Publication date:
February 2024
Author:
Alexander Harrowell

Generative AI At The Cutting Edge

Prospects, Applications, And The Impact On Hardware

Brought to you by Informa Tech

Commissioned by:

Ambarella®
AI envisioned™

Contents

Ambarella Foreword

With the advent of AI and neural network processing over recent years, we have witnessed the capabilities of devices increase many times over. We are now on the cusp of witnessing another even more significant increase in performance and capabilities through the deployment of Large Language Models and Generative AI at the edge.

For 20 years Ambarella's focus has been on solving the unique power, performance, latency, and privacy demands at the edge – and this is now being applied to the significant GenAI workloads.

The progress in Generative AI so far has primarily focused on servers and the training of ever-growing language models. At Ambarella, an edge AI semiconductor company, we view this as the initial phase, leading to widespread technology adoption through scalable integration at the edge, encompassing everything from robots and consumer devices to security systems and autonomous vehicles. When integration at scale occurs, technological challenges such as power consumption, fine tuning on device, reliability at scale, and cost, come to the forefront, all necessitating the right dedicated SoC architecture.

At Ambarella, we have been delivering video and AI products to consumers and businesses for over two decades. From security systems to consumer sports cameras to automotive vision safety systems, we have always focused our SoC architecture on solving the challenges of delivering high performance at low power. We recognize that the evolution of AI hinges on more than just performance. It's about balancing power efficiency with scalability. The progression of generative AI, powered by increasingly sophisticated language models, calls for specialized, energy-efficient hardware. This must be crafted for high efficiency in a variety of applications from multi-modal AI boxes & hubs for smart cities to robotics and autonomous driving.

Semiconductor technology is set to play a vital role in integrating generative AI into edge computing, opening new pathways, and contributing to the evolution of AI.

To offer insights into the present and future of generative AI, highlighting the importance of efficient, high-performance solutions in edge computing, we've commissioned this independent analyst whitepaper with Omdia to give their view on the market needs and trends. We invite you to join us in this journey as we explore AI's potential and pave the way for a smarter world.

Executive Summary

Generative AI has seized everyone's attention through the combination of creativity, versatility, and scale. However, the boom in scale that has rendered this possible has become a problem in itself, with AI training and inference becoming the pace-setting compute workload of the 2020s. This is a major barrier to bringing the possibilities of the new AIs to the edge – which might be where much of their potential value lies.

2022's surge of open-source AI innovation has opened the door a crack, offering a huge variety of AI models that fall into the so-called "missing middle" category between 5 and 50 billion parameters, thus being just about in reach for edge inference. However, we still need robust hardware solutions for edge AI – and especially generative AI at the edge – that are capable of serving this class of model within the especially demanding price/power/area constraints of edge and embedded systems if the privacy, latency, and bandwidth cost advantages of edge AI are to be realized.

Vendors from the mobile space are keen to widen their repertoire and also to re-use their hardware architecture for PC or data center products, but they are making slower progress than expected in 2021 as they try to address markets such as cameras, robotics, developer workstations, or AI boxes supporting multiple sensors. Similarly, there's little evidence that they're addressing the emerging demand for hardware to cover both inference and fine-tuning of smaller LLMs as we move beyond the initial AI training boom.

The three Ps of AI at the edge are privacy, performance, and power efficiency, but a fourth P – developer productivity – is crucial to delivering them.

A solution will need strong performance per watt, which in turn will require optimization as a whole system across the CPU, GPU, and neural processing unit (NPU) domains. Although the mobile players excel at this, it will also need a strong showing across a wide variety of Transformer-based models, where they do much less well, and scalability both up to powerful systems such as edge servers/boxes or automotive central units and down to Internet of Things applications such as basic cameras. And most of all, it will need an outstanding developer experience. The three Ps of AI at the edge are privacy, performance, and power efficiency, but a fourth P – developer productivity – is crucial to delivering them.

