFlow Lite), convert the trained network to ONNX format, and then use the chip/SIP companies' tools to prepare and deploy the network to run on their technology. Nonetheless, standard tools and inference engines are available. Our respondents generally claim to use at least one, as Figure 25 shows.

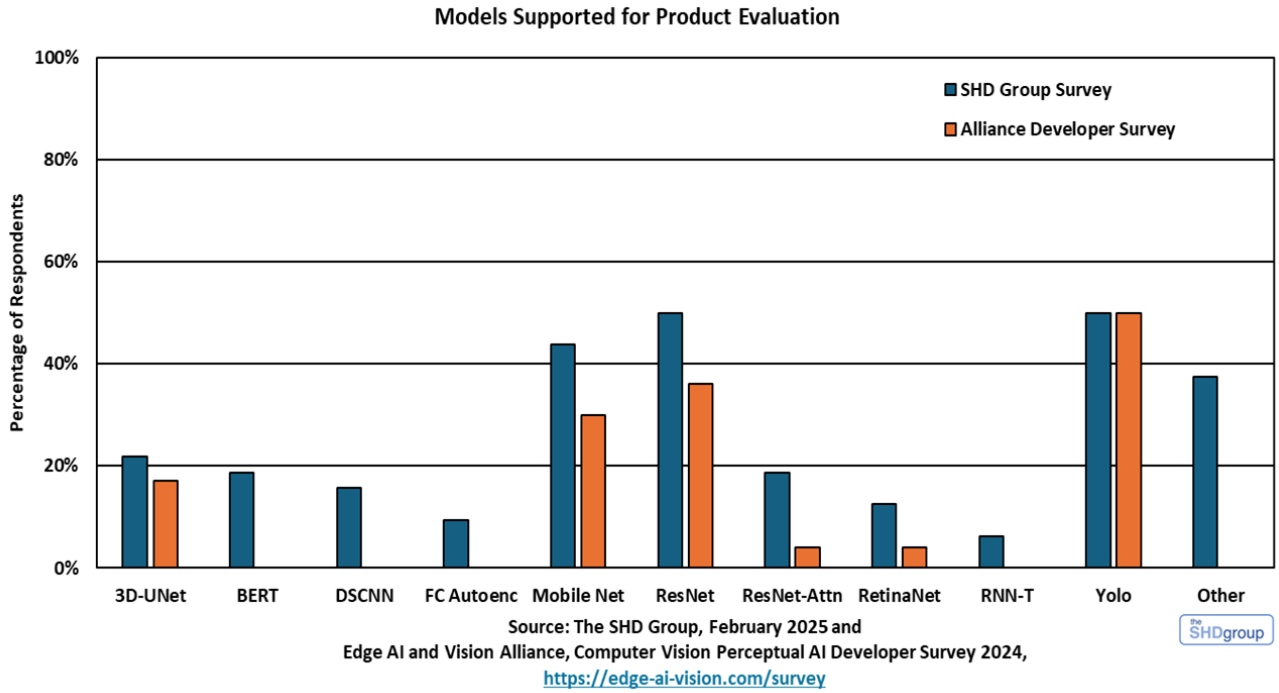
Figure 25: What inference and deployment frameworks/libraries/tools do you use with your products?



Responses mostly track the developer survey, but a few exceptions stand out. Both show widespread support for ONNX Runtime. We suspect that these respondents are confusing ONNX Runtime with the ONNX format. More than 20% of developers employ TensorRT, which is from Nvidia, the leading AI chip company. From our survey, we see chip, SIP, and software companies stating they support it. The developer survey didn't ask about PyTorch for inference. Although our survey did, we are skeptical that respondents would use it in the development phase, much less for deployment. Despite their responses, we suspect they only support PyTorch indirectly for training. As our interviews revealed, AI-technology suppliers support specific models to communicate their products' performance and help customers evaluate them. Question 13 of our survey asks about specific models chosen from among those referenced in the developer survey or used in MLPerf benchmarks. Various CNN models rank highly on both surveys, including 3D-UNet, MobileNet, ResNet, and Yolo. The lattermost is unfortunately ambiguous given the disparate Yolo versions.

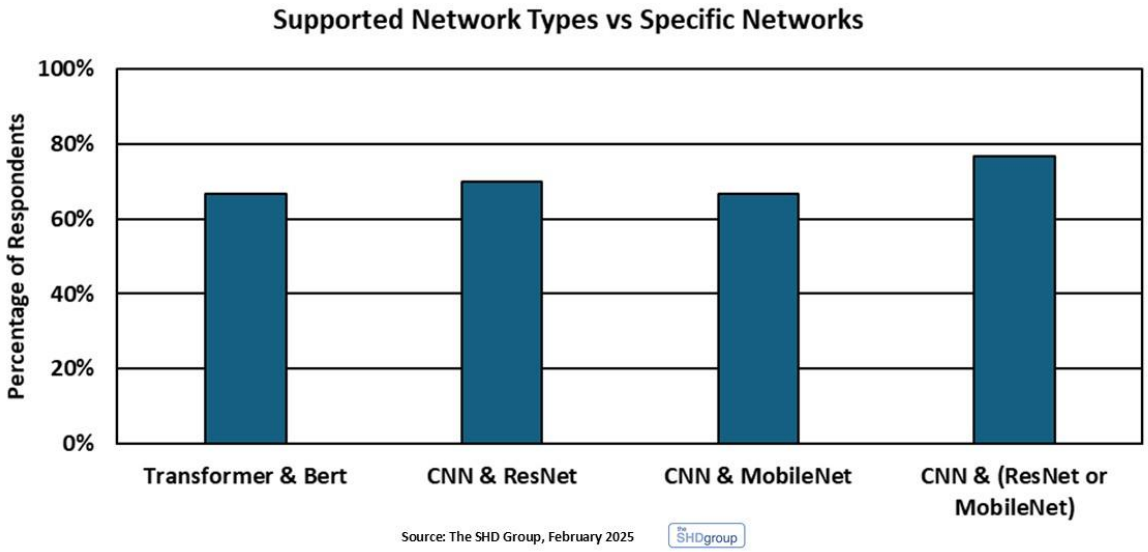
A key finding is that despite suppliers touting transformer support, few respondents support Bert, as Figure 26 shows. Although it is dated, it's a standard transformer benchmark. Our findings also suggest our respondents could probably eliminate supporting ResNet-Attention and RetinaNet.

Figure 26: Which models do you support for product evaluation?



Although we didn't explicitly ask respondents to map network types to supported networks, we can join their responses to the separate questions and—with liberal interpretation—draw some conclusions. As alluded to above, the proportion of respondents claiming to support transformers and Bert is low, only about 70%. Surprisingly, the proportion claiming to support CNNs and either ResNet or MobileNet is also low, about 75%, as xxx shows. Despite their flaws, these are clearly standard benchmarks serving a role analogous to Dhrystone for CPUs.

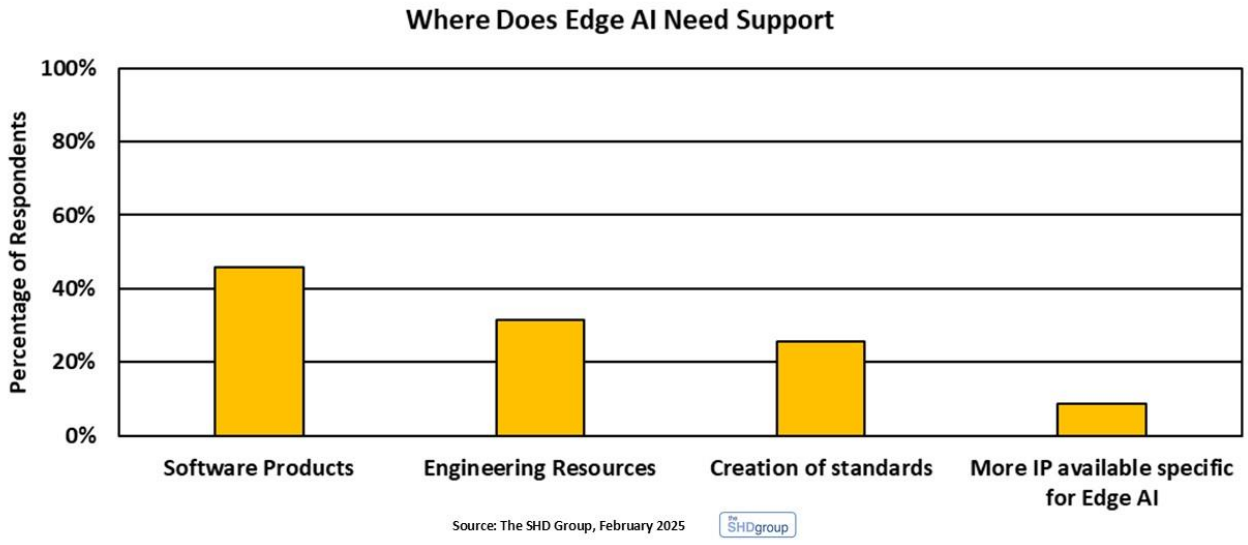
Figure 27: For specific network types, which models do you support for evaluation



Looking Forward

Our survey's final question asked respondents to look forward and consider where they would like to see the Edge-AI ecosystem direct more effort. Perhaps reflecting the preponderance of chip and SIP suppliers in our sample, relatively few are asking for more Edge-AI-SIP, as Figure 28 shows. The top vote getter was software, indicating that few embedded design teams will have data scientists on staff or programmers skilled in AI.

Figure 28: Where would you like to see more support directed into the Edge-AI Ecosystem?



Additional Discussion on Neural Networks

One of the very interesting facets of the work occurring around Neural Networks today is the constant evolution and improvement in the performance of these algorithms. Advancements in the field are moving very quickly with silicon architectural advancements moving equally as fast to accommodate the new neural networks being created and the applications that will use them.

A case in point is the recent trend towards multimodal Transformers and LLMs that combine two or more different neural networks such as video and touch or video, audio and touch, creating the capability to accomplish a set of pre-defined actions in reaction to an event or condition instead of only reporting on that event or condition. This is the basis towards the introduction of Agentic (AI agent) applications that will soon be available on PCs, Laptops, Tablets, Smartphones, Automotive and many other systems, all performing AI in Edge devices.

To accommodate this evolutionary step, SoC device architectures are becoming more complex with higher performance levels and capability as the neural networks become more refined and the silicon architectures become more defined.

There is historical precedent for this in the semiconductor market as seen in the advancement of PCs and other systems becoming capable of video playback starting in the late 1980s. As silicon architecture became more capable, it took on more functionality, at ever-lowering cost structures. As part of this evolutionary path where previously an entire PCB with many parts was needed, due to refinements in software, creation of standards and advancements in silicon architectures, the parts count and footprint were reduced down to a single part or even to zero parts through use of more capable silicon and software.

Today, no one who is buying a smartphone, PC, Laptop or tablet even asks if those systems will be able to have video playback capabilities. It is just a given that they do so.

The following table shows this evolution:

Table 7: Timeline foe Evolution of Video playback Capability

1985	Intel 386 introduced 32bit computing enabling more robust multimedia applications
1986	VGA cards enabled 256-colod displays at 320x200 resolution, improving video output
1987	VGA cards introduced with 640x480 resolution and 256 colors; card size reduced
1991	MPEG-1 standards allowed digital and audio compression suitable for PC playback
	Early MPEG-1 video playback required add-on MPEG Daughter decoder cards
1993	Intel Pentium CPU significantly improved floating-point calculations for video decoding
	A single CPU now handled tasks that previously required separate hardware
1995	Windows 95 bundled multimedia APIs and support for MPEG-1. Daughter cards specialized for MPEG-2 playback
1996	MPEG-2 became the standard for DVD playback. DVD-ROM drives started appearing in PCs
	DVD playback introduced higher resolutions (720x480) requiring external DVD decoder cards
1998	Integration of video decoding in graphics chipsets
2000	Introduction of DivX Codec; GPU cards now included built-in hardware for video decoding – eliminating need for separate decoder cards
2003	H.264 standard provided high-quality video compression with low bit rates. Intel integrated GPU functionality into their CPU – entire video processing moved into a single CPU package
2004	GPU-based video acceleration – offloaded video decoding from the CPU, enabled video decoding in a single chip
2008	Integration of GPU functionality into CPU
2010	VPS Codec and WebM standard, multicore CPUs handle video playback through software decoding
2013	H.265 standard (HVEC). GPUs included specialized hardware blocks for decoding H.265 to play 4K content
2015	VP9 Codec – introduced by Google for high-efficiency streaming on multiple platforms
2017	AV1 Codec -open source codec with superior compression for 4K and 8K video
2020	AI assisted video decoding enabled smooth 8K video playback
2023	AV1 hardware decoding support became standard in SoCs and GPUs – enabled ultra-compact devices like smartphones, laptops
1980s	Full PCs required for basic video
1990s	Dedicated decoder cards shrank but still needed expansion slots
2000s	Decoding integrated into GPUs and CPUs enabling compact systems
2010s	SoCs and miniaturization allowed high-quality video playback in tablets / other handheld devices
2020s	Ultra-HD playback in handheld devices with AI enhanced performance

Source: The SHD Group, March 2025

The point to be made here is that, as the silicon architecture for video evolved, they became more capable of running more functions. As the functionality increased, the number of discrete parts necessary to obtain the same level of functionality decreased. Over time, the market settled on a few different types of encoding / decoding software applications, most of which became standards.
However, this is not the case with Neural Networks today as there are literally hundreds of different NNs in the market now that can be used for AI.

So, the key question for device manufacturers, SIP companies and software developers today is which neural networks are going to be the most in demand for the various applications Edge-AI solutions are created for. Most of the systems where inference for Edge-AI is being performed are either low-cost or have power constraints. While an Edge-AI silicon architecture can be created to be able to run all the neural networks in the market today, such an architecture would probably be more expensive and power hungry than what most Edge-AI applications can tolerate. A more focused approach is necessary to meet the requirements of the target applications. This limits how much performance, and capability can be built into the silicon for a given range of applications.

System developers would typically look for the best solution to build their system around by looking at several parameters that are important for the functioning of the system, are important to their customers in their

applications, are within the expertise of the system company and that are cost-competitive within the target market.

The following table lists some of the parameters that were noted by our survey respondents. It is not an all-inclusive list, but gives some perspective on how companies choose between different approaches and solutions

As Table 8 shows, the parameters shift between different device types based on their individual strengths and weaknesses and the demands of the specific markets the solution is geared toward. Given the range of applications that have emerged in the Edge-AI market, there can be more than one device type that can be applied to the system to perform the necessary functions demanded by the market.

Table 8: Different Parameters for Choosing a Silicon Solution by Device Type

Feature	CPU	GPU	NPU	DSP	MCU	FPGA	Custom ASIC/SoC
Best For	General-purpose computing, some AI tasks	AI model training, high-performance AI workloads	AI inference, low-power AI applications	Signal processing, real-time AI computations	Simple embedded AI applications	AI acceleration, reconfigurable AI workloads	Highly optimized AI applications, mass production
Cost	Low to Medium (Widely available)	High (AI-optimized GPUs are expensive)	Variable (Low for embedded NPUs, high for enterprise-class NPUs)	Low to Medium (Efficient, but not as widely used for AI)	Low (Very affordable for simple AI tasks)	High (Development cost is high, but mass production can be cost-effective)	Very High (Significant upfront cost, but low per-unit cost at scale)
Power Efficiency	Low (Consumes high power for AI tasks)	Medium (Power-hungry, but efficient for training)	High (Optimized for low-power AI inference)	High (Designed for efficient real-time signal processing)	Very High (Designed for low-power embedded systems)	Medium (Can be power-efficient depending on design)	Very High (Optimized for specific workloads)
AI Inference Performance	Slow (Software-optimized AI inference)	High (Accelerated inference but power-hungry)	Very High (Designed specifically for inference)	High (Optimized for real-time AI computations)	Low (Can run small AI models but very slow)	Variable (Depends on FPGA implementation)	Very High (Designed for AI inference at maximum efficiency)
AI Training Performance	Poor (Slow for deep learning models)	Excellent (Best for training large neural networks)	Limited (Mostly designed for inference, not training)	Low (Not suitable for AI training)	Poor (Not designed for AI training)	Moderate (Can be programmed for AI, but not ideal for large-scale training)	High (Custom-designed for AI training in some cases)
Parallel Processing	Low (Few cores, optimized for serial tasks)	High (Thousands of cores for parallel tasks)	Very High (Optimized for tensor operations)	Medium (Parallel SIMD operations for DSP)	Low (Single or few cores)	High (Custom parallelism based on design)	Very High (AI-optimized, custom-designed for workload)
Memory Bandwidth	Low (Uses DDR4/DDR5 RAM)	High (Uses HBM, GDDR6 for faster data access)	Medium to High (On-chip SRAM or LPDDR)	Medium (Optimized memory access for DSP workloads)	Low (Embedded Flash/RAM)	Variable (Depends on FPGA architecture)	Very High (Optimized memory for AI models)
Ease of Programming	High (Well-known architecture, optimized for software development)	Moderate (Requires CUDA, OpenCL, or AI SDKs)	Low (Requires specialized SDKs and AI frameworks)	Moderate (DSP programming knowledge required)	High (Simple software programming)	Low to Moderate (Requires expertise in FPGA programming)	Very Low (Fixed-function, not reprogrammable, requires hardware expertise)
Software Ecosystem	Excellent (Broad compatibility with AI frameworks)	Excellent (CUDA, ROCm, TensorRT, AI frameworks)	Limited (Vendor-specific frameworks, e.g., Edge TPU, Intel Movidius)	Moderate (DSP-optimized AI libraries)	Limited (MCU AI frameworks exist but limited)	Low to Moderate (Custom programming required)	Vendor-Specific (Highly optimized but fixed)

Source: The SHD Group, March 2025

VIII. Effects of AI on SoC Designs

While this report focuses on the Edge-AI market and not the entire AI market, it is essential to briefly address this burgeoning technology within the semiconductor industry. The SHD Group underscores the significance of AI as a compelling opportunity for SoC designs in the foreseeable future, anticipating it to be a primary driving force for all SoCs in the market for an extended period.

The advent of AI is reshaping the landscape of contemporary silicon design and the utilization of AI functionalities in various applications. A noteworthy aspect of this evolution is the current classification of any chip featuring AI capabilities as an AI device. The SHD Group observed that this categorization stems from the early stages of the market development. As device complexity continues to escalate, architectural definitions become more precise and AI algorithms become more sophisticated, our perceptions of these devices will evolve over time.

Presently, low-complexity devices executing only a limited set of AI functions are deemed cutting-edge. However, as performance enhancements unfold in the coming years, even budget-friendly devices are anticipated to incorporate a diverse array of AI functionalities while maintaining cost-effectiveness. Instances already exist where devices can concurrently perform multiple AI functions at Levels 1 and 2, and this trend is poised to expand in the future.