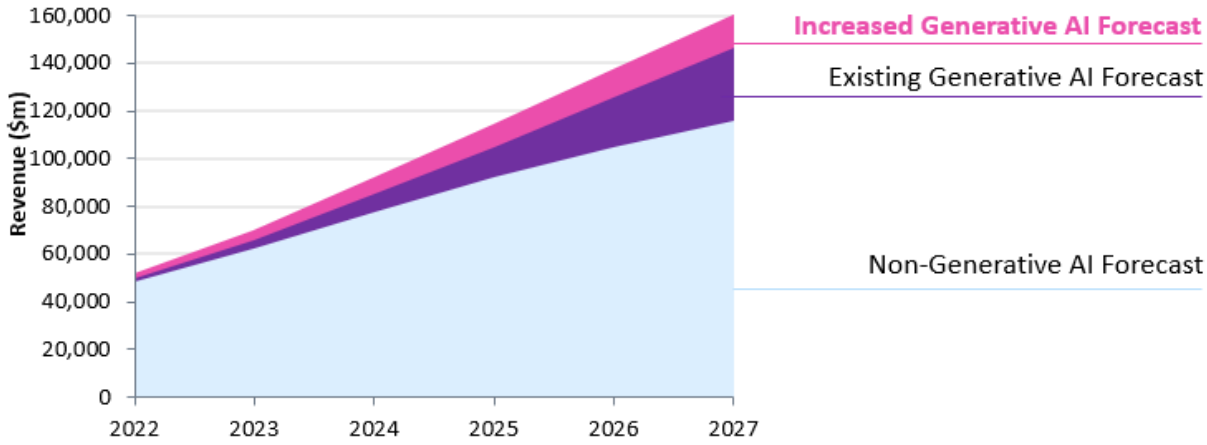
The Rise of Generative AI

From 2015's DreamDeeply to today's StableDiffusion, ChatGPT, or Mixtral, generative artificial intelligence models are perhaps the iconic technology of our time. Defined by the ability to create text, images, or other output ex-novo as a function of their training data and user prompts, these AIs have fascinated the world and sparked a surge of innovation and excitement, building on the key insight that models of natural language could encode very rich information about nearly anything, including features that are only implicit in the original training data.

Generative AI Is Changing The Game

Omdia's *Generative AI Software Market Forecast* estimates that the generative AI submarket is growing three times as fast as the broader AI opportunity and consequently driving faster growth as well as becoming a bigger share of the total. From around 9% of the market in 2023, we expect it will reach 27% by 2027, a $43bn opportunity of which $13bn is net-new and the rest ($31bn) represents a shift from older AI technologies towards generative AI.