Figure 29: Different Levels of AI Functionality

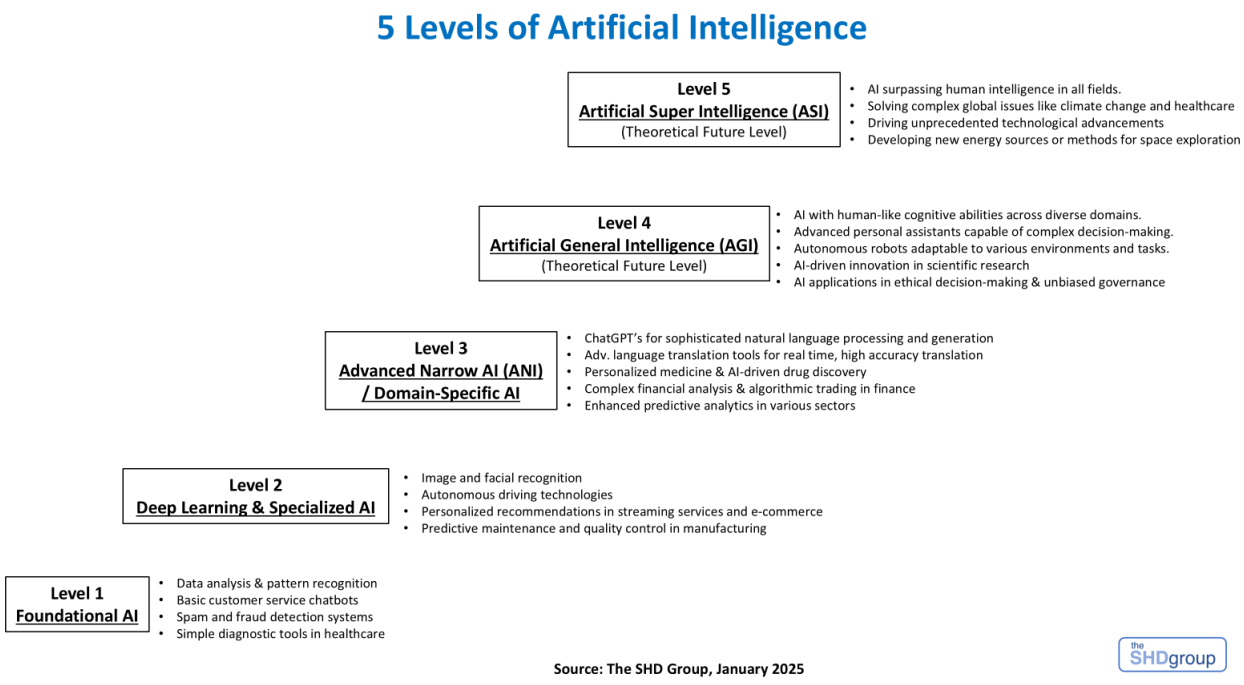


Figure 29 represents The SHD Group's perspective on the potential evolution of AI anticipated over the next several years as a series of levels. In it, we outline existing functions achievable from Level 1 "Foundational" capabilities and project movement towards Artificial Super Intelligence (ASI) at Level 5. There are many ways to try and illustrate the levels in the evolution of AI, but we have created the above approach as a somewhat simple way to look at progress – and consider the impact on SoCs in general and markets and applications in particular. It is at this point that the evolution of the technology takes over and helps produce something compelling and potentially extremely useful to people. The largest LLMs today are exceeding 1 trillion parameters in order to accomplish and conduct human-like conversations and responses. This is truly amazing but does not satisfy the

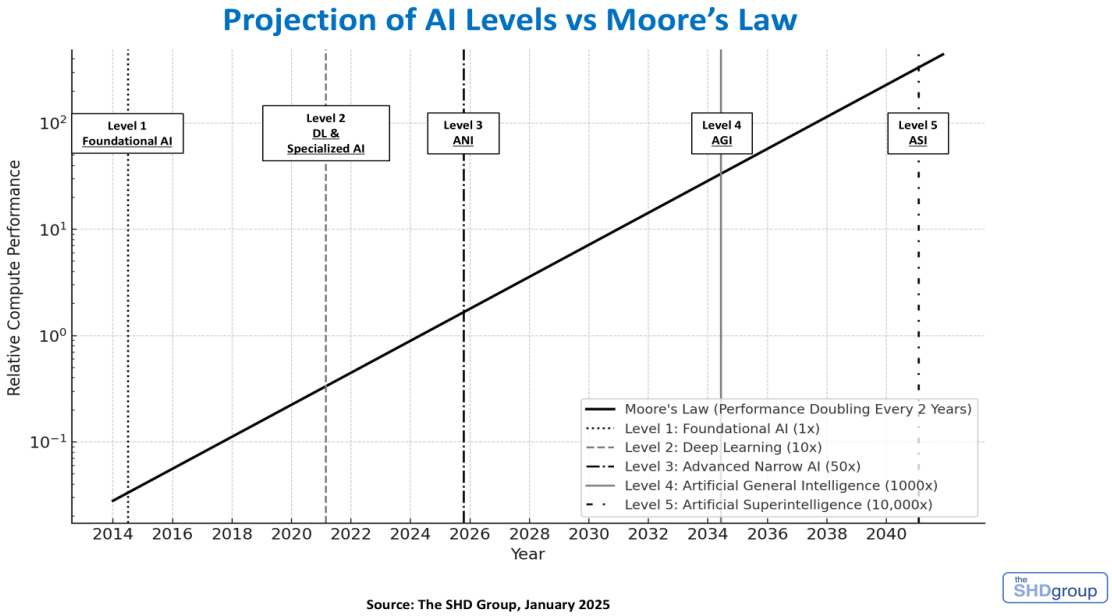
market need to be able to run these applications on our personal or mobile devices where the real demand resides.

If these applications can only ever run while connected to the Cloud due to the enormous number of parameters to be executed, the potential benefits of LLMs will be limited – and volumes of SoCs needed constrained.
However, many companies are attempting to pare down the number of parameters needed for their LLM to not be dependent on connection back to the Cloud for execution. If connection to the Cloud becomes unnecessary, this becomes a massive market driver for future SoC devices with AI capabilities integrated into everything from cell phones to appliances.

We are already seeing Microsoft adding AI functionality to their Office applications for Desktop PCs and Laptops. Apple and Qualcomm and others are adding AI functionality to Smartphones. And many other companies have joined OpenAI in introducing multimodal LLMs that can carry out predetermined actions on the behalf of the user. In addition, these same companies are also introducing AI Agents into the market that can act as personal assistants in the areas of social secretary, accountant, security expert and many other functions that everyday people require in their daily lives.

There is certainly great interest in these types of products and services withing the business community and the general public. If these products can deliver tangible benefits to the end users, we anticipate an acceleration in the market revenues for AI-enabled devices, systems and software.
As depicted in Figure 30, an evolutionary trajectory exists, leading to increased complexity and functionality. The SHD Group anticipates that silicon solutions introduced to the market will generally align with this trend, adapting to evolving market demands for AI capabilities. We believe the industry is hovering between Levels 2 and 3 today and can project forward to AGI in the middle of the next decade.

Figure 30: AI Processing vs. Moore's Law 2014 - 2040



Source: The SHD Group, January 2025

With the introduction of ChatGPT tools by OpenAI in December 2022, AI started becoming a household word. It is estimated that 300M people are utilizing AI through ChatGPT and thus becoming familiar with AI-enabled tools and applications by leveraging the benefits of Generative AI and Large Language Models. This rapid rate at which AI functions like ChatGPT is being integrated into new applications is dramatic and changing the way we use computers as tools. Moreover, this is just the tip of the iceberg –driven by essentially one "product" - ChatGPT.

As more and more different types of AI products come to market, we can anticipate new AI-enabled applications from additional companies that could result in millions more customers using AI. When taken all together, this presents a great incentive for the semiconductor industry to respond with diverse solutions including Domain-Specific AI capabilities.

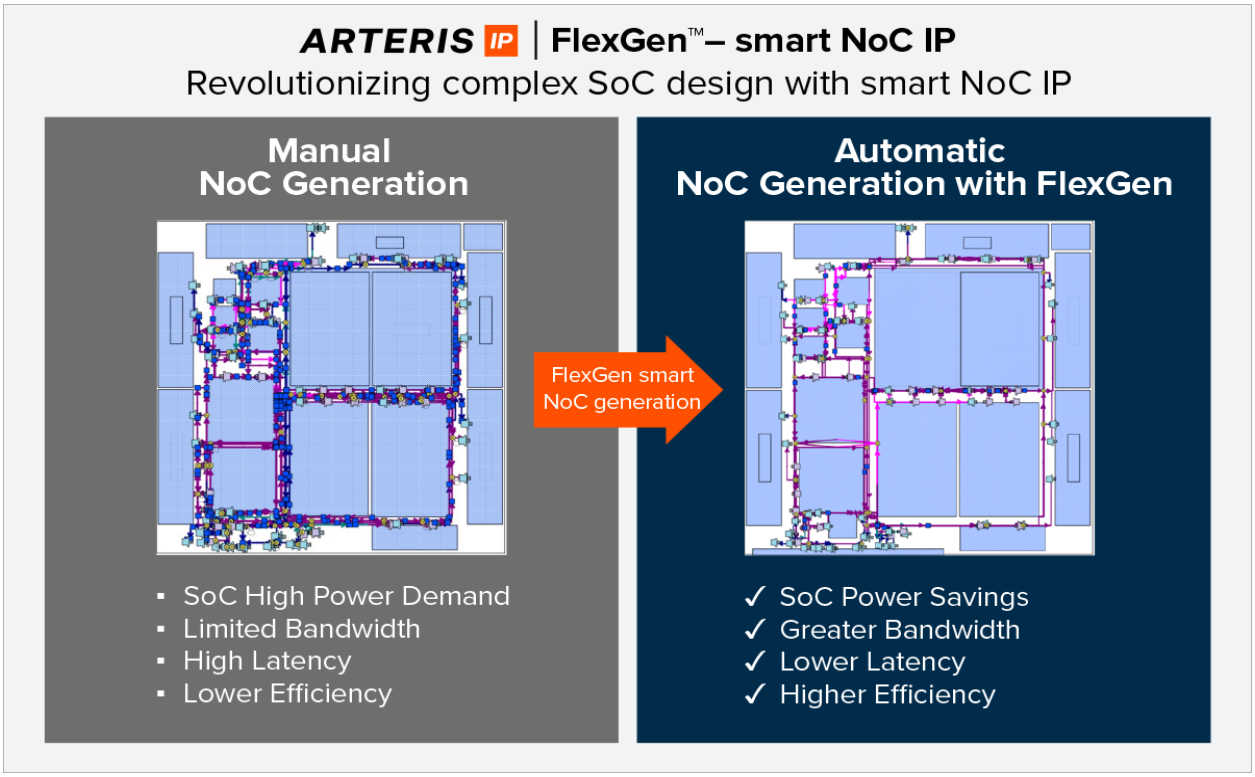
Apple and Qualcomm have already enabled their smartphones to run ChatGPT applications. Intel and AMD have introduced new CPUs for Desktop and Laptop PCs that incorporate NPUs to accelerate AI functions and Microsoft has added AI functionality, including the ability to run LLMs to their Windows OS and Office applications. At the time of this writing, several companies; Anthropic, DeepSeek, Google, Baidu, Mistral and others have joined Open AI by introducing their own LLMs that are available to the public and are generating great interest in the market.

This is the real reason we have commented on the AI market and the emergence of technologies such as LLMs in this report. These developments represent a great opportunity for AI-enabled designs and designers going forward in both market revenues and unit volumes. This is especially true because there are no real "legacy constraints" for AI-enabled applications; it is still an evolving field.

It is likely that new, high-performance silicon will be necessary to enable AI applications in PCs, laptops, tablets, and other computer systems, whether those are based on LLMs or another form of AI. In addition, it is also being deployed into the Industrial, Consumer, Automotive, Networking and Military markets, which are growth areas for SoCs of all types.

In keeping with the premises laid out in Figure 30 above, data appliances created specifically for the execution of AI-driven applications are starting to enter the market as more companies become familiar with the AI technologies and are directing their efforts into the Edge markets profiled in this report.

Figure 31: Arteris FlexGen – smart NoC IP



Source: Courtesy of Arteris IP, March 2025

Arteris FlexGen™ smart NoC IP automates NoC design. Using AI/ML-driven automation, FlexGen enables the efficient generation of optimized NoC designs with reduced manual effort, shorter iterations, and expert-level quality including the optimization of power, performance, and area (PPA). As SoCs scale to 500+ IP blocks and beyond, AI, HPC, and multi-die architectures push the limits of traditional interconnect design methodologies, making NoC development increasingly complex.

IX. Architectural Definition for a SoC

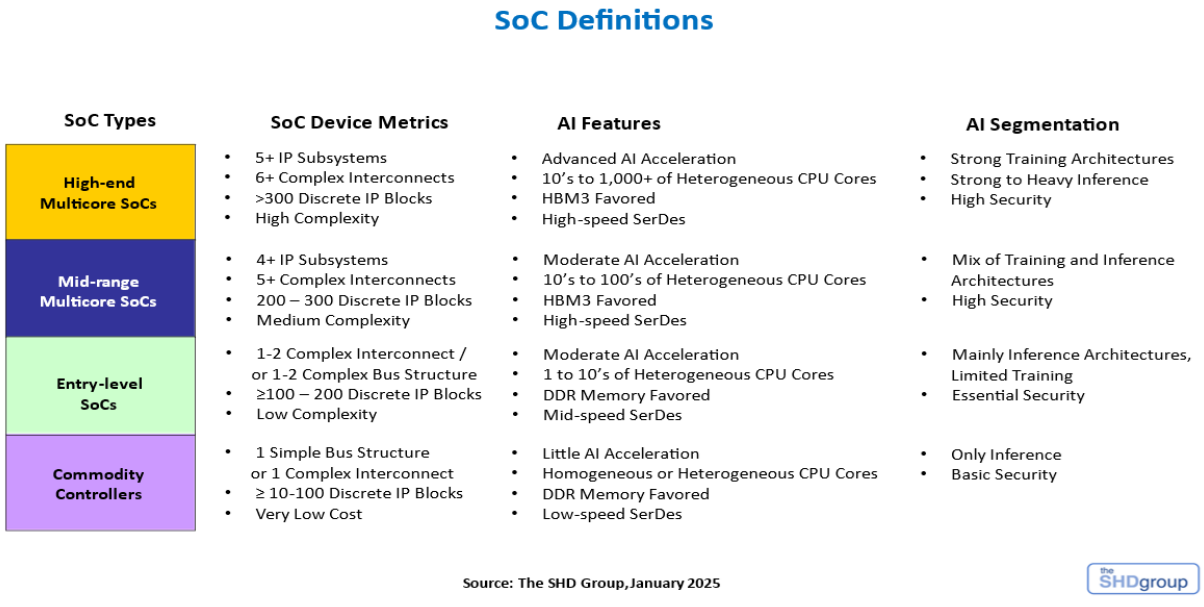
This study looks at specific types of SoC silicon solutions that are being used in the Edge-AI market today. These device types include:

- CPU – General Purpose
- GPUs
- NPUs
- DSPs
- FPGAs
- MCUs
- Custom ASIC / SoC

The SHD Group has compiled an architectural definition of the three SoC types we collect data for:

- High-end multicore SoC – silicon that functions with the most complexity, representing the highest cost range.
- Mid-range multicore SoC – silicon that represents the mid-range in device complexity and has a more moderate cost range.
- Entry-level SoC – silicon that operates with the least device complexity and occupies the lowest cost structure.

Figure 32: SoC Defined by IP Content



As this diagram in Figure 32 shows, we consider SoC architecture to be more about how the design is created than about the silicon itself. Contemporary SoCs today rely heavily on the use of 3rd party SIP and realistically cannot be crafted any other way. With this thought in mind, we created this diagram to give some granularity to these three silicon solutions to better understand the design and market dynamics that impact each type. We have extended these definitions to include device features and segmentation of the emerging AI market solutions since this new development in the semiconductor marketplace is going to be a main driver of applications, revenues, and unit volumes for many years to come.

We expect to amend this diagram as evolving market requirements provoke responses from silicon architects and designers alike.

X. Silicon Design Trends: Rising Design Complexity

In the last several years, contemporary SoCs have become very complex silicon solutions. They now consist of hundreds of millions or billions of gates, a hundred or more discrete semiconductor IP blocks, high-speed data channels, megabytes of volatile and non-volatile embedded memory, increasing amounts of analog/mixed-signal functionality, multiple CPU cores and multiple operating systems. In addition, robust, high-speed wireless connectivity can also be a prime requirement. All these features require millions of lines of application code written to provide the rich feature sets the market demands. In addition, AI functionality is finding its way into many SoCs in almost every application niche.

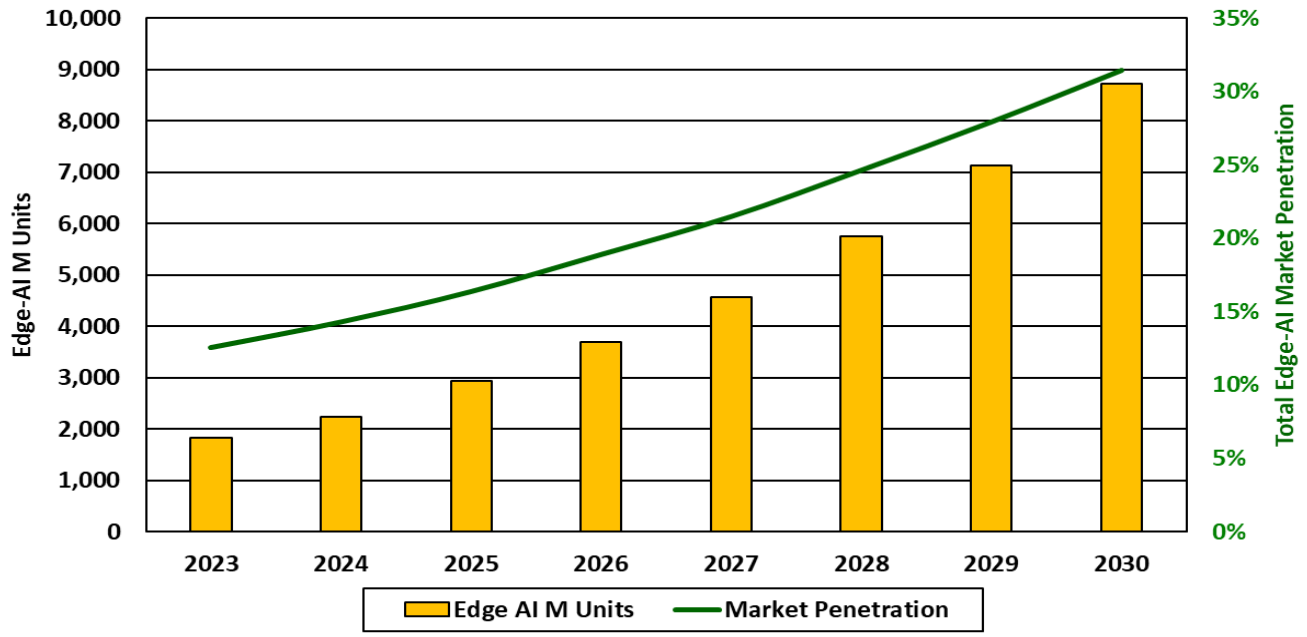
The following metrics show how design complexity has grown over time and how it is forecast to increase in support of rising performance requirements by most market applications.

The SHD Group tracks SoC design starts, and the average gate counts resident on each type of SoC. The following chart is derived from data we have collected through primary research over the course of many years.

Forecast

The widening array of AI applications is catalyzing global SIP and chip revenue and shipment growth as Figure 33 shows. The SHD Group is projecting that market penetration by Edge-AI devices will exceed 31% by 2030.

Figure 33: Edge-AI SoC Unit Shipments



Forecast Years 2025 - 2030

Source The SHD Group, February 2025

Please Note: A Spreadsheet Addendum for this report contains the fine granularity used to create the tables and charts in this report. The Addendum contains an additional 193 Tables and 187 charts that fully explore the Edge-AI market. This Addendum is available to Sponsors and Purchasers of the report.

Table 9: Device Complexity in K Gates 2023 - 2030

									CAGR %
K Gates	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**25 - 30**
High-end	1,227,526	1,652,490	2,199,856	2,974,333	4,032,966	5,338,623	6,974,571	9,198,229	26.9%
Mid-range	702,827	798,039	988,551	1,132,825	1,314,539	1,555,130	1,832,284	2,197,237	14.2%
Entry-level	15,814	19,401	24,055	30,337	38,965	50,023	64,191	82,334	22.8%
Avg. Gates	648,722	823,310	1,070,821	1,379,165	1,795,490	2,314,592	2,957,015	3,825,933	23.6%
Growth	30.2%	26.9%	30.1%	28.8%	30.2%	28.9%	27.8%	29.4%	

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

Table 9 shows the rise in device complexity as calculated in K Gates per device over the forecast period.

Figure 34: Rising Device Complexity

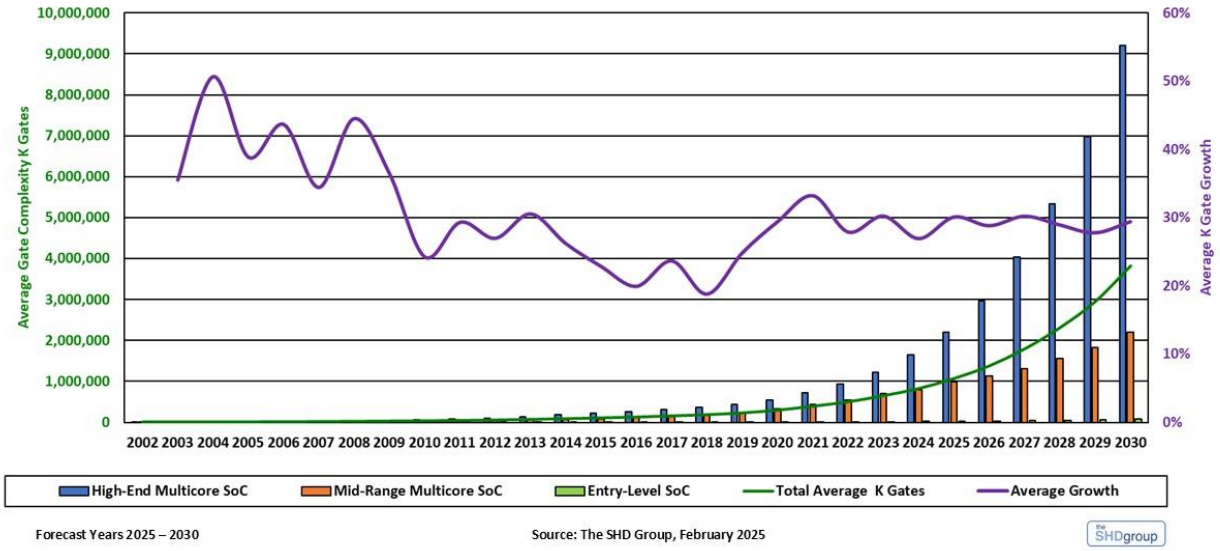


Figure 34 looks at the three types of SoC that The SHD Group tracks by gate count and calculates how the complexity levels have risen over time. The gate counts for each device type in 2002 were used as a baseline, and then each succeeding year was divided by that value to arrive at the growth rates shown above. As this chart shows, a representative design for a high-end multicore SoC in 2024 was 347X larger than its counterpart in 2002. The mid-range multicore SoC value of 366X is larger than the high-end multicore SoCs by 2020 because it started at a much lower gate count in 2002 and, therefore, has a higher growth rate over time. By 2030, advanced parts are forecast to be 1,933X larger than their 2002 counterparts, and value designs are expected to be 1,009X larger than in 2002. Entry-level SoC designs started in 2007 and are projected to be 56X larger by 2030 than in 2008. Total complexity for all SoC types in 2030 is projected to be 1,654X higher than it was in 2002.

Looking at the SoC design landscape in this way gives some perspective on how complex silicon has become today and what those trends will look like going into the future. It is safe to say that design costs will also continue to rise, driven by the device complexity levels shown above.

As device complexity levels and silicon and software design costs have risen, the demands put upon the Electronic Design Automation (EDA) industry to continuously improve their tools have been intense. Fortunately for SoC designers, the EDA industry has delivered higher performance tools to the semiconductor industry to aid in designing these very complex silicon solutions.

XI. Edge-AI Semiconductor Market Analysis

The market for Edge-AI devices is growing quickly, reflecting the increasing usefulness of AI functionality in edge applications. In our analysis, The SHD Group expects this trend to continue over the term of our forecast and even strengthen as device architectures are defined and AI algorithms are refined over time.

A gating factor to our forecast is the recent emergence of multimodal LLMs and their use in the creation of AI agents (Agentic AI). Given the premise that Agentic AI will enter the market and be useful to end users in all market segments, we believe this could provide a strong boost to the acceptance of Edge-AI and an increase in both units and revenues for the semiconductor market.

Total Edge-AI Revenues by Market Segment

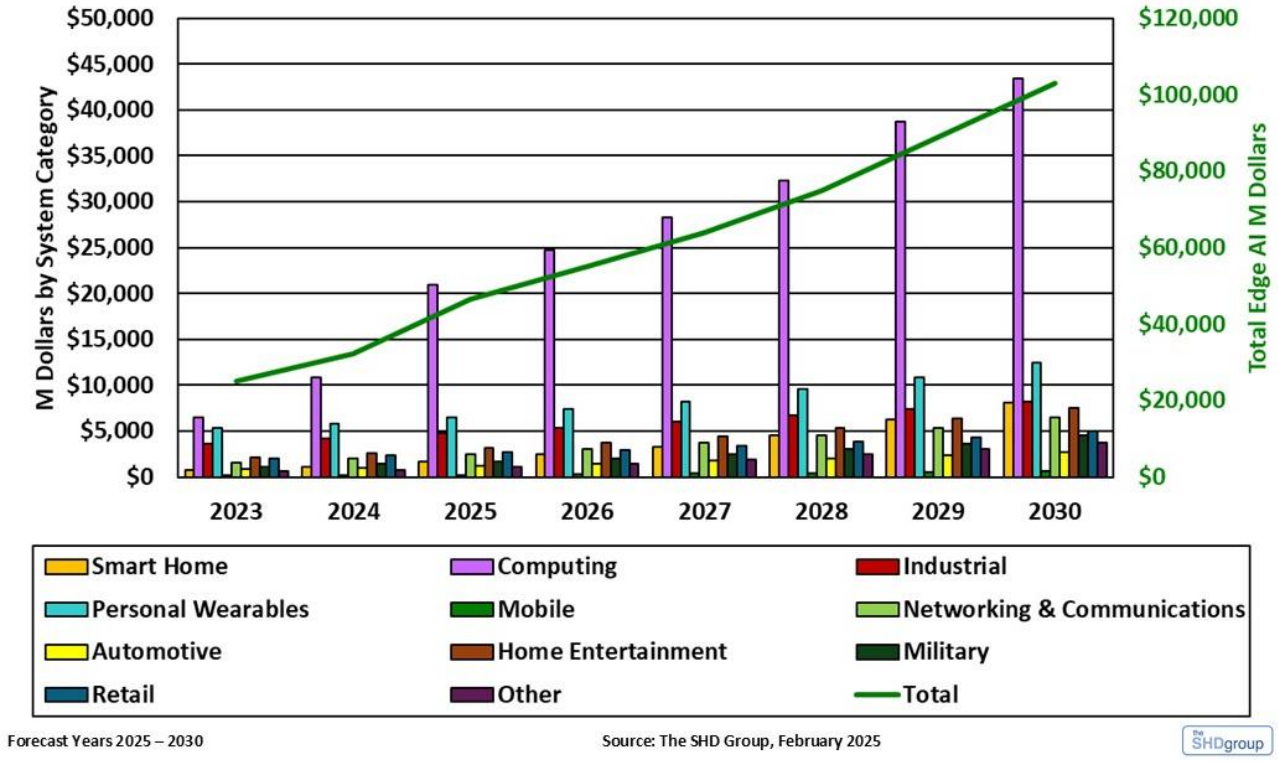
Table 10: Total M Dollars for Edge-AI Device Revenues by Market Segment

									CAGR %
M Dollars	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**24 - 30**
Smart Home	$802.2	$1,107.1	$1,679.8	$2,431.7	$3,320.0	$4,556.4	$6,257.6	$8,146.1	39.5%
Computing	$6,551.6	$10,840.7	$20,977.8	$24,781.3	$28,268.9	$32,335.4	$38,743.3	$43,408.7	26.0%
Industrial	$3,659.7	$4,192.0	$4,777.4	$5,372.2	$6,014.9	$6,731.4	$7,373.3	$8,172.2	11.8%
Personal Wearables	$5,365.4	$5,807.1	$6,503.9	$7,407.9	$8,251.1	$9,557.2	$10,832.7	$12,410.8	13.5%
Mobile	$164.8	$207.4	$241.0	$312.8	$395.3	$458.0	$575.0	$689.7	22.2%
Networking & Communications	$1,603.2	$2,008.2	$2,455.4	$3,075.0	$3,793.1	$4,585.6	$5,395.4	$6,460.4	21.5%
Automotive	$883.1	$1,039.0	$1,246.9	$1,479.9	$1,755.2	$2,047.7	$2,343.5	$2,697.5	17.2%
Home Entertainment	$2,188.6	$2,587.5	$3,123.3	$3,696.3	$4,427.0	$5,341.1	$6,337.5	$7,543.6	19.5%
Military	$1,122.4	$1,408.7	$1,711.9	$2,050.5	$2,504.0	$3,008.2	$3,678.4	$4,608.3	21.8%
Retail	$2,076.4	$2,367.1	$2,666.0	$2,972.8	$3,363.2	$3,862.9	$4,360.4	$4,966.7	13.1%
Other	$675.5	$812.7	$1,102.5	$1,505.9	$1,917.0	$2,447.3	$3,007.5	$3,786.1	29.2%
Total M Dollars	$25,092.9	$32,377.5	$46,485.8	$55,086.3	$64,009.7	$74,931.3	$88,904.6	$102,890.1	21.3%
Percent Growth		29.0%	43.6%	18.5%	16.2%	17.1%	18.6%	15.7%	
Market Penetration	18.9%	23.5%	32.7%	36.9%	40.7%	45.0%	50.5%	55.9%	

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

- In the Edge-AI market, the Computing category is forecast to reach the highest revenues out of all the device types analyzed by The SHD Group. In 2024, the device revenues are forecast to reach $10,840.7M dollars and are forecast to grow to $43,408.7M dollars by 2030. This represents a CAGR of 26.0%.
- In the Edge-AI market, the total device revenues were $32,377.5M dollars in 2024 and are forecast to grow to $102,890.1M dollars by 2030. This represents a CAGR of 21.3%.
- Total market penetration for Edge-AI device revenues was 23.5% in 2024 and are projected to reach 55.9% by 2030.

Figure 35: Total M Dollars for Edge-AI Device Revenues by Market Segment



Forecast Years 2025 – 2030

Source: The SHD Group, February 2025

Total Edge-AI Unit Shipments by Market Segment

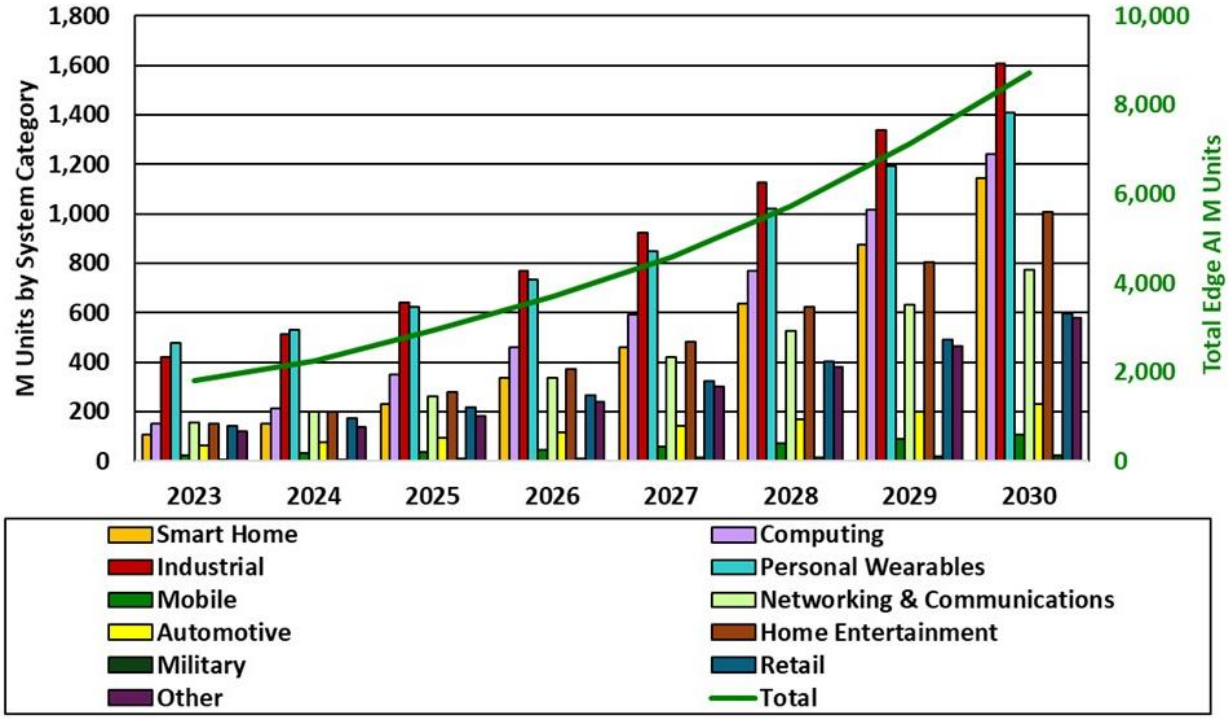
Table 11: Total M Unit Shipments for Edge-AI Devices by Market Segment

									CAGR %
M Units	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**24 - 30**
Smart Home	109.3	152.0	232.4	336.6	460.8	635.5	876.7	1,145.2	40.0%
Computing	151.8	212.3	349.7	460.2	591.8	767.6	1,015.9	1,242.9	34.3%
Industrial	420.6	513.5	643.1	770.8	923.3	1,127.3	1,337.1	1,607.2	20.9%
Personal Wearables	476.8	531.8	622.3	735.1	850.1	1,021.5	1,194.9	1,410.2	17.6%
Mobile	23.9	30.3	36.3	47.2	59.5	74.1	88.3	106.4	23.3%
Networking & Communications	155.5	202.0	261.2	337.1	422.1	524.6	634.4	775.3	25.1%
Automotive	63.9	77.1	95.5	117.4	142.1	169.8	197.8	232.2	20.2%
Home Entertainment	152.9	200.6	279.0	370.6	481.7	625.3	805.0	1,007.1	30.9%
Military	6.3	7.8	9.3	11.3	13.4	16.4	20.0	24.4	20.8%
Retail	140.8	174.2	215.7	265.5	324.3	401.6	492.0	599.3	22.9%
Other	118.6	139.2	181.9	241.4	302.5	379.7	464.4	579.9	26.8%
Total M Units	1,820.3	2,240.9	2,926.5	3,693.2	4,571.6	5,743.3	7,126.4	8,730.0	25.4%
Percent Growth		23.1%	30.6%	26.2%	23.8%	25.6%	24.1%	22.5%	
Market Penetration	12.7%	14.5%	17.2%	19.6%	22.2%	25.3%	28.4%	31.8%	

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

- In the Edge-AI market, the Industrial category is forecast to reach the highest unit shipments out of all the device types analyzed by The SHD Group. In 2024, the unit shipments were 513.5M units and are forecast to grow to 1,607.2M units by 2030. This represents a CAGR of 20.9%.
- In the Edge-AI Market, the total unit shipments were 2,240.9M units in 2024 and are forecast to grow to 8,7300.0M units by 2030. This represents a CAGR of 25.4%.
- Total market penetration for Edge-AI device revenues was 14.5% in 2024 and are projected to reach 31.8% by 2030.