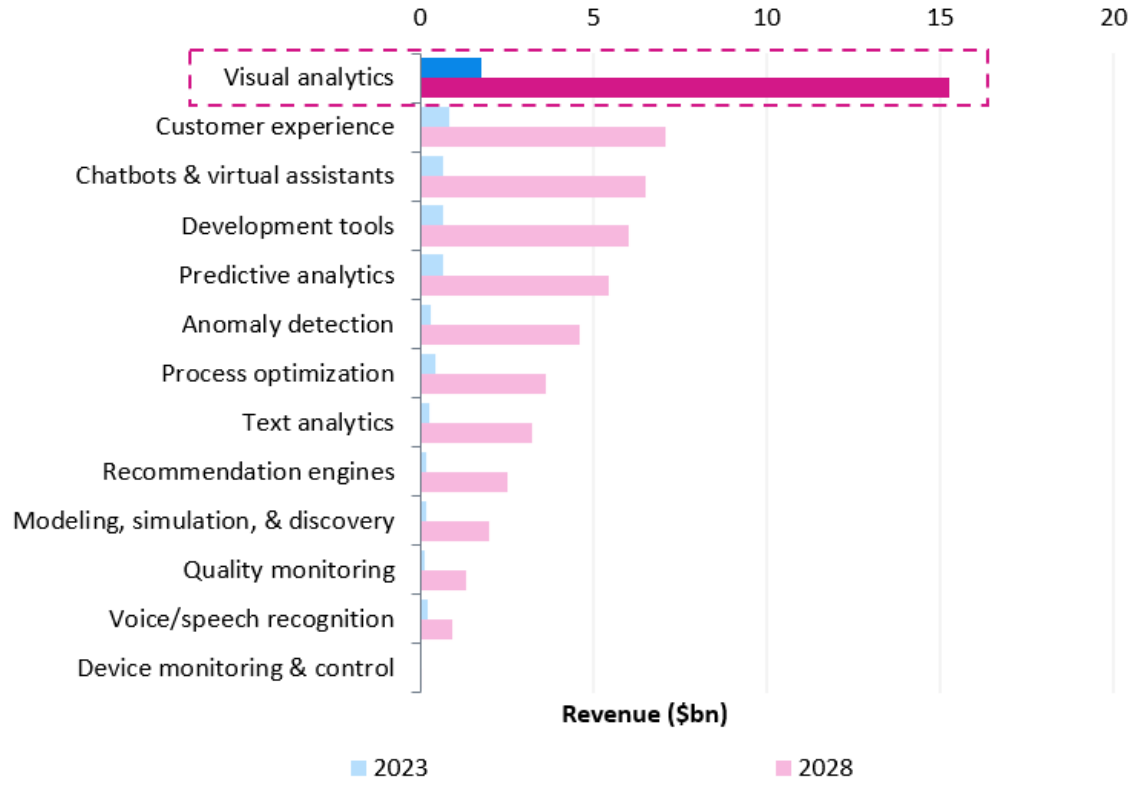
Figure 1: Generative AI is growing faster than the broader AI opportunity (GAI and AI software revenue forecast, world markets: 2022–27)



Source: Omdia AI Software Market Forecast 2023

Within this, the stand-out opportunity is the "Visual analytics" category, which Omdia expects to grow from $1.76bn in 2023 to $15bn in 2027, a CAGR of 54%. This is handily bigger than the next biggest opportunity ("Consumer experience") and includes a wide range of applications that essentially use a camera as a sensor, although industrial quality inspection and anomaly detection are covered separately. As such, this segment is unusually likely to have some processing at the edge or even further out into the Internet of Things domain.

Figure 2: "Visual analytics" is the top GAI opportunity
(AI software revenue by horizontal, world markets: 2023, 2028)



Source: Omdia AI Software Market Forecast 2023

Three Defining Factors Shaped Generative AI

Three crucial factors have shaped the era of generative AI so far:

Creativity:

- Generative AI is called that because it generates new content that wasn't explicitly present in the training data, giving at least the appearance of creativity.
- This has given the field enormous glamour and something that is historically precious for innovations – playfulness and fun.
- This creativity applies to input as well as output. Generative language models respond to free-text prompts, giving us a new and radically different paradigm for applications development. Rather than explicitly writing code, users craft natural-language prompts to shape the model's behavior. This can be interactive, like asking a chatbot a string of questions to refine its answers, or declarative, like giving it examples of answers and letting it figure out how to complete the rest.

Versatility:

- One of the most important and surprising characteristics of the large language models (LLMs) that underlie generative AI is so-called transfer learning, in which the model demonstrates skill on different tasks to the ones it was trained on. Few-shot transfer learning was an unexpected discovery but has become a canonical feature of LLMs.
- If creativity makes LLMs fun, transfer learning makes them useful. The same model can retrieve information, organize it in a stated format, display it to users, and summarize their conversation about it – and it may be able to do this across multiple languages.
- Transfer learning enables us to create complex applications across multiple domains through the prompt, making generative AI a true general-purpose technology.