Figure 36: Total M Unit Shipments for Edge-AI Devices by Market Segment



Forecast Years 2025 – 2030

Source: The SHD Group, February 2025

Total Edge-AI by Device Type

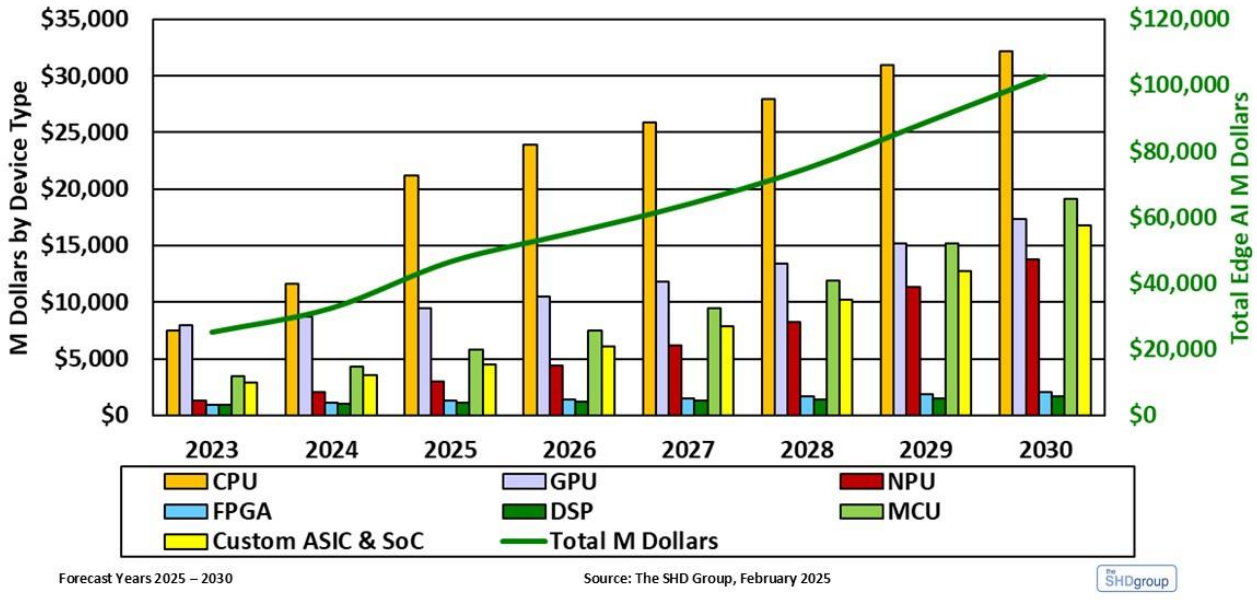
Table 12: Total M Dollars for Edge-AI by Device Type

									CAGR %
M Dollars	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**24 - 30**
CPU	$7,531.1	$11,602.6	$21,207.4	$23,959.6	$25,860.4	$27,954.0	$30,962.4	$32,140.0	18.5%
GPU	$7,973.1	$8,686.8	$9,504.6	$10,479.7	$11,817.3	$13,387.0	$15,167.8	$17,372.0	12.2%
NPU	$1,322.3	$2,020.9	$3,025.4	$4,387.8	$6,195.9	$8,286.1	$11,368.5	$13,814.4	37.8%
FPGA	$969.3	$1,118.3	$1,270.9	$1,395.2	$1,534.5	$1,674.7	$1,846.2	$2,023.2	10.4%
DSP	$917.6	$1,046.3	$1,153.7	$1,242.7	$1,335.1	$1,436.7	$1,538.1	$1,645.0	7.8%
MCU	$3,462.5	$4,318.5	$5,796.5	$7,495.2	$9,431.9	$11,928.8	$15,238.9	$19,134.8	28.2%
Custom ASIC & SoC	$2,917.0	$3,584.1	$4,527.4	$6,126.1	$7,834.6	$10,264.0	$12,782.7	$16,760.8	29.3%
Total M Dollars	$25,092.9	$32,377.5	$46,485.8	$55,086.3	$64,009.7	$74,931.3	$88,904.6	$102,890.1	21.3%
Percent Growth		29.0%	43.6%	18.5%	16.2%	17.1%	18.6%	15.7%	
Market Penetration	18.9%	23.5%	32.7%	36.9%	40.7%	45.0%	50.5%	55.9%	

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

- In the Edge-AI market, CPUs are forecast to reach the highest revenues out of all the device types analyzed by The SHD Group. In 2024, the device revenues are forecast to reach $11,602.6M dollars and are forecast to grow to $32,140.0M dollars by 2030. This represents a CAGR of 18.5%.
- In the Edge-AI market, the total device revenues were $32,377.5M dollars in 2024 and are forecast to grow to $102,890.1M dollars by 2030. This represents a CAGR of 21.3%.

Figure 37: Total M Dollars for Edge-AI by Device Type



Total Edge-AI Revenues by Region

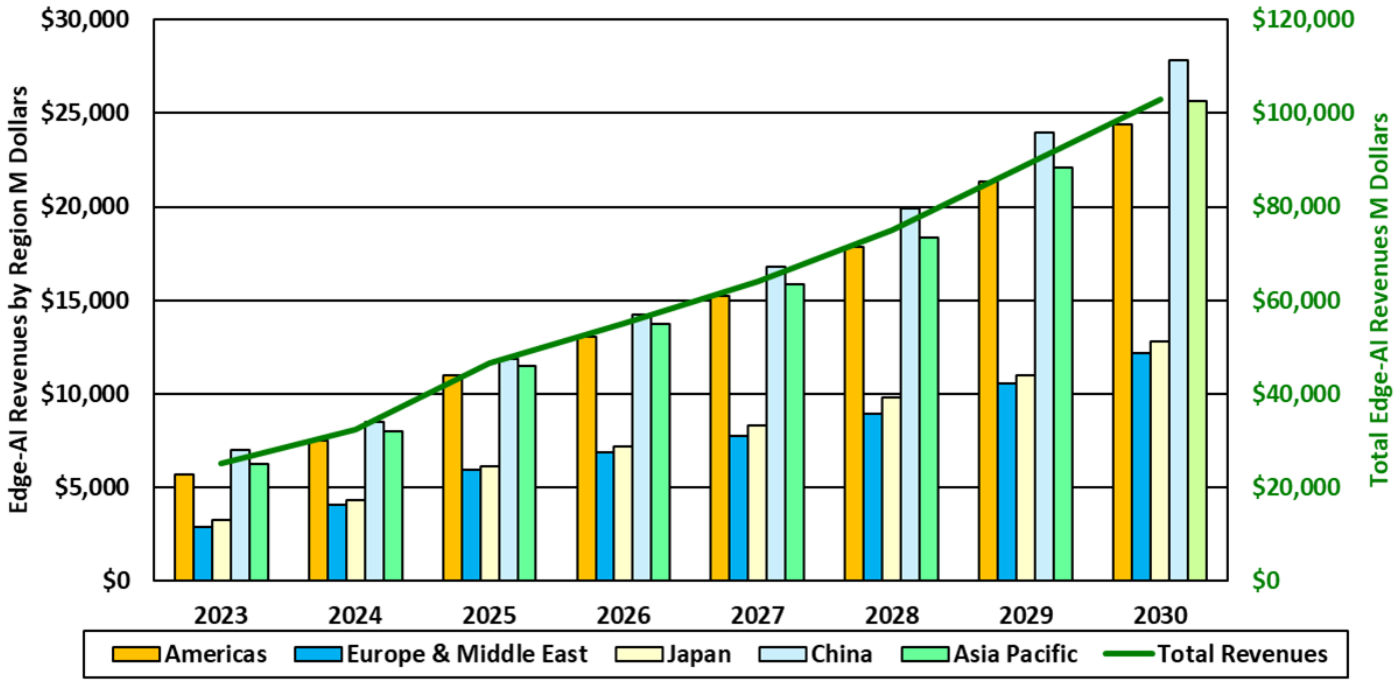
Table 13: Total Edge-AI Revenues by Region

									CAGR %
M Revenues	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**24 - 30**
Americas	$5,671.5	$7,533.7	$10,986.7	$13,061.9	$15,250.1	$17,884.1	$21,331.0	$24,387.4	21.6%
Europe & Middle East	$2,911.8	$4,054.7	$5,968.8	$6,902.5	$7,785.3	$8,926.8	$10,530.8	$12,196.1	20.1%
Japan	$3,238.6	$4,310.9	$6,116.2	$7,194.9	$8,313.2	$9,823.4	$10,976.7	$12,804.7	19.9%
China	$6,995.5	$8,497.5	$11,887.2	$14,215.0	$16,822.4	$19,927.3	$23,957.9	$27,843.7	21.9%
Asia Pacific	$6,276.5	$8,001.0	$11,526.9	$13,712.1	$15,838.3	$18,370.2	$22,109.5	$25,658.1	21.4%
Total Revenues	$25,093.8	$32,397.9	$46,485.8	$55,086.4	$64,009.3	$74,931.7	$88,905.8	$102,890.1	21.2%
Percent growth		29.1%	43.5%	18.5%	16.2%	17.1%	18.6%	15.7%	

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

- In the world Edge-AI regional market, China is forecast to reach the highest revenues out of the five regions analyzed by The SHD Group. In 2024, the regional revenues for China reached $8,497.5M dollars and are forecast to grow to $27,843.7M dollars by 2030. This represents a CAGR of 21.9%.
- In the total world regional Edge-AI market, the revenues were $32,397.9M dollars in 2024 and are forecast to grow to $102,890.1M dollars by 2030. This represents a CAGR of 21.2%.

Figure 38: Total Edge-AI Revenues by Region



Forecast Years 2025 - 2030 Source: The SHD Group, February 2025

Market Penetration for Edge-AI

Table 14: Edge-AI Market Penetration in All Segments by Revenue

Market Penetration	2023	2024	2025	2026	2027	2028	2029	2030
Smart Home	16.1%	19.3%	25.0%	29.5%	33.9%	39.8%	45.5%	51.5%
Retail	34.6%	35.0%	36.2%	37.0%	39.1%	41.9%	43.9%	46.9%
Personal Wearables	34.8%	37.9%	41.5%	45.5%	49.1%	53.3%	57.1%	62.0%
Mobile	10.9%	13.1%	15.2%	19.9%	24.1%	27.1%	33.0%	38.6%
Home Entertainment	10.9%	12.5%	14.3%	16.6%	19.3%	22.5%	26.0%	30.6%
Automotive	21.3%	24.0%	27.0%	31.2%	35.5%	39.7%	43.9%	48.9%
Industrial	27.8%	30.1%	32.6%	35.1%	37.6%	40.3%	42.7%	45.4%
Military	20.7%	22.6%	24.9%	27.0%	30.1%	32.8%	36.6%	41.6%
Computer	13.7%	23.0%	46.4%	53.2%	58.6%	64.7%	74.3%	82.7%
Networking & Communications	13.2%	14.8%	16.7%	19.8%	23.3%	26.9%	30.3%	34.5%
Other	26.8%	30.8%	32.4%	36.5%	41.2%	47.2%	52.6%	59.6%
Total Market Penetration	18.9%	23.5%	32.7%	36.9%	40.7%	45.0%	50.5%	55.9%

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

Market penetration is calculated by comparing the device revenues of non-AI SoCs to SoCs that have some level of AI functionality.

- In the Edge-AI market, the Computer segment is forecast to reach the highest market penetration for revenues out of all the segments analyzed by The SHD Group. In 2024, the market penetration was 23.0% and is projected to grow to 82.7% by 2030.
- In the Edge-AI market, the Personal Wearables segment is forecast to reach the second highest market penetration for revenues out of all the segments analyzed by The SHD Group. In 2024, the market penetration was 37.9% and is projected to grow to 62.0% by 2030.
- In the Edge-AI market, the Smart Home segment is forecast to reach the third highest market penetration for revenues out of all the segments analyzed by The SHD Group. In 2024, the market penetration was 19.3% and is projected to grow to 51.5% by 2030.
- Overall, the Edge-AI market is forecast to reach a market penetration for revenues in 2024, of 23.5% and is projected to grow to 55.9% by 2030.

Figure 39: Edge-AI Market Penetration in All Segments by Revenue

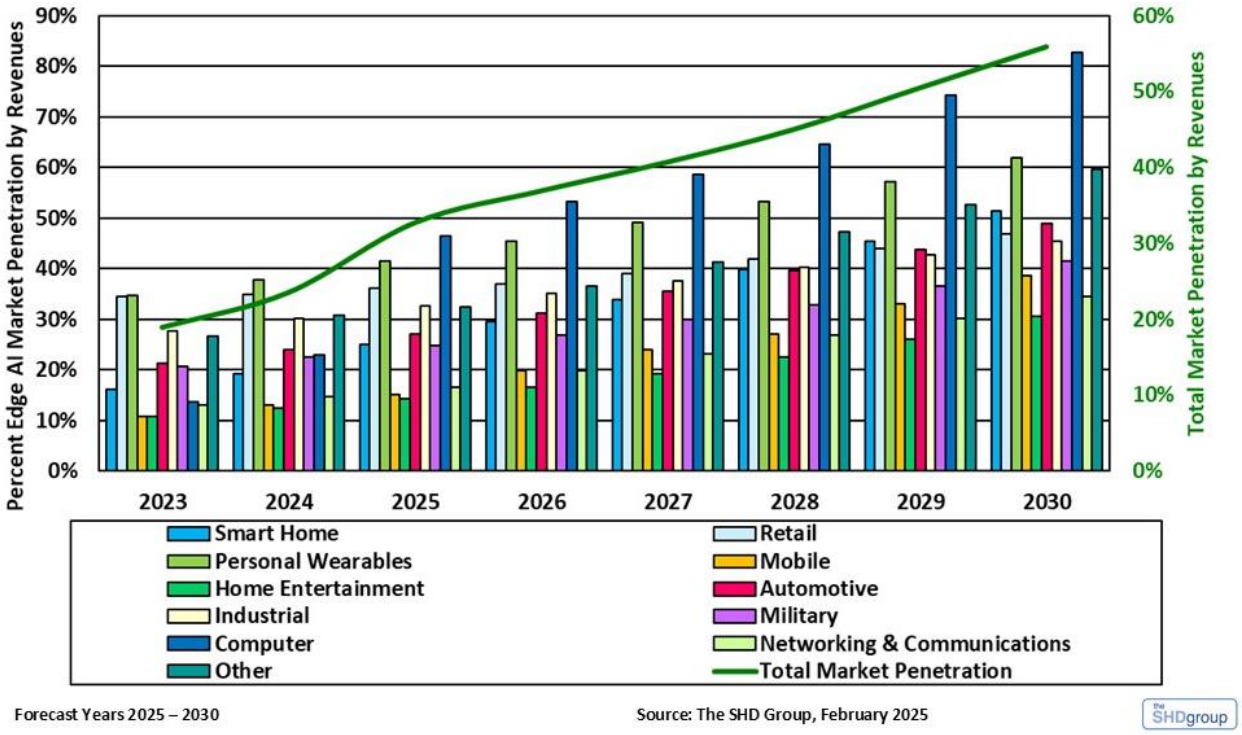


Table 15: Edge-AI Market Penetration in All Segments by Units

Market Penetration	2023	2024	2025	2026	2027	2028	2029	2030
Smart Home	11.6%	13.9%	18.1%	21.3%	24.5%	28.7%	32.7%	36.9%
Retail	12.9%	13.9%	15.3%	16.9%	18.6%	20.7%	22.8%	25.1%
Personal Wearables	18.3%	20.3%	23.0%	25.9%	28.5%	31.8%	34.8%	38.4%
Mobile	8.1%	9.8%	12.2%	15.4%	18.5%	22.2%	25.6%	29.9%
Home Entertainment	8.9%	10.6%	13.1%	15.6%	18.2%	21.4%	24.8%	28.6%
Automotive	11.6%	13.2%	15.3%	17.9%	20.5%	23.3%	25.8%	28.9%
Industrial	15.1%	16.6%	18.9%	20.9%	23.0%	25.5%	28.1%	30.8%
Military	18.9%	19.9%	22.2%	24.0%	26.1%	28.2%	31.6%	35.0%
Computer	10.9%	14.6%	18.4%	21.5%	24.9%	29.6%	34.7%	40.3%
Networking	8.7%	10.0%	11.9%	14.3%	16.6%	19.3%	21.7%	24.7%
Other	10.8%	12.3%	13.7%	15.7%	17.9%	20.5%	23.0%	26.2%
Total Market Penetration	12.7%	14.5%	17.2%	19.6%	22.2%	25.3%	28.4%	31.8%

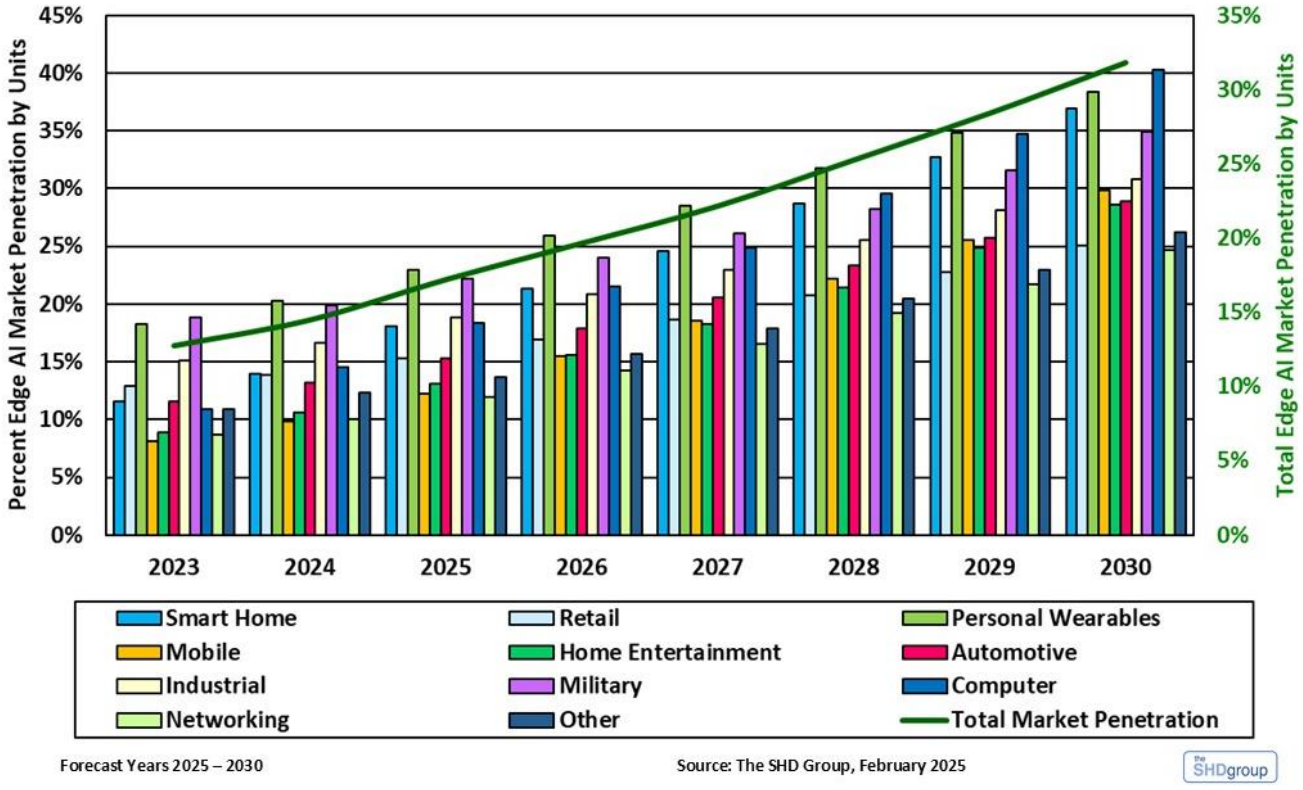
Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

Market penetration is calculated by comparing the device Unit shipments of non-AI SoCs to SoCs that have some level of AI functionality.

- In the Edge-AI market, the Computer segment is forecast to reach the highest market penetration for unit shipments out of all the segments analyzed by The SHD Group. In 2024, the market penetration was 14.6% and is projected to grow to 40.3% by 2030.

- In the Edge-AI market, the Personal Wearables segment is forecast to reach the second highest market penetration for unit shipments out of all the segments analyzed by The SHD Group. In 2024, the market penetration was 20.3% and is projected to grow to 38.4% by 2030.
- In the Edge-AI market, the Smart Home segment is forecast to reach the third highest market penetration for unit shipments out of all the segments analyzed by The SHD Group. In 2024, the market penetration was 13.9% and is projected to grow to 36.9% by 2030.
- Overall, the Edge-AI market is forecast to reach a market penetration for unit shipments in 2024, of 14.5% and is projected to grow to 31.8% by 2030.

Figure 40: Edge-AI Market Penetration in All Segments by Units



XII. 3RD Party Semiconductor Intellectual Property Market Analysis

The 3rd Party Semiconductor Intellectual property (SIP) market is one of the key components of the semiconductor market today with every SoC using multiple SIP blocks of one type or another to accomplish and enhance the final silicon solution. The reality today is that no SoC of any complexity level can be designed without the use of 3rd Party SIP, and also that the number of blocks and the complexity levels of those blocks in these designs is growing.

Table 16 shows the revenues for the Total Worldwide SIP market by revenue category.

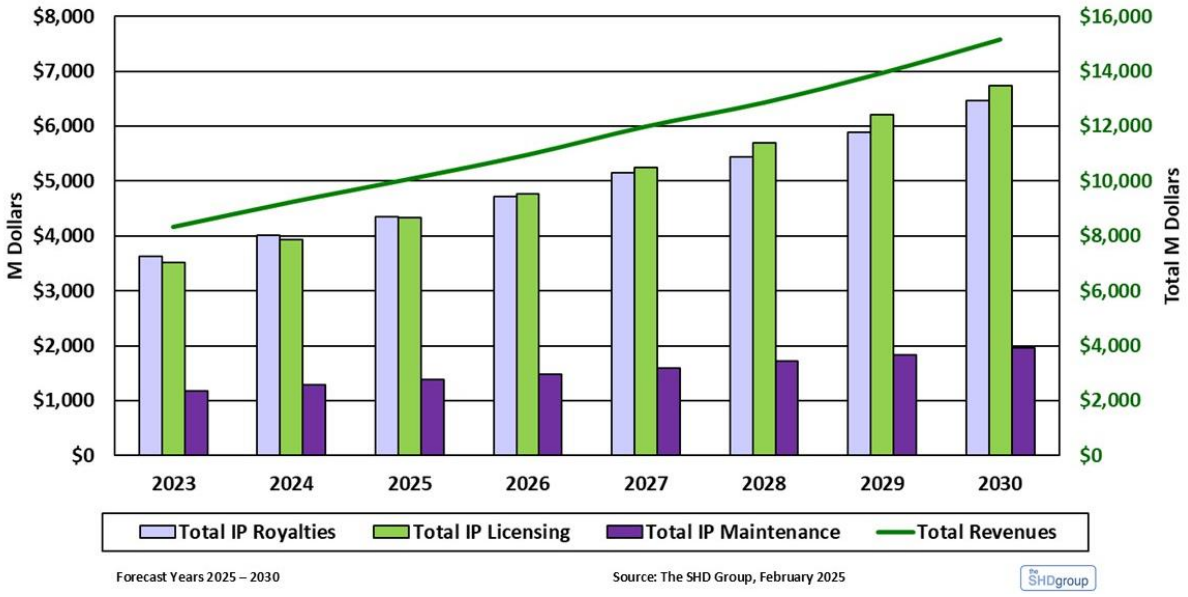
Table 16: Total M Dollars for the Worldwide SIP Market by Revenue Category

									CAGR %
M Dollars	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**24 - 30**
Total SIP Royalties	$3,624.4	$4,010.8	$4,353.1	$4,718.5	$5,149.3	$5,441.4	$5,894.0	$6,463.7	8.3%
Total SIP Licensing	$3,519.8	$3,938.1	$4,340.8	$4,761.6	$5,243.0	$5,696.4	$6,210.1	$6,741.6	9.4%
Total SIP Maintenance	$1,178.8	$1,284.1	$1,379.3	$1,479.6	$1,597.5	$1,727.3	$1,841.6	$1,964.3	7.3%
Total Revenues	$8,323.0	$9,233.0	$10,073.2	$10,959.6	$11,989.9	$12,865.1	$13,945.8	$15,169.6	8.6%
Percent Growth		10.9%	9.1%	8.8%	9.4%	7.3%	8.4%	8.8%	

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

- 2024 saw the total worldwide IP market continue its growth reaching $9,233.0B dollars and is forecast to reach $15,169.6B by 2030, a CAGR of 8.6%.
- The Royalties associated with 3rd Party SIP reached $4,010.8B in 2024 and is forecast to climb to $6,463.7B by 2030, a CAGR of 8.3%.
- The Licensing revenues in 2024 were $3,938.1B and are forecast to reach $6,741.6B by 2030, a CAGR of 9.4%.
    - Licensing revenues exceed Royalty revenues in 2026 driven in part by the new designs and SIP needed for AI-enabled SoC s of all types.
- SIP Services and maintenance revenues were $1,284.1B in 2024 and are projected to reach $1,964.3B by 2030, a CAGR of 7.3%.

Figure 41: Total M Dollars for Worldwide SIP Market by Revenue Category



SoCs created to function in the Edge-AI market necessarily use multiple SIP blocks just as their non-AI counterparts do in the broader semiconductor market. Most SIP blocks can be considered to be dual-use in nature – meaning that they will function equally well in both types of SoC; non-AI and AI-enabled SoCs.

However, some blocks are now being created with AI functionality that are intended to be used only in designs that will be AI-enabled SoC designs. This is being done to aid designers in achieving the right mix of AI functions in their final silicon solutions. The following section looks at the range of SIP blocks The SHD Group tracks and analyzes from the point of view of incorporating some level of AI functionality into them.

The intent here is to examine the specific SIP blocks that have incorporated AI functionality. These blocks generally either aid in the performance of inference or are performing the inference themselves. Other types of blocks are present in the design, but do not themselves contain any AI functionality.

In addition, some system functions have an AI component (such as some discrete DRAM memories) but are not SIP blocks themselves. Given how much interest and urgency there is in creating AI silicon that is high-performing, cost-effective and has a tolerable power budget, we believe that eventually what are discrete system-level parts today could become SIP blocks that do incorporate some level of AI functionality tomorrow.

Embedded Memory AI-SIP Blocks

AI inference operations Edge devices require processing very large amounts of data and creating weights that determine how important each data point is. This requires moving the weights between the CPU and the memory where the weights are stored. This process can use a great deal of power and takes time – which can create latency, delaying the final answer to be acted upon.

With this process in mind, some embedded memory SIP vendors have added AI functionality into their SIP blocks to cut down on some of the data movement, power consumption and latency inherent in inference operations. This has taken the form of a 'Processing-in-Memory approach and is used to reduce the need to transfer data between memory and CPUs / GPUs, significantly improving AI performance.

- Embedded Memory SIP vendors Etron Technology, Semidynamics and others offer SIP memory blocks that include AI functionality.

These products range from the Processor-in-Memory approach mentioned above to using a RISC-V CPU IP block or to include Machine Learning (ML) algorithm specific features to AI algorithms to enhance memory bandwidth.

The SHD Group believes research in this area will continue and more products will be introduced in the embedded memory SIP market as AI silicon designers look for more ways to increase bandwidth and reduce power consumption and area in AI-enabled SoCs.

CPU AI-SIP Blocks

Many of the discrete CPU vendors and many of the CPU SIP vendors have incorporated AI functionality into their products to make their products higher-performing and more efficient when handling AI applications. These product offerings extend across all applications and markets as Edge-AI continues to grow into these application segments.

- Discrete CPU vendors like AMD and Intel have added AI functionality to their products.
- In the SIP market, ARM is the largest CPU SIP vendor and has an extensive product portfolio of AI SIP blocks.
- In addition, many of the RISC-V CPU SIP vendors have AI-enabled products. Andes Technology, SiFive, Nuclei Technology, MIPS Technologies, Codasip, Synopsys and many others have entered this market.

In our forecast, the AI-CPU SIP market obtains the highest revenues over the forecast period.

DSP AI-SIP Blocks

DSP AI SIP blocks provide flexible instruction sets optimized for AI/ML operations, enabling seamless integration of AI and DSP tasks within a single application. They are used in both audio and video applications that are found in automotive and industrial systems. In automotive applications, the DSP AI SIP blocks are architectured to combine vision, radar, and AI processing within a single architecture.

- Ceva is the largest DSP AI SIP vendor with an extensive list of products.
    - In addition, Cadence Design Systems-Tensilica, Synopsys, IP Cores, Inc., and DSPIA, Inc., among others have entered this market with AI-enabled DSP SIP blocks.

NPU AI-SIP Blocks

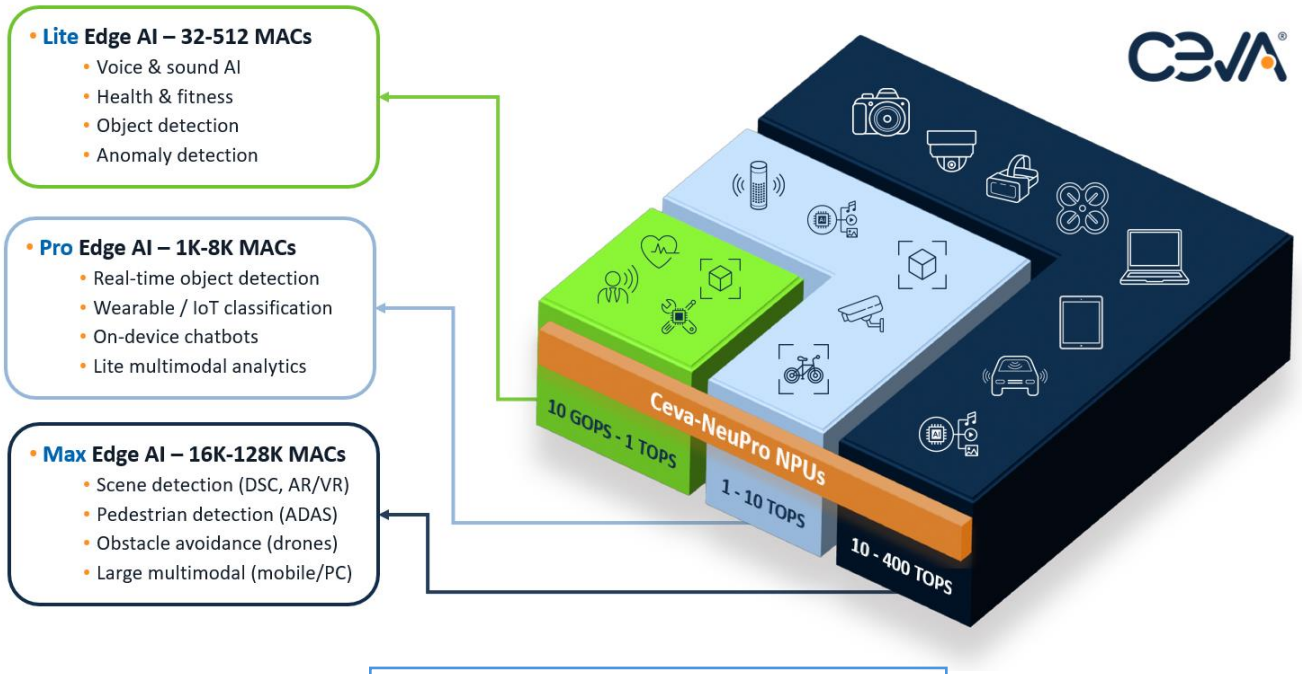
Several of the current CPU AI SIP vendors have also created NPU AI SIP blocks, joining other SIP vendors who are concentrating only on NPU AI SIP. This is due to the great demand for AI inference in Edge-AI devices many different silicon architectures have emerged and aimed at this market with many more planned in the near future. NPUs are valued for their power and area efficiency and are heavily used in SoCs of all types.

Some vendors have combined NPU, DSP, and real-time CPU functionalities into a single programmable core. This unified architecture simplifies system-on-chip (SoC) hardware design and software programming, efficiently handling AI and machine learning inference tasks.

Several SIP vendors have introduced NPU AI-SIP in the last few years.

- ARM, Ceva, Quadric, Synopsys, Tenstorrent and VeriSilicon are SIP vendors in this market as it is growing in importance.
- The SHD Group feels other SIP vendors will introduce NPU AI SIP products in the near future.

Figure 42: Ceva-NeuPro NPU



Source: Courtesy of Ceva, Inc., March 2025

Ceva-NeuPro-Nano (as shown in Figure 42) is a highly efficient and self-sufficient Edge NPU designed for Embedded ML applications. With more than 4 billion inference chips for Embedded ML (TinyML) devices forecasted to ship annually by 2029, this Edge NPU SIP is the smallest of Ceva's NeuPro NPU product family. It delivers the optimal balance of ultra-low power and high performance in a small area to efficiently execute Embedded ML workloads across AIoT product categories, including hearables, wearables, home audio, smart home, smart factory, and more. Ranging from 1 GOPS up to 200 TOPS per core, Ceva-NeuPro-Nano is designed to enable energy-efficient, always-on audio, voice, vision, and sensing use cases in battery-operated devices across a wide array of end markets.

Ceva-NeuPro-Nano is a stand-alone NPU, not an AI accelerator, and therefore does not require a host CPU/DSP to operate. The IP core includes all the processing elements of a standalone NPU, including code execution and memory management. The Ceva-NeuPro-Nano architecture is fully programmable and efficiently executes neural networks, feature extraction, control code and DSP code. It also supports the most advanced machine-learning data types and operators including native transformer computation, sparsity acceleration, and fast quantization to efficiently execute a wide range of neural networks, delivering a highly optimized solution with excellent performance.

Ceva-NeuPro-M is a scalable NPU architecture, ideal for transformers and generative AI applications, with an exceptional power efficiency of up to 3500 Tokens-per-Second/Watt for a Llama 2 and 3.2 models. With 30% of generative AI inference predicted to be on-device in the next two years, the Ceva-NeuPro-M NPU IP delivers exceptional energy efficiency tailored for edge computing while offering scalable performance to handle AI models with over a billion parameters. Its award-winning architecture introduces significant advancements in power efficiency and area optimization, enabling it to support massive machine-learning networks, advanced language and vision models, and multi-modal generative AI.

Even mid-range AI workloads such as computer vision (object detection and classification), speech recognition, and small-scale NLP (keyword spotting), are becoming dominated by the use of transformers (e.g., ViT, BERT).

Transformers support in edge NPUs is becoming mandatory for local text generation, context-aware AI assistants, and multimodal models for AR/VR, robotics, and advanced user interfaces applications.

With a processing range of 10 to 400 GOPs per core, leading area efficiency, support for advanced transformers, sparsity, and compression, the Ceva-NeuPro-M optimizes key AI models seamlessly, and with its highly scalable design, it provides an ideal SIP solution for embedding high performance AI processing in SoCs across a wide range of Edge-AI applications.

GPU AI-SIP Blocks

GPU AI-SIP is also used in Edge-AI silicon solutions as an alternative to performing these functions in the CPU or the NPU. GPU SIP blocks integrate tensor cores or AI acceleration units to handle tasks such as image classification, speech recognition, and deep learning inference and offloads intensive workloads from the main CPU. This type of SIP can be power-hungry and can be limited by how much embedded memory is available in Edge-AI devices.

- There are many companies in this market today such as ARM, Imagination Technologies, Silicon Arts, VeriSilicon and many others.

Analog AI-SIP Blocks

In general Analog SIP blocks today do not include any AI functionality. There are certain Analog SIP functions that could possibly benefit from performing inference functions on the data they are seeing. However, this would require, at the very least, a controller of some sort and a substantial amount of embedded memory - > 1MB, to be incorporated into the Analog SIP block. This approach is not favored by SoC designers today wanting to keep their designs as simple as possible and also wanting to keep the inference functions solidly in the digital domain. This may change in the future as demand for more performance from every part of the system continues to increase and designers look to maximize functionality.

Interface AI-SIP Blocks

At the time of this writing, Interface SIP blocks do not contain any AI functionality of their own, relying instead on other circuits in the system to perform these functions. However, it is our belief that given the sheer amount of data being moved and the very large number of calculations being performed, designers may look to add AI (or Machine Learning) functionality to Interface SIP blocks in the future.

This could be especially beneficial for MCUs, or other types of parts, where they are required to support a very wide range of interface protocols. The issue for parts of this type is that they are usually pin-limited in most of the systems where they reside. This could be due to it being a legacy application where the customers of these systems cannot tolerate higher costs for the higher-performing solutions.

The new interface AI SIP blocks would look to allow the use of multiple communications protocols over the same device pinout and provide greater functionality with little to no changes to the device footprint.

The SHD Group expects products like this to enter the market starting in the future.

Interconnect AI-SIP Blocks

Currently, Interconnect SIP blocks do not incorporate any AI functionality. They are used extensively on all types of AI and Non-AI SoCs and are an essential SIP block in most designs. There is the possibility of incorporating ML capabilities into Interconnect SIP for the purpose of fine-tuning the network performance. However, this has not happened yet. The SHD Group believes this may happen in the out years of our AI-SIP forecast as SoC designers seek to get the most performance out of their silicon solutions.

Security AI-SIP Blocks

In the last several years the semiconductor market has started placing greater emphasis on security functionality as more devices – especially at the edge, have been connected to the Internet and to each other. With the rise

of AI and more SIP vendors entering the Security SIP market, AI functionality is starting to be added to security SIP blocks.

Many of these capabilities have existed as software that runs on the system processor, but with the rise of AI, now the trend is to incorporate these capabilities, in the form of ML, directly into the SIP blocks themselves to reduce latency and improve responsiveness.

This addition of AI functionality focuses on enhancing the threat detection and response capabilities of Security SIP blocks. The idea is to identify and mitigate sophisticated cyber threats more effectively. This trend is now extending into using quantum -resistant algorithms to protect systems from future quantum attacks.

Companies such as Rambus, Arista, Synopsys, Cadence (having acquired Secure IC in February 2025) and others are engaged in the Security market with multiple products.

eFPGA AI-SIP Blocks
eFPGA SIP blocks have emerged as an important and useful tool for SoC designers interested in adding a programmable logic capability to their designs. This has now extended to eFPGA SIP vendors adding AI capabilities to their products. This has taken the form of an AI / ML compute engine that can reuse stored of cached weights and data or tensor processing blocks to create a reconfigurable AI and DSP accelerator SIP block Achronix , Flex Logix (acquired by Analog Devices in November 2024), Menta and Quick Logic are companies that have added AI functionality to their eFPGA products.

Embedded Analytics AI-SIP Blocks
With the rising complexity of SoC designs today and the demand for more performance, a new type of SIP has entered the market that gives SoC designers the capability of embedding analytic functions directly into their designs. These analytic functions give designers insights into, and control over, different aspects of the design such as monitoring and controlling the CPU – Memory interactions, the numerous power islands and power rails and various security functions, are a few areas where these blocks come into play. These SIP blocks collect and then transmit data collected during the operation of the part back to designers, giving them better information on device operations and the ability to change device operating parameters if necessary.

Companies have started to add AI functionality to their products to provide better insights into device operations with ML functions. This data is then fed into a ML Analytics platform to detect anomalies, predict potential failures and optimize performance.
protean Tecs, Siemens EDA and Synopsys have embedded analytics products in this market.