Scale:

- The story of the LLMs has so far been marked by the so-called "bitter lesson" that more data, bigger models, and more training compute tend to beat more sophisticated modelling.
- Synergy between datasets, models, and hardware has driven a race for scale since the 2000s. Assembling very large data sets from the whole Internet made it possible to train large models, large models benefited from more training data, and big compute clusters were necessary to train the models. The big models became the pace-setting workload for the semiconductor industry, driving the creation of bigger GPUs and clusters, and making even bigger models possible.
- The journey from the 2017 paper that introduced the Transformer architecture to 2023's GPT-4 has primarily been a question of scale, building bigger models, bigger data sets, and bigger compute clusters of bigger GPUs. The versatility and creativity that characterize generative AI seem to be emergent properties that kick in once enough scale is achieved.

Moving Beyond Giant Models: The Rise of "Small Large Language Models"

Because scale has been so central to the generative AI story so far, we tend to think about giant clusters in hyperscale data centers, boosted by flagship GPUs or exotic wafer-scale accelerators. Historically, though, digital innovations have tended to start centralized and move edgewards.

Computing in general began by being organized around mainframe systems before minicomputers, PCs, smartphones, and Internet of Things devices led progressive waves of decentralization. The development of cloud computing looks superficially like a reversal of this trend, but under the surface, cloud infrastructure is made up of globally distributed data centers that themselves contain thousands of modular servers linked by deliberately decentralized

Historically, though, digital innovations have tended to start centralized and move edgewards.

network designs, and much of the point of the infrastructure is to support decentralized, distributed applications. It might be more accurate to say that these systems display centralized command with decentralized execution.

Something similar is happening with LLMs and generative AI. If ChatGPT rang in 2023, it was almost immediately upstaged by the leak of LLaMa, a multi-purpose, midsize LLM developed by Meta whose weights and biases were published on the Internet by an anonymous hacker. GPT-4 is thought to have around 1 trillion parameters, although not all of those may be active on each inference, while GPT-3 has 175 billion. The LLaMa family ranges between 7 billion and 70 billion, and it soon became clear that it could be customized to achieve comparable or even better performance than the GPTs on specific tasks.

LLaMa was the first "small LLM", but it would definitely not be the last. Meta's choice to release the successor model, LLaMa-2, as open-source software kicked off a year of unparalleled innovation in AI, as a vast variety of new models appeared from the open-source community. Many of these projects focused on delivering performance that matched or beat the GPT-3.5 model behind ChatGPT while keeping the model small enough to run locally on a PC. Even more importantly, some of them achieved major advances in the technique of training and fine-tuning generative AIs, such as May 2023's QLORA, opening up AI development to teams without the resources to build giant clusters.

By breaking through the barriers the race for scale created, these developments will bring the enormous versatility and flexibility of generative AI to the edge.

The big players got involved, too; in December 2023, Google announced its Gemini model family, which includes both their biggest model yet, Gemini Ultra, and also Gemini Nano, a version trained by distillation learning from Ultra that runs on some Pixel smartphones. Apple open-sourced the MLX framework for AI projects on its Apple Silicon SoCs. It's also worth remembering here that Stability AI's iconic Stable Diffusion image generator was a pioneer of scaling down generative AI – the model requires 5GB of GPU RAM to run and consequently fits on any Apple Silicon Mac, while Qualcomm has demonstrated it running on a smartphone. In December, Stability founder Emad Mostaque predicted on his X feed that an open-source project would match or better GPT-4 on an edge computing platform some time in 2024. In January 2024, Stability AI launched StableLM 2, a language model in only 1.6B parameters – slightly bigger than the original StableDiffusion.

By breaking through the barriers the race for scale created, these developments will bring the enormous versatility and flexibility of generative AI to the edge.

Real Applications Are At The Edge, And Small LLMs Will Deliver Them

Table 1 shows examples of some edge applications that are more likely to produce real revenue than an amusing chatbot and that are likely to be enhanced with generative AI in the next 18 months – for example with Large Language and Vision Assistant (LLaVa) models.