Logic AI-SIP Blocks
Logic SIP blocks are basic building blocks widely used in the design of Custom ASICs / SoCs. These blocks by themselves are not offered with any AI functionality built into them. Instead, SoC designers combine these blocks to create the AI functionality they need for their silicon solutions.
Because of their very basic nature, The SHD Group does not anticipate any Logic SIP blocks to be offered to the market in the future.

Audio AI-SIP Blocks
Audio SIP was one of the first areas to integrate AI functionality for functions like key-word spotting, wake word detection, speech recognition, voice-controlled user interfaces, language translation, noise cancellation and sound identification. Given the propensity of Edge-AI devices to be relatively small, a voice interface used for control allows the elimination of many manual controls, helping to cost reduce such devices and improve reliability with the removal of electromechanical controls.

Table 17: Total Worldwide AI-SIP Revenue Forecast

									CAGR %
M Dollars	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**24 - 30**
Memory	$0.0	$13.8	$19.8	$29.0	$45.6	$60.3	$85.4	$104.9	40.3%
CPU	$255.4	$295.0	$413.7	$505.8	$598.8	$693.6	$814.5	$980.0	22.2%
DSP	$37.6	$44.1	$56.0	$63.5	$76.6	$93.2	$108.9	$131.2	19.9%
NPU	$80.5	$109.3	$141.7	$178.8	$223.5	$282.0	$374.9	$503.0	29.0%
GPU	$111.4	$126.0	$141.7	$161.2	$184.1	$221.1	$255.9	$310.2	16.2%
eFPGA	$2.5	$6.4	$12.2	$13.2	$18.8	$28.0	$44.2	$68.6	48.3%
Audio	$30.6	$38.6	$45.4	$58.4	$75.1	$91.4	$104.8	$121.4	21.1%
Total Revenues	$517.9	$633.2	$830.5	$1,009.8	$1,222.5	$1,469.6	$1,788.7	$2,219.2	23.2%
Percent Growth		22.3%	31.2%	21.6%	21.1%	20.2%	21.7%	24.1%	

Forecast Years 2025 - 2030 **Source: The SHD Group, March 2025**

- In the Worldwide market for AI-SIP, total revenues for CPU AI-SIP are forecast to grow quickly as the Edge-AI continues to grow in volume. In 2024, the total AI-SIP revenues reached $295.0M dollars and are forecast to grow to $980.0M dollars by 2030. This represents a CAGR of 22.2%.
- In the Worldwide AI-SIP market for AI-SIP, Total AI-SIP revenues are forecast to grow quickly as the Edge-AI continues to grow in volume. In 2024, the total AI-SIP revenues reached $633.2M dollars and are forecast to grow to $2,219.2M dollars by 2030. This represents a CAGR of 23.2%.

Figure 43: Total Worldwide AI-SIP Market Forecast

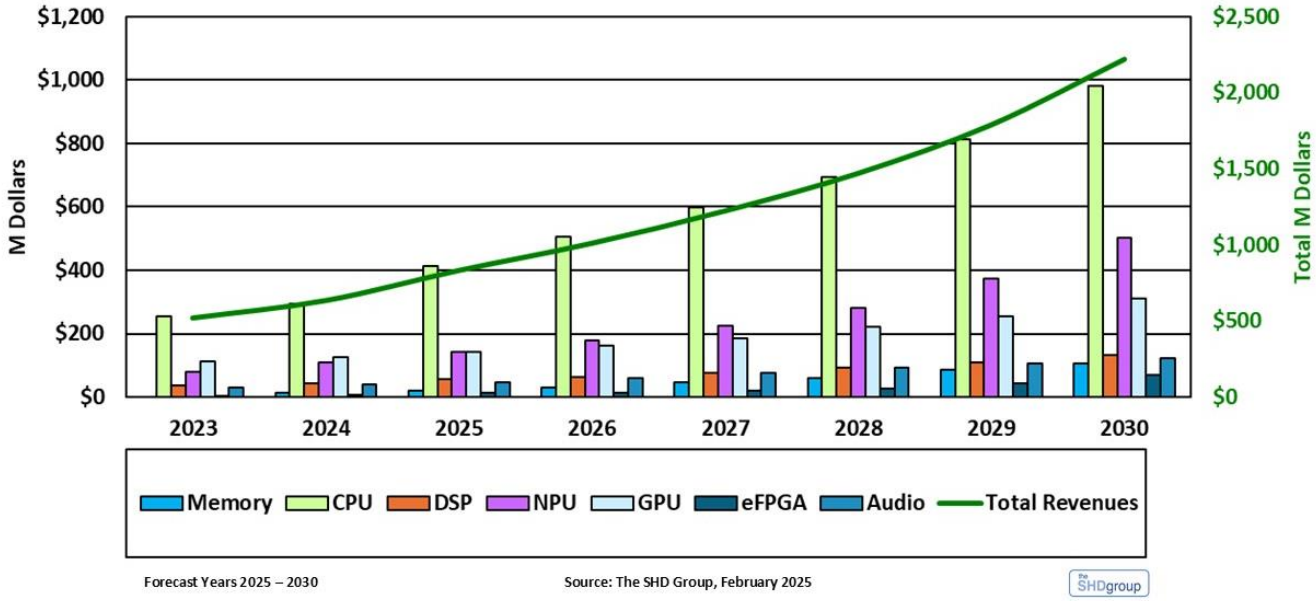


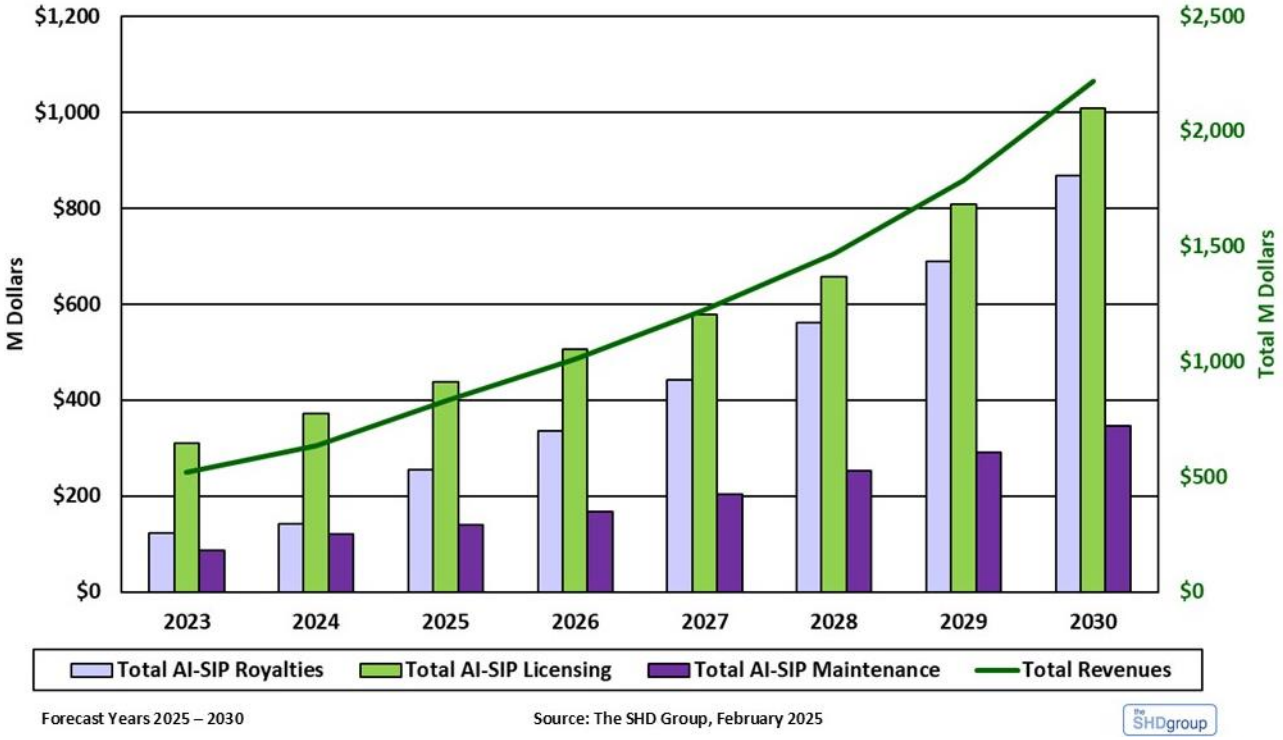
Table 18: Total Worldwide AI-SIP Market Forecast by Revenue Category

									CAGR %
M Dollars	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**24 - 30**
Total AI-SIP Royalties	$122.0	$142.2	$254.2	$335.9	$441.5	$560.8	$688.8	$866.6	35.1%
Total AI-SIP Licensing	$309.9	$371.3	$437.2	$506.4	$577.4	$656.6	$808.7	$1,007.1	18.1%
Total AI-SIP Maintenance	$86.1	$119.7	$139.1	$167.5	$203.6	$252.2	$291.2	$345.5	19.3%
Total Revenues	$517.9	$633.2	$830.5	$1,009.8	$1,222.5	$1,469.6	$1,788.7	$2,219.2	23.2%
Percent Growth		22.3%	31.2%	21.6%	21.1%	20.2%	21.7%	24.1%	

Forecast Years 2025 - 2030 **Source: The SHD Group, March 2025**

- In the Worldwide 3rd Party SIP market, Total SIP Licensing revenues are forecast to continue the growth of the last few years. In 2024, the Licensing SIP revenues reached $413.7M dollars and are forecast to grow to $1202.0M dollars by 2030. This represents a CAGR of 19.5%.
- In the Worldwide AI SIP market, Total AI SIP revenues for AI SIP are forecast to grow quickly as the Edge-AI continues to grow in volume. In 2024, the total AI SIP revenues reached $712.5M dollars and are forecast to grow to $2764.7M dollars by 2030. This represents a CAGR of 25.4%.

Figure 44: Total worldwide AI-SIP Market Forecast by Revenue Category



As the use of AI SIP in SoC designs targeted at Edge-AI systems, it will grow to be a larger and larger part of the Total 3rd Party SIP market. The following table shows what the market penetration will be over the forecast period.

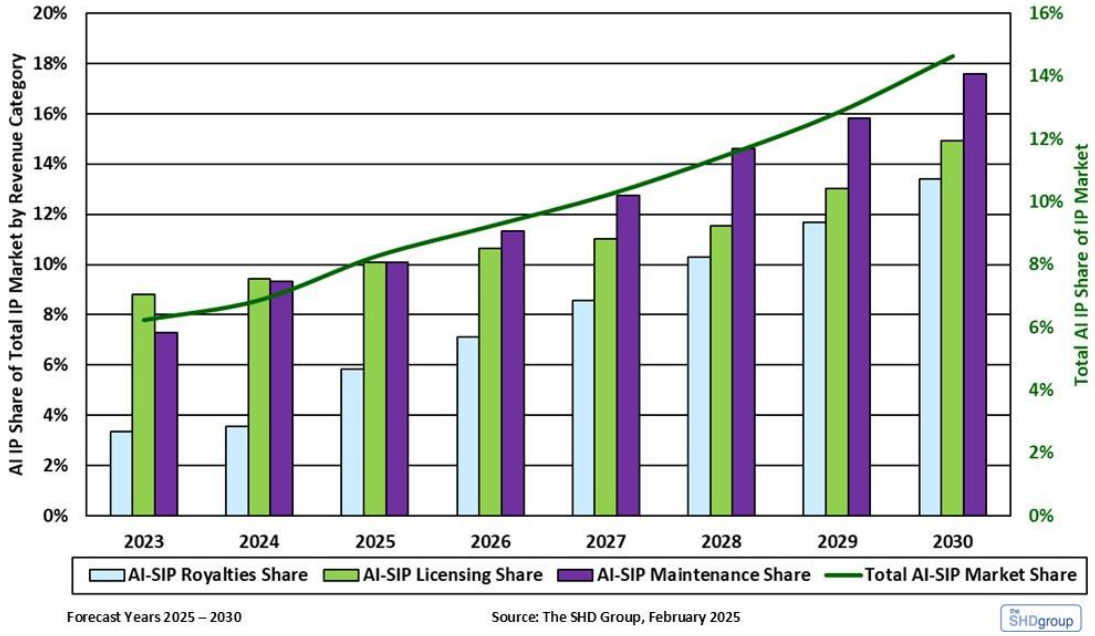
Table 19: AI-SIP Share of Total SIP Market by Revenue Category

Percent of Market	2023	2024	2025	2026	2027	2028	2029	2030
AI-SIP Royalties Share	3.4%	3.5%	5.8%	7.1%	8.6%	10.3%	11.7%	13.4%
AI-SIP Licensing Share	8.8%	9.4%	10.1%	10.6%	11.0%	11.5%	13.0%	14.9%
AI-SIP Maintenance Share	7.3%	9.3%	10.1%	11.3%	12.7%	14.6%	15.8%	17.6%
Total AI SIP Market Share	6.2%	6.9%	8.2%	9.2%	10.2%	11.4%	12.8%	14.6%

Forecast Years 2025 - 2030 **Source: The SHD Group, March 2025**

- In the Worldwide SIP market, total AI-SIP Royalty revenues are projected to be a small portion of total SIP Royalty revenues then grow quickly as more units ship into the market. In 2024, the AI-SIP Royalties revenues reached 3.4% percent and are forecast to grow to 13.4% percent by 2030.
- In the Worldwide SIP market, total AI-SIP Licensing revenues are projected to be a growing portion of total SIP Licensing revenues as many new designs are started to meet market demand. In 2024, the AI-SIP Licensing revenues reached 9.4% percent and are forecast to grow to 14.9% percent by 2030.
- In the Worldwide SIP market, total AI-SIP Maintenance revenues are projected to be a growing portion of total SIP Maintenance revenues as many companies look to gain expertise over this new type of SIP. In 2024, the AI-SIP Maintenance revenues reached 9.3% percent and are forecast to grow to 17.6% percent by 2030.
- In the Worldwide 3rd Party AI-SIP market, total AI-SIP revenues are forecast to continue the growth of the last few years with acceleration in the out-years of this forecast. In 2024, the total AI-SIP revenues reached 6.9% percent of the total SIP market revenues and are forecast to grow to 14.6% percent of the total SIP market by 2030.

Figure 45: AI-SIP Share of Total SIP Market by Revenue Category



Total Design Starts for Edge-AI

Table 20: Total Design Starts by Market Category

Design Start Units	2023	2024	2025	2026	2027	2028	2029	2030	24 - 30
Industrial	105	116	131	149	143	170	172	190	8.6%
Automotive	66	83	105	135	151	148	149	160	11.6%
Networking	74	107	125	142	145	128	128	131	3.4%
Computer	67	87	109	121	125	127	130	139	8.1%
Consumer	119	147	172	174	173	180	183	201	5.4%
Other	11	16	18	17	22	24	30	32	12.2%
Total	442	556	660	738	759	777	792	853	7.4%
Percent Growth	45.8%	25.8%	18.7%	11.8%	2.8%	2.4%	1.9%	7.7%	

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

- In the Edge-AI market, design starts in the Consumer category are forecast to reach the highest design starts out of all the device types analyzed by The SHD Group. In 2024, the design starts were 147 units and are forecast to grow to 201 units by 2030. This represents a CAGR of 5.4%.
- In the Edge-AI Market, the total design starts were 556 units in 2024 and are forecast to grow to 853 units by 2030. This represents a CAGR of 7.4%.

Figure 46: Total Edge-AI Design Starts by Market Category

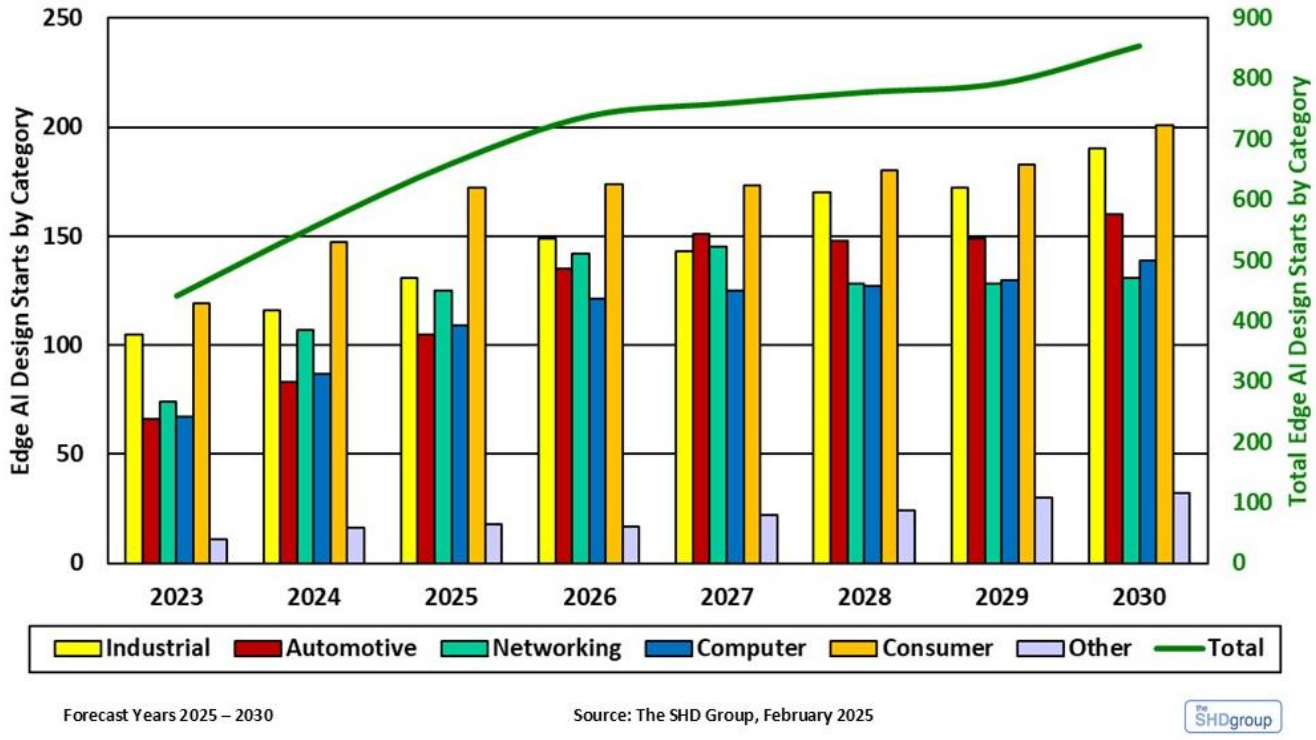


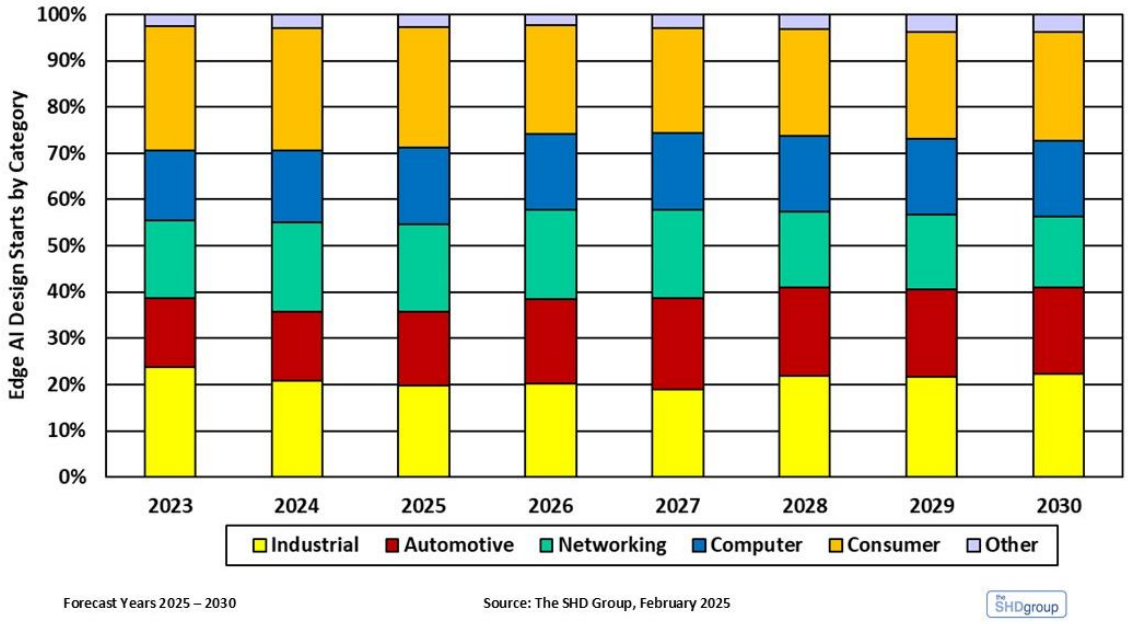
Table 21: Total Market Share of Design Starts by Market Category

Design Start Share	2023	2024	2025	2026	2027	2028	2029	2030
Industrial	23.8%	20.9%	19.8%	20.2%	18.8%	21.9%	21.7%	22.3%
Automotive	14.9%	14.9%	15.9%	18.3%	19.9%	19.0%	18.8%	18.8%
Networking	16.7%	19.2%	18.9%	19.2%	19.1%	16.5%	16.2%	15.4%
Computer	15.2%	15.6%	16.5%	16.4%	16.5%	16.3%	16.4%	16.3%
Consumer	26.9%	26.4%	26.1%	23.6%	22.8%	23.2%	23.1%	23.6%
Other	2.5%	2.9%	2.7%	2.3%	2.9%	3.1%	3.8%	3.8%
Total	23.8%	20.9%	19.8%	20.2%	18.8%	21.9%	21.7%	22.3%

Forecast Years 2025 – 2030 **Source: The SHD Group, March 2025**

- In the Edge-AI market, design starts in the Consumer category are forecast to reach the largest share of design starts out of all the device types analyzed by The SHD Group. In 2024, the Custom ASICs / SoCs design starts were 26.3% of the total and are forecast to have a 24.7% share by 2030.

Figure 47: Total Market Share of Design Starts by Market Category



XIII. Conclusion and Recommendations

Artificial intelligence and machine learning have emerged over the past decade as transformative technologies. Their role at the edge differs from that in computing environments, enabling or improving specific functions for consumer, industrial, and other markets. The SHD Group's research has found that computer vision is the most common function employing AI at the edge, but audio processing and condition monitoring are also important. Consequently, CNNs are the dominant network type, with ResNet, MobileNet, and Yolo being specific models that companies support. Suppliers are interested in transformers and multimodal networks, but their role at the edge remains unclear. Few are aware of state-space models and other emerging technologies.

The technology suppliers we interviewed and surveyed are mostly aligned with developer practices revealed by the Edge-AI and Vision Alliance survey, but they could be overextending themselves. By necessity, they are ahead of most developers when it comes to transformers, but they also support older tools and models with little demand.

We recommend that chip, SIP, and other suppliers seeking a broad customer base focus first on CNNs. If they haven't already, they must also add transformer support and research how these models apply to tasks beyond language models. They should also explore new technologies that could prove useful, such as multimodal and state-space models. They will have to educate developers, however, on these technologies' benefits and how to use them. Small companies may do better to focus on a specific niche, such as audio processing for consumer applications, to avoid the extensive competition targeting CNNs for many designs.

In addition, we also recommend paying close attention to the introduction of AI Agents to the market. These applications will likely be run on a growing number Edge-AI devices and systems as their utility and value is proven to end users. The key here is the ability of AI Agents to enhance productivity and the decision-making capabilities of their users. Given their ability to take action based on different sets of inputs the user has predetermined, they hold the promise of providing personalized, real-time and scalable support to their users across multiple domains.

If the promise of multimodal LLMs and AI Agents is realized, then we believe the impact on the semiconductor market will be profound and long-lasting.

XIV. Edge-AI Company Ecosystem Guide

The following is a list of companies The SHD Group interviewed and surveyed for data to create this report. We wish to extend our thanks to our sponsors and all these companies and individuals without whose support this report would not have been possible.

Andes Technology

Website
www.andestech.com

Headquarters
Hsinchu Science Park, Taiwan

Founded: 2005

Product Categories
RISC-V CPU SIP
Complier / development software
Peripheral SIP

Target Segments
AI, 5G, Automotive, IoT, Networking, SSD and Wearables

Edge AI Products

Andes N-Series RV32/64C

Status: In production
TOPS: >10MOPS
Key Selling Points: Designed for cost-sensitive, low-power applications and has configurable extensions for efficiency and flexibility. Focused on embedded and IoT applications that require high code density and efficiency.

Andes D/A-Series RV32/64P

Status: In Production
TOPS: >10MOPS – 500MOPS
Key Selling Points: Designed for high-performance AI and compute-intensive tasks, featuring DSP and Vector processing. Target applications include AI, Automotive, and high-end embedded systems.

Andes V-Series RV64GCPV

Status: In production
TOPS: 500MOPS – 100TOPS
Key Selling Points: Designed for high-performance AI, HPC and large-scale data processing, leveraging full RISC-V Vector Extension(RVV) support for parallel computing. Target applications include 5G & Telecommunications, Automotive & ADAS, AR/VR & graphics processing, AI & Machine Learning and Cloud computing.

Apple, Inc.

Website
www.apple.com

Headquarters
Cupertino, CA

Founded: 1976

Product Categories
Semiconductors

Target Segments
Consumer Products, Edge Computing, (PCs, Laptops & Tablets)

ARM Holdings plc

Website
www.arm.com

Headquarters
Cambridge, UK

Founded: 1990

Product Categories
CPU SIP, GPU SIP, NPU SIP, Security SIP SoC Platforms, MCU & Embedded SIP, Interconnect SIP, Software & Development Tools, Customizable SIP

Target Segments
Mobile & Consumer systems, IoT, Automotive, Data Center, Networking, Computer, AI & Machine Learning

Arteris IP

Website
www.arteris.com

Headquarters
Campbell, CA

Founded: 2004

Product Categories
Interconnect Network-on-Chip IP
Coherent and Non-Coherent Cache NoC IP
FlexGen smart NoC IP

Target Segments
Automotive, Communications, Consumer, Enterprise Computing, Industrial

FlexGen™ smart Network-on-Chip IP

Status: In production

Key Selling Points: FlexGen™ smart NoC IP automates NoC design. Using AI/ML-driven automation, FlexGen enables the efficient generation of optimized NoC designs with reduced manual effort, shorter iterations, and expert-level quality including the optimization of power, performance, and area (PPA). As SoCs scale to 500+ IP blocks and beyond, AI, HPC, and multi-die architectures push the limits of traditional interconnect design methodologies, making NoC development increasingly complex.

Axelera AI

Website
https://axelera.ai

Headquarters
Eindhoven, The Netherlands

Founded: 2021

Product Categories
Metis AI Processing Unit
Metis AI Evaluation Systems
M.2 AI Edge Acceleration Card
Voyager SDK Stack

Target Segments
Security, Retail, Industrial

Ceva, Inc.

Website
www.ceva-ip.com

Headquarters
Rockville, MD, USA

Founded: 2002

Product Categories
Semiconductor IP
Software

Target Segments
Automotive / Transportation
Consumer Products
Edge Computing (PCs, inc. tablets)
Networking / Communications
Industrial / Mfg.

Edge AI Products

Ceva-NeuPro-M

Status: sampling now
TOPS: 4 - 1200 TOPS
Key selling points: Ceva-NeuPro-M redefines high-performance AI processing for smart edge devices with heterogeneous coprocessing, targeting generative and classic AI models. Ceva-NeuPro-M is a scalable, self-contained NPU architecture that concurrently processes diverse AI workloads using mixed precision MAC array, sparsity engine, weight and data compression, and a programmable VPU for future networks.

Ceva-NeuPro-Nano

Status: sampling now
TOPS: Up tp 200 GOPS (0.2 TOPS)
Key selling points: Ceva-NeuPro-Nano is a highly efficient, self-sufficient Edge NPU designed for TinyML applications. It delivers the optimal balance of ultra-low power and the best performance in a small area to efficiently execute TinyML workloads across AIoT product categories, including Hearables, Wearables, Home Audio, Smart Home, Smart Factory, and more.

Ceva-SensPro2

Status: in production
TOPS: 0.2 - 4 TOPS
Key selling points: Ceva-SensPro2 is a highly scalable, high-performance Vision and AI DSP for multitasking sensing and AI of multiple sensors. Ceva-SensPro2 maximizes performance-per-watt for multi-sensor processing by utilizing a combination of single and half precision floating point math, neural network processing, along with parallel processing capacity for Vision, SLAM, Radar, and LiDAR.

Ceva-NeuPro Studio

Status: in production
Key selling points: The Ceva-NeuPro Studio is a comprehensive AI SDK for creating fully optimized runtime software for Ceva-NeuPro-Nano and Ceva-NeuPro-M NPUs, and Ceva-SensPro DSPs. Ceva-NeuPro Studio supports multiple open AI inference frameworks and incorporates a broad range of network optimizations and quantization algorithms, data flow management and libraries into a holistic solution.

ESWIN Computing

Website
www.eswincomputing.com

Headquarters
Beijing, China

Founded: 2019

Product Categories
Semiconductors
Semiconductor IP

Target Segments
Edge Computing (PCs, inc. tablets)

Expedera Inc

Website
www.expedera.com

Headquarters
Santa Clara, CA, USA

Founded: 2018

Product Categories
Semiconductor IP

Target Segments
Automotive / Transportation
Consumer Products
Edge Computing (PCs, inc. tablets)
Industrial / Mfg.

Horizon.cc

Website
www.horizon.cc

Headquarters
Rockville, MD, USA

Founded: 2015

Product Categories
Semiconductors
Software
AI Services
Systems, boards, other OEM products

Target Segments
Automotive / Transportation

Edge AI Products

Journey 5 ADAS Solution
Status: in production

Huawei

Website
www.huawei.com

Headquarters
China

Founded: 1992

Product Categories
Semiconductors
Semiconductor IP
Software
AI Services
Systems, boards, other OEM products

Target Segments
Automotive / Transportation
Consumer Products
Edge Computing (PCs, inc. tablets)
Networking / Communications
Industrial / Mfg.

Inuitive

Website
www.inuitive-tech.com

Headquarters
Israel

Founded: 2012

Product Categories
Semiconductors
Systems, boards, other OEM products

Target Segments
Automotive / Transportation
Consumer Products
Edge Computing (PCs, inc. tablets)
Industrial / Mfg.

Edge AI Products

NU4000

Status: in production
TOPS: 1.6
Key selling points: Integrated Perception Processor - offer multiple vision functions in a single cost-efective chip: 6 Cameras I/F, Depth Processing, Computer Vision, AI, VSLAM, and more.

NU4500

Status: expected to sample in Quarter-Year
TOPS: 8
Key selling points: Integrated Perception Processor - offer multiple vision functions in a single cost-efective chip: 6 Cameras I/F, Depth Processing, Computer Vision, AI, VSLAM, Dual Camera ISP and more.

NU4500

Status: expected to sample in Quarter-Year
Key selling points: Integrated Perception Application Processor - offer multiple vision functions in a single cost-efective chip: 10 Cameras I/F, Depth Processing, Computer Vision, AI, VSLAM, High Throughput/Resolution Dual Camera ISP, Power CPU, High throughput Video Encode/Decode and more. Fast interfaces such as PCIe Gen4, USB3.0, LPDDR5 etc.

Krispan Incorporated

Website
www.krispan.com

Headquarters
Santa Clara, CA, USA

Founded: 2006

Product Categories
Semiconductor IP
Software
AI Services

Target Segments
Edge Computing (PCs, inc. tablets)
Networking / Communications.

LIST Semiconductor (Leading Interconnect Semiconductor Technology)

Website
www.leadingics.com/en

Headquarters
Shenzen, China

Founded: 2019

Product Categories
Semiconductor Packing Substrates

Target Segments
Automotive / Transportation
Consumer Products
Networking / Communications

MACSO Technologies

Website
www.macso.ai

Headquarters
Auckland, New Zealand

Founded: 2021

Product Categories
AI Services

Target Segments
Edge Computing (PCs, inc. tablets)
Industrial / Mfg.

Edge AI Products

Animal Health Monitor

Status: in production
Inference rate: Variational auto-encoder
Key selling points: Uses neural networks to perform noise filtration and classification of audio signals that may represent the presence of respiratory disease.

Smoke/Vape Detector

Status: in production
Inference rate: LSTM
Key selling points: Uses neural network to classify whether the recorded time series of measured aerosol particles represents presence of smoke or vape

Air Quality Monitor

Status: in production
Inference rate: Variational autoencoder
Key selling points: Uses neural network to quantify with uncertainty the effective air change rate based on measurements of ambient aerosol particles. This is used to quantify the efficacy of HVAC systems

Fall detector

Status: expected to sample in Quarter-Year
Key selling points: Uses sensor fusion between audio and motion sensors to classify whether a movement and consequential audio signal represents a harmful fall

Matsuada

Website
www.matsusada.com

Headquarters
Japan

Founded: 1978

Product Categories
Semiconductors
Semiconductor IP
AI Services

Target Segments
Automotive / Transportation
Consumer Products
Networking / Communications

OpenMV

Website
www.openmv.io

Headquarters
Atlanta, GA, USA

Founded: 2015

Product Categories
Machine Vision Cameras, Camera Modules
Expansion Shields

Target Segments
System Developers, Education & Research,
Industrial Automation, Drones & Robotics,
AI & IoT, Security & Surveillance

OpenMV N6

Status: In Production

Key Selling Points: The OpenMV N6 is a small, low-power microcontroller board which allows you to easily implement applications using machine vision in the real world. The OpenMV Cam can be programmed using high level Python scripts instead of C/C++, making it easier to deal with the complex outputs of machine vision algorithms and working with high level data structures.

Piera Systems

Website
www.pierasystems.com

Headquarters
Mississauga, Canada

Founded: 2018

Product Categories
Software

Target Segments
Automotive / Transportation
Consumer Products
Industrial / Mfg.
Environmental Sensing

Quadric

Website
www.quadric.io

Headquarters
Burlingame, CA, USA

Founded: 2016

Product Categories
General Purpose Neural Processing Unit (GPNPU)

Target Segments
Automotive, Robotics, Surveillance and Augmented Reality

Edge AI Products

Chimera GPNPU

Status: In production
Key Selling Points: The Chimera GPNPU family provides a unified processor architecture that can handle matrix and vector operations and sclar (control) code in one execution pipeline. The Chimera GPNPU is a single software-controlled core, allowing for simple expression of complex parallel workloads.

Samsung

Website
www.samsung.com

Headquarters
Suwon, S. Korea

Founded: 1950

Product Categories
Semiconductors
Semiconductor IP
Software
AI Services
Systems, boards, other OEM products

Target Segments
Automotive / Transportation
Consumer Products
Edge Computing (PCs, inc. tablets)
Networking / Communications
Industrial / Mfg.

Sondrel

Website
www.sondrel.com

Headquarters
Reading, UK

Founded: 2002

Product Categories
Custom Soc, ASIC, Semiconductor Design Services

Target Segments
Automotive / Transportation
Edge Computing (PCs, inc. tablets)
Networking / Communications
Industrial / Mfg.

Sophgo

Website
www.sophgo.com

Headquarters
Beijing, China

Founded: 2016

Product Categories
Semiconductors

Target Segments
Consumer Products
Edge Computing (PCs, inc. tablets)

Edge AI Products

G2380

Status: in production
TOPS: 32T@int8, 16T@fp16
Inference rate: Focus LLM
Memory: LPDDR5
Key selling points: focus on risc-v +ai aipc market and edge market

SG2300

Status: in production
TOPS: 32T@int8, 16T@fp16
Inference rate: LLM opencv
Memory: LPDDR4x
Key selling points: focus on AI market

TetraMem Inc.

Website
www.tetramem.com

Headquarters
Fremont, CA, USA

Founded: 2018

Product Categories
Semiconductors

Target Segments
Automotive / Transportation
Consumer Products
Edge Computing (PCs, inc. tablets)

Texas Instruments

Website
www.TI.com

Headquarters
Dallas, TX, USA

Founded: 1930

Product Categories
Semiconductors
Software

Target Segments
Automotive / Transportation
Consumer Products
Networking / Communications
Industrial / Mfg.

MagikEye, Inc.

Website
www.magik-eye.com

Headquarters
Stamford CT, USA
Tokyo, Japan
Prague, Czech Republic

Founded: 2014

Product Categories
FPGA Prototyping Platform, FPGA-Assisted Verification tools, Multi-debug Modules, Rapid prototyping software, High speed interface daughter cards

Target Segments
Facial Recognition, Robotics, AR/VR, Industrial Automation, Access Control & Security, Automotive.

Edge AI Products

ILT Evaluation Systems

Status: available now for customer evaluations and integration

Key Selling Points: Licensable 3D depth sensor technology based on using an infrared laser and a CMOS image sensor, using a unique algorithm called Invertible Light™ Technology (ILT) developed by MagikEye. ILT can create point cloud data at high speeds and with very low power using low cost hardware. With this, MagikEye products can measure depth as close as 5cm and reaching to 5m.

S2C EDA Solutions

Website
www.s2cinc.com

Locations
San Jose CA, USA
Tokyo, Japan
Seoul, Korea
Hong Kong
Beijing, China
Shanghai, China
Shenzhen, China
Xi'an, China

Founded: 2003

Product Categories
FPGA Prototyping Platform, FPGA-Assisted Verification tools, Multi-debug Modules, Rapid prototyping software, High speed interface daughter cards AI Services

Target Segments
Autonomous driving, Surveillance
5G, Data Center, HPC and AI/ML.

Edge AI Products

Prodigy™ S8-100 Logic System

Status: In production
Key selling points: Our flagship system is the Prodigy™ S8-100 Logic System is designed to meet the demands of AI and HPC. Available in Single, Dual and Quad FPGA configurations, the S8-100 supports medium-scale to hyperscale designs with ease, making it a versitile choice for advanced chip development.