Table 1: Generative AI applications at the edge

Application	Media	Typical Model	Future Model	Use Case	Compute Demand
Smart City "AI Box"	Video	YOLO, ResNet-50	LLaVa, Vision Transformer	Analyze video with natural language queries	Large increase
Home Security	Video	MobileNet	LLaVa	Define alerts in natural language	Large increase
Industrial Robot	Video, Image	YOLO	DreamerV3, Vision Transformer	Imitation learning for CoBots working with humans	Large increase
Developer Kit	Text/Code	BERT	Code-LLaMa	Rapid prototyping of AI applications	Small increase
Medical Instrument	Image, Tabular	U-Net	LLaVa	Medical assistant using data beyond just imagery	Moderate increase

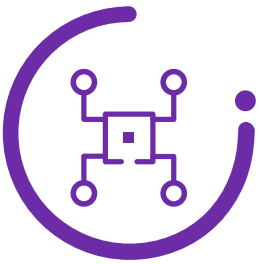
Source: Omdia

These applications have usually had basic AI features – for example, object detection in a security camera, or an industrial robot that can check if the right part numbers have been packed for shipment – but now, the arrival of generative and specifically multi-modal systems such as LLaVa, an extension of LLaMa that can work with images and text concurrently, is going to change them.

For example, it will be possible to write free-text queries over the video feed coming into a smart city application, relying on the multi-modal AI's ability to segment incoming scenes and encode information about the relationship between objects in them, or generate diagnoses or candidate drugs based on both medical imaging and tabular data. Creativity in what they generate as output, and flexibility in what they accept as input, will change these applications dramatically.

All the use cases in Table 1 have something in common – as well as being ones that are likely to migrate from older AI or classical computer vision solutions to generative and multi-modal AI, they are all ones that run locally on cameras, robots, industrial PCs, or medical instruments. This brings with it both opportunities – for snappy low-latency response and resilient, scalable distributed operations - and also constraints. Operating at the edge is always especially challenging in terms of price, power, and area, and as we will see, this is especially difficult for AI workloads because of our third key factor, scale.

The Hardware Impact



Bringing generative and multi-modal AI to the edge will mean changing the hardware that currently operates there. In Table 1, Omdia has tried to give some idea of how the computing requirement for each of these use cases is likely to change.

The developer workstation use is already being served by GPUs drawn from the intersection of entry-level server and top-end PC hardware, which are quite capable of running and even fine-tuning models up to around 34 billion parameters, so this is unlikely to change much although they will become more common. Similarly, most medical image analysis solutions that use AI have adopted powerful GPUs, although not all.

All the others, though, are currently either CPU-only, using video-specific ASIC or FPGA solutions, or using the lowest tier of AI acceleration. As a result, they will have to add more compute, moving up towards the performance levels we are seeing in current premium smartphones and top-of-the-line surveillance cameras, 30-50 TOPS, or higher in the case of devices like the "AI Box", usually an industrial PC supporting a number of cameras for smart city, surveillance, or industrial automation applications, where either a more powerful accelerator or multiple smaller ones will be necessary.

They will also have to do this while keeping power consumption down, close to the levels mobile devices achieve, and keeping the cost of the processor from wrecking the bill of materials. This is going to be difficult, as nowhere has the race for scale had more impact than in AI accelerator hardware.