Edge AI and Vision Alliance

Website
www.edge-ai-vision.com

Headquarters
Walnut Creek CA, USA

Founded: 2011

Product Categories
Conferences, Seminars, Workshops, Member Surveys, Panel discussions, Member presentations, all things related to Edge AI and Vision applications. Vibrant and growing membership of Semiconductor companies, Software developers, SIP vendors, Product developers, and Systems companies in the Embedded Vision and Edge AI markets. Quarterly conferences and monthly update meetings.

Target Segments
Automotive, Consumer, Computer, Edge Computing, Industrial, Networking, Retail, Security

Edge AI Products

Status: All areas active now

Key selling points: The Edge AI and Vision Alliance is a worldwide industry partnership bringing together technology providers and end-product companies to accelerate the adoption of edge AI and vision in products. Alliance membership is open to any company that supplies or uses technology for AI or vision systems and applications. Member benefits include early insights into new markets, technologies, applications and standards; increased influence over the direction of the industry; and increased visibility to analysts and media. Members are eligible to sponsor, exhibit and present at the highly regarded and well-attended Embedded Vision Summit conference, and have access to the Alliance's Members-only educational and networking events, the Edge AI and Vision Innovation Forums, held throughout the year.

XV. Definitions

Edge-AI Definitions

Artificial intelligence (AI): the technology and science of computers performing tasks that historically required human intelligence. The term AI has come to also include machine learning, and this report uses the term accordingly.

Machine learning (ML): the technology and science of **models** that can learn and adapt without explicit programming. ML models include neural networks and **classical machine-learning** approaches (statistical analysis algorithms) such as regression, clustering, and support vector networks.

Neural network: an AI/ML modeling technique patterned after the human brain and nervous system and based on interconnected nodes typically organized in layers (stages).

Training: the ML process whereby a system such as a neural network learns. Because it's computationally intense, training often occurs on a high-performance system, and the results are transferred to a different system that performs inferencing.

Inference: the application of a trained AI/ML model.

Edge: a vague term referring to systems outside a data center or more specifically those located near real-world data sources. **Edge computing** more specifically refers to PCs and servers in the home, enterprise, or network service-providers' access networks.

TOPS: an abbreviation for tera-operations per second, a simple performance metric for computing systems.

Quantization: converting a model from using one data type to a less precise one. Neural-network models are typically trained using 16-bit floating-point (FP16) data. In some cases, less-precise data types can be used without unduly sacrificing accuracy. These types include 8-bit integer (INT8), 8-bit floating point (FP8), or 8-bit block floating point (BF8). Because these formats are smaller than FP16, they require less memory and less hardware to process.

Agentic AI: evolution of artificial intelligence programs to have the ability to self-organize, be self-reflective, be self-running and be proactive. It can take data inputs that are pre-defined from pre-defined sources and make decisions based on those inputs and then execute actions based on those decisions. In this process the agentic AI is self-learning, improving its decision-making capabilities as it is used.

Semiconductor Device-Type Definitions

Chip: a finished semiconductor product.

SoC (System on Chip): a chip integrating various technologies embodying the core functions of a system.

SIP (Semiconductor Intellectual Property): a licensable design.

CPU (Central Processing Unit): a general-purpose microprocessor chip or an IP block that executes instructions. Examples: Arm Cortex-A series cores.

GPU (Graphics Processing Unit): a chip or an SIP block that executes functions related to computer graphics. GPUs have been adapted to perform other functions, such as those related to AI.

MCU (MicroController Unit): Originally a microprocessor with on-chip code storage but now a low-performance CPU (e.g., an Arm Cortex-M core) or a microprocessor or SoC.

NPU (Neural Processing Unit): a chip or SIP block for processing AI functions, such as neural networks. We use the NPU term generally, including GPUs adapted to AI processing.

DSP (Digital Signal Processor): a chip or SIP block that executes instructions for performing mathematical functions on digitized real-world signals.

FPGA (Field-Programmable Gate Array): a chip or SIP block containing logic functions that can be altered after manufacturing.

ASIC (Application-Specific Integrated Circuit): a chip designed for a single customer. Note that some people use the term to apply to ASSPs as well.

ASSP (Application-Specific Standard Product): a chip designed for a single purpose and available on the merchant market to multiple customers.

Neural Network Types Definitions

An **autoencoder** is a special encoder-decoder case that attempts to copy its input to its output. Obtaining the intermediate representation during training can be one reason for a user to employ an autoencoder. Another reason is to compare the input to the output, which is essentially a cleaned-up version of the input; large differences indicate the input is somehow anomalous.

A **Convolutional Neural Network (CNN)** applies a feature detector to the input (convolves the input with a filter) and passes the results to subsequent layers that apply various activation functions. Convolution employs matrix math, driving performance measured in TOPS. CNNs also use nonlinear **activation functions;** examples include the sigmoid function, hyperbolic tangent (tanh), rectified linear unit (ReLU), and softmax. To maximize performance, AI accelerators, therefore, must offload these functions as well as matrix operations. Computer vision functions such as object detection and classification commonly use CNNs, and the network type can be used for other tasks, such as audio classification. Because computer vision is common at the edge, many Edge-AI processors accelerate CNNs.

Diffusion models are a generative AI technology primarily used for image generation but are also expanding into language processing, 3D modeling, and audio synthesis. They can incorporate CNNs or transformers, depending on the architecture. Beyond image generation, diffusion models can also perform inpainting (filling missing parts of images), denoising, and super-resolution. Recent research explores their use in text generation and structured data synthesis.

Encoder-decoder Networks perform sequence-to-sequence transforms, such as translating English text to French. The encoder transforms input data into an intermediate representation that the decoder transforms to create the output. Since the advent of transformer networks, they have become the dominant encoders and decoders. However, other neural-network types, such as CNN, LSTM, and RNN have been employed. For example, an image-caption generator can encode with a CNN and decode (generate the caption text) with an RNN.

A **Generative Adversarial Network (GAN)** is a machine-learning approach involving two neural networks: a **generator** that creates synthetic data and a **discriminator** that evaluates whether the data is real or fake. During training, the discriminator provides feedback to the generator, refining its ability to produce realistic outputs. GANs have been **widely used for generating images**, including tasks such as **image synthesis, super-resolution,**

inpainting, and style transfer. However, their applications extend beyond images to **video synthesis, text-to-image generation, data augmentation, and even audio generation** (e.g., music and speech).

Multimodal networks, including multimodal transformer models, process multiple data types (modalities), such as images, text, and audio. Some consist of separate models whose outputs feed into a final model, while others jointly learn from all inputs in an end-to-end manner. An example application is image captioning, where the system identifies objects, their relationships, and actions in a scene, then translates this understanding into text. Multimodal transformers like BLIP-2, Flamingo, and SimVLM handle this task.

A **Recurrent Neural Network (RNN)** operates on sequential data. Unlike in a CNN, the processing of each data point depends upon those earlier in the sequence. RNN activation functions include sigmoid, tanh, and ReLU. Applications include music generation, time-series forecasting, and natural-language processing (NLP) tasks such as text translation and sentiment classification. Transformers have largely replaced RNNs for NLP, however.

Although an RNN processes the current input based on previous inputs, it can overemphasize recent inputs over older ones. **Long Short-Term Memory (LSTM)** networks are a type of RNN that is better at factoring in long-term dependencies. LSTM applications overlap those of traditional RNNs, performing better on tasks such as speech recognition and largely obviated for NLP. Edge-AI processors for smart speakers and other systems responding to wake words typically accelerate LSTM models and other RNNs.

A **Spiking Neural Network (SNN)** differs from Artificial Neural Networks (ANNs) such as CNNs and RNNs by processing pulses (spikes) instead of numerical values, with pulse timing conveying information. Their operation resembles that of a biological brain more closely than ANNs, and thus SNNs are a type of neuromorphic computing. Training entails either converting a trained ANN model or through SNN-specific techniques analogous to biological processes. Training technology lags hardware development, but TensorFlow and the PyTorch-related SpykeTorch simulator support SNNs.

Advantages of SNNs include their low power, low latency, and ability to learn in the field as they infer. Their low power and latency derive from their processing events instead of the status quo; for example, reacting to differences (if any) in sequential video frames instead of processing all pixels in each frame. SNN applications include time-series predictions, speech recognition, computer vision, motor control, and robot motion. Despite SNNs' advantages, their immaturity narrows their applicability to the rare designs focused on neuromorphic computing.

Transformer Neural Networks include encoder models (e.g., Bert) and decoder models (e.g., GPT). The former primarily focuses on understanding text and is used for tasks like question answering and sentiment analysis, while the latter is autoregressive and specializes in text generation. Some encoder-decoder models (e.g., T5, BART) can both understand and generate text by leveraging both architectures.

Transformers excel at Natural Language Processing (NLP) and have also been adapted to fields such as protein folding (AlphaFold), drug discovery, and image segmentation. The latter belongs to a family of models called Vision Transformers (ViTs), which apply Transformer architecture to image patches instead of text tokens.

Transformer models tokenize inputs into discrete units (tokens or patches) and map them to embeddings, which are high-dimensional vectors that capture meaning. These embeddings represent relationships between words (e.g., "king" - "man" = "queen" - "woman").

A key feature **Multimodal Networks**, including multimodal transformer models, process multiple data types (modalities), such as images, text, and audio. Some consist of separate models whose outputs feed into a final model, while others jointly learn from all inputs in an end-to-end manner. An example application is image captioning, where the system identifies objects, their relationships, and actions in a scene, then translates this understanding into text. Multimodal transformers like BLIP-2, Flamingo, and SimVLM handle this task.

State-Space AI Models (SSMs) emerged in 2021, four years after the seminal transformer paper. They can be applied to text, vision, audio, and time-series processing. They're better at handling long input sequences than transformers due to their efficient memory usage. However, they struggle with tasks requiring precise copying or retrieval of input data. While SSMs are not yet widely deployed in data-center-based AI, recent research suggests they could be competitive with transformers for certain workloads. They are not well known among Edge-AI companies, but their low memory footprint makes them a potential fit for edge applications in the future. There is growing interest in mapping SSMs onto Spiking Neural Networks (SNNs), though this remains an early-stage research area. Our survey did not ask about state-space models.

Specific Neural Networks

3D-UNet is a CNN for volumetric (3D) images and is used in medical imaging applications such as MRI and CT scan analysis. It is included in some MLPerf benchmarks from MLCommons medical image segmentation.

Bert is a transformer-based language model. Developed by Google, it was first applied to search ranking and snippet extraction. It is included in MLPerf benchmarks for NLP tasks.

DS-CNN (Depthwise Separable CNN) is a lightweight CNN for spotting keywords in speech samples. It is optimized for low-power AI applications like wake word detection and is included in the MLPerf Tiny benchmark.

FC Autoencoder is a neural network for anomaly detection in machine sounds and is included in MLPerf.

MobileNet is a CNN for edge-based computer vision functions such as image classification and object detection. It prioritizes efficiency and low power usage, using depthwise separable convolutions. It has three major versions (V1–V3), with EfficientNet-Lite as its modern successor.

MobileNet SSD is a variant that integrates MobileNet with Single Shot Detector (SSD) for object detection. The MLPerf Tiny benchmark employs MobileNet V1 for visual wake word detection, while the MLPerf Mobile benchmark uses MobileNet SSD and MobileNetEdgeTPU for object detection and classification.

ResNet (Residual Network) is a family of CNNs developed by Microsoft for image classification and object detection. ResNet-50, a 50-layer variant, is a common benchmark for AI processors.

Attention-enhanced ResNet variants (e.g., SE-ResNet, CBAM-ResNet) add attention steps to improve features extraction and performance in deep networks.

RetinaNet is an object-detection network from Facebook AI Research (FAIR) that enhances traditional one-stage detectors by adding subnetworks for classification and bounding box prediction. It is also included in the MLPerf benchmarks for evaluating object detection models.

RNN-T (RNN Transducer) is a neural network architecture for automatic speech recognition (ASR), optimized for real-time streaming applications. It's smaller and more efficient than many previous ASR models, making it well suited for edge devices like smartphones and voice assistants.

Training Frameworks and Tools

Darknet is an open-source framework known for implementing YOLO, but development has ceased.

JAX is a high-performance numerical computing library with JIT compilation for efficient AI workloads on CPUs, GPUs, and TPUs.

Keras is a high-level Python API for AI model building, primarily used with TensorFlow, but also supporting JAX and PyTorch through Keras Core.

MXNet is an Apache deep learning framework once backed by Amazon, but development has largely stalled in favor of PyTorch.

NVIDIA TAO Toolkit is a transfer learning tool that fine-tunes pretrained models for domain-specific tasks.

MATLAB is a scientific computing environment used for AI, ML, and system modeling, supporting both classical ML and deep learning.

OpenCV is an open-source computer vision and ML library with Python, Java, C++, and MATLAB bindings.

PyTorch is a flexible AI/ML framework developed by Meta, now governed by the PyTorch Foundation. It features dynamic computation graphs (eager execution) and can offload computations to GPUs, TPUs, and other accelerators.

TensorFlow is a Google-originated AI/ML framework extended by various tools and libraries (e.g., Keras). A key function is to process tensor-based computations and define data flow through nodes performing operations like matrix multiplication. TensorFlow supports both eager execution and graph compilation (tf.function). It handles training, inference, and deployment to accelerators (GPUs, TPUs, and other AI hardware), including multi-device distribution.

TensorFlow Lite (TFLite) is a lightweight inference framework for edge devices, supporting a subset of TensorFlow's operations (~700 vs. ~7,000 ops). A typical workflow involves training in TensorFlow and converting to TFLite for deployment.

Triton (OpenAI) is a Python-based parallel programming framework that compiles to CUDA for optimized GPU performance, serving as an alternative to NVIDIA CUDA for AI acceleration.

Inference Frameworks and Tools

Darknet is a neural network framework recognized for its implementation of the Yolo model. An open-source project that has seen no activity in a few years, it's defunct.

Jax is a Python library for array-based computation. It can execute on CPUs or NPUs (e.g., Nvidia GPUs and Google TPUs) and employs just-in-time compilation. Various other Python libraries rely on Jax.

Keras is a Python API to a function library to facilitate building AI models. Initially designed to work with TensorFlow, it also works with Jax and PyTorch.

TensorFlow is a Google-originated AI/ML framework extended by various tools and libraries (e.g., Keras). As its name implies, a key function is to enable a developer to specify how data stored in tensor objects (similar to matrices) flow through nodes that operate on the data (e.g., by performing matrix multiplication). TensorFlow ultimately compiles the neural network into a graph. It then optionally performs the iterative process that trains the model and subsequently performs inferencing (runs the trained model). TensorFlow also handles offloading networks to accelerators (NPUs), even distributing work among multiple processing units.

TensorFlow Lite is a set of tools for inference on edge devices and supports a subset of TensorFlow capabilities. For example, it defines more than 700 operators compared with nearly 7,000 for TensorFlow. A typical workflow would be to train a model with TensorFlow and convert it to TensorFlow Lite. TensorFlow Lite includes a runtime environment, enabling it to execute the model.

MXNet is a neural network framework for training and deployment. An Apache project once backed by Amazon, it's no longer developed and is defunct.

Nvidia TAO is a set of software tools to facilitate transfer learning, which is the process of adapting a pretrained model to a new task such as modifying a generic object-recognition model to recognize only a specific object type not in the original training data.

Matlab is a programming environment used by scientists and engineers for math-intensive tasks and visualization. Well known for signal processing, it's also possible to use it for AI/ML. Matlab can help with preparing data, developing and running AI models (including classical non-neural-network models), and modeling a bigger system that incorporates AI.

OpenCV is an open-source computer-vision library that can be accessed from Python, Java, Matlab, and C++. In addition to vision functions, it also provides functions for classical machine learning and neural networks.

PyTorch is an AI/ML Python library providing a framework for describing and training neural networks. Originally defined by Meta (Facebook), it's now governed independently. It's more flexible and easier to use than TensorFlow. Like TensorFlow, PyTorch has tensors and graphs as its core components, although it handles graphs differently. It also can offload computation to hardware accelerators (e.g., NPUs and GPUs). PyTorch has more than 2,000 operators.

Triton from OpenAI wasn't included in our survey. It's a Python-based parallel-programming environment and comprehends a Python-like language and compiler. It's associated with data-center GPUs and is an alternative to Nvidia Cuda, automating memory handling and scheduling computations to improve developer productivity. Like PyTorch, it can also perform inference.

Inference Frameworks and Tools

The **Android NN API** is a soon-to-be deprecated software layer for frameworks such as TensorFlow Lite to enable them to perform hardware-accelerated or CPU-based inference operations.

Caffe was developed at the University of California, Berkeley, and Meta (Facebook) later developed Caffe2, which was subsequently merged into PyTorch, rendering it defunct. In addition to Python, C++ can call Caffe2, enabling it to be employed in the field.

DeepStream is a pipeline-based framework from Nvidia for analyzing digital media. It helps developers create stream-processing pipelines incorporating processing functions, including neural networks, to enable real-time analytics.

ONNX Runtime executes neural networks converted to the **ONNX format**, a neutral format also used with chip and SIP vendors' proprietary runtimes. Developed by Microsoft, the ONNX Runtime runs on the host CPU and can optionally offload computation to an Nvidia GPU. It runs under Windows, Android, and iOS; it is accessible from Python, C++, and other languages.

OpenCV DNN is an optional OpenCV module that supports inference, particularly using x86 CPUs. It supports various frameworks, including TensorFlow and PyTorch, as well as models in the ONNX format. The DNN module focuses on supporting vision-related models.

OpenVino is software for optimizing and deploying neural networks. Backed by Intel, it's an open-source project mainly supporting Intel hardware (e.g., x86 CPUs) but it also supports Arm CPUs and Intel encourages other contributions. A developer employing OpenVino typically ingests a model in a standard format (e.g., in ONNX

format or from PyTorch or TensorFlow), optimizes it using OpenVino, and executes inference with the software's runtime engine.

PyTorch can perform inference in addition to being used for model training. It also integrates with the ONNX Runtime, and the PyTorch Foundation has developed the **ExecuTorch Runtime** that runs on edge devices and runs compiled models. It would be unusual to employ PyTorch for inference on an edge device.

TensorFlow can perform inference in addition to model training. It would be unusual to employ TensorFlow for inference on an edge device.

TensorFlow Lite includes a runtime engine for on-device inference. Target devices include smartphones, microcontrollers, and other edge devices.

TensorFlow Lite Micro is a version of TensorFlow Lite for devices such as microcontrollers that can spare only kilobytes of memory. The runtime requires 16 KB for an Arm Cortex-M3 CPU and can run without an OS.

Tensor RT is an inference library and runtime from Nvidia. Related software supports large language models and model optimization.

Table 22: Addendum Spreadsheet Tables, 1 – 193 and Figures 1 - 187

Addendum Spreadsheet Tables (available to Sponsors via Addendum)

Addendum Spreadsheet Figures (available to Sponsors via Addendum)