The Hardware Lottery

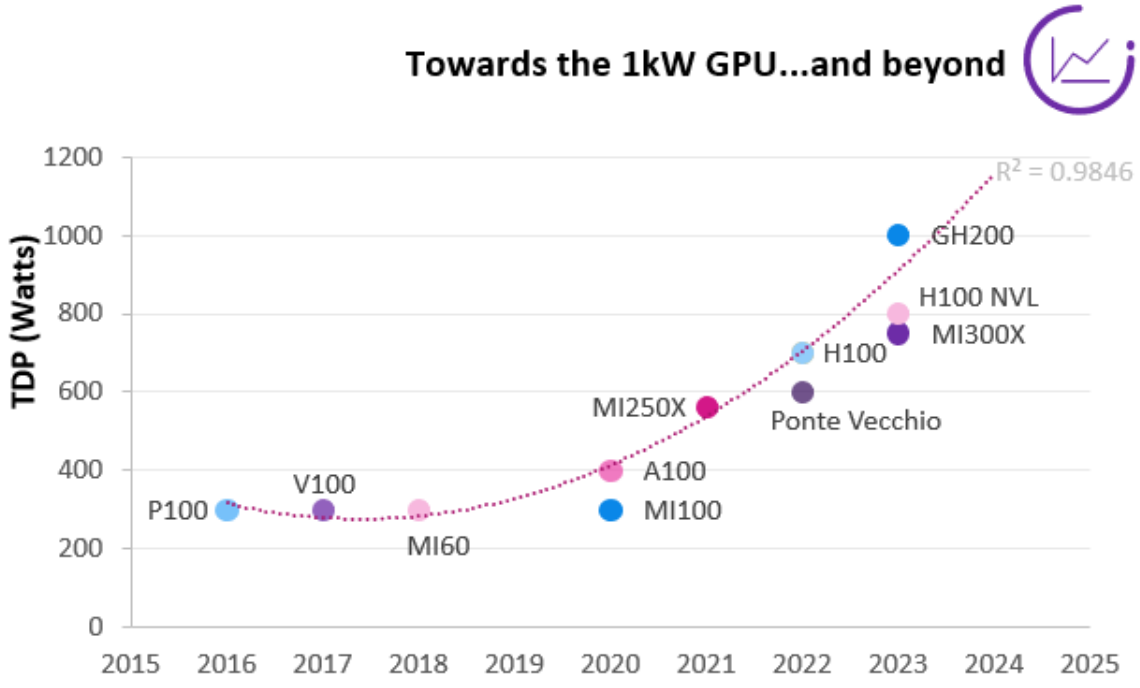
Sara Hooker, now of Cohere for AI and previously of Google Brain, wrote in a 2021 paper that fundamental AI research tends to be shaped by whatever the fastest computing solution available happens to be. One line of research wins the hardware lottery, turning out to work particularly well with the current best solution, and drags the rest of the field along with it.

The combination of deep backpropagating neural networks with GPUs has been a massive case in point; it wasn't anything the AI researchers were planning, nor was it a major goal of the chip designers, but it turned out to be a marriage made in heaven. As the Chinese GPU startup MooreThreads' name suggests, when Moore's law seemed to be struggling, the move to massively parallel GPU computing gave us more threads. As a result, AI has become nearly synonymous with GPUs and specifically, NVIDIA's GPUs, thanks to their long-term investment in the CUDA software development kit.

They will also have to do this while keeping power consumption down, close to the levels mobile devices achieve, and keeping the cost of the processor from wrecking the bill of materials. This is going to be difficult, as nowhere has the race for scale had more impact than in AI accelerator hardware.

The flipside of this effect is that after the GPUs shaped AI, AI shaped the GPUs. The dash for ever bigger models demanded ever bigger GPUs, as well as server configurations with higher GPU:CPU ratios and bigger clusters of machines. AI model training, especially LLM training, has become the pacing workload for the entire industry, and the industry rapidly adapted to deliver what the customer wanted. GPUs, and other AI accelerators, ran at around a thermal design power of 300W up to 2020; since then, it's taken them three years to break the kilowatt, while current market leading server CPUs are pushing into power consumption territory that used to be the domain of the GPUs.

Figure 3: The surge in accelerator power consumption

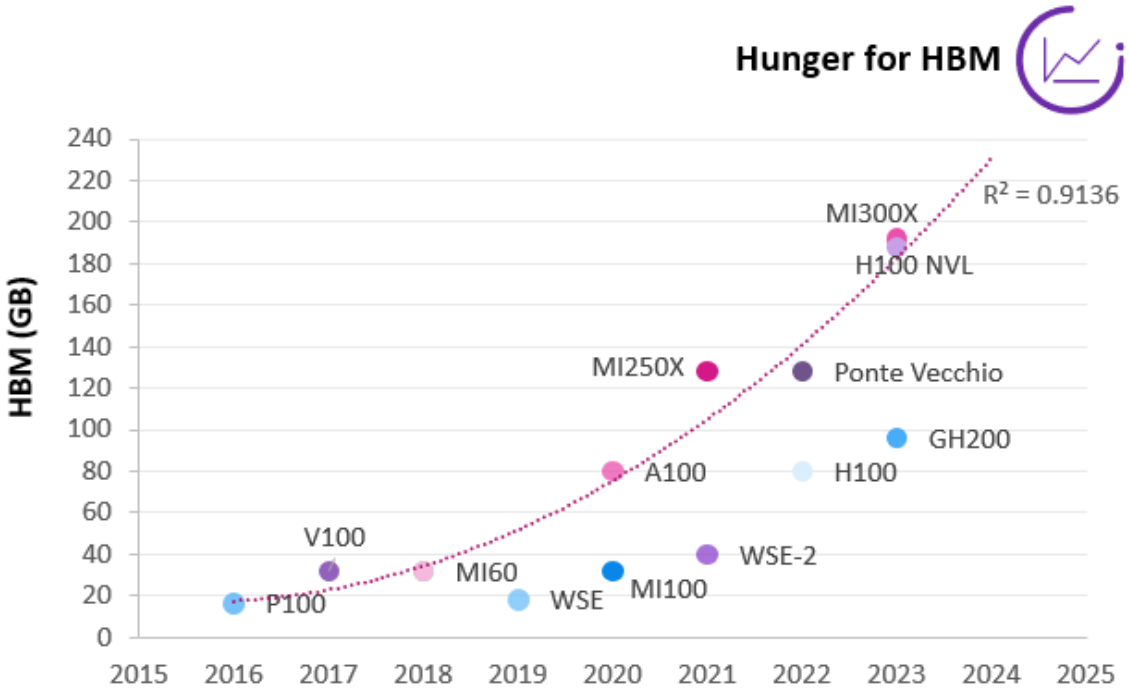


Notes: Most power-dense option (usually SXM) shown. MI300X, H100NVL, and GH200 are multichip modules
Source: Omdia

Compute was the easy bit…

The problem was not really the computing as such. Input-output and memory operations are more costly in energy terms than additional TOPS, and there is little point in having more computing power if it has to idle waiting for data to be loaded. It has been said that a modern GPU is a device that transforms compute-bound problems into I/O-bound problems. As such, the key driver of the increase in accelerator size and power consumption was memory, specifically on-chip high bandwidth memory (HBM). This created a number of downstream problems – for a start, the surge in power consumption tracks the increasing HBM content closely.

Figure 4: GPUs and AI ASICs' HBM content has dramatically increased to support bigger models



Notes: Most RAM-dense option shown. H100NVL and GH200 are multichip modules
Source: Omdia

This drove AI progress forward – if only down the one street of building bigger neural networks – but built in a crucial dependency on advanced packaging processes such as TSMC's CoWoS (Chip on Wafer on Substrate) and SoIC (System on Integrated Chip) to integrate the HBM stacks. It also creates a problem for latency-sensitive applications, in that creating large batches of inference work means that the latency is dependent on where your request falls in the batch. If you are the last token in the batch, you are processed immediately, while if you are the first, you wait until the batch fills, generating jitter.

Most of all, it drove up the power bill. Typical AI server configurations, with 8 flagship GPUs or other accelerators and 2 top-performance CPUs, are now literally as hot as a restaurant-grade gas stove with all the burners lit up. Most of this power consumption is used either to shuttle data to and from memory, or to keep the memory itself alive.

One way out of the bind: customization and specialization

Microelectronics has historically advanced through three major strategies. The first, Moore's law, barely needs introduction. The second, More-than-Moore, revolves around higher levels of integration, building more functions into the die and eliminating off-chip connections. A relevant example is the move to chiplet designs for flagship CPUs and GPUs, or the increasingly common

architecture derived from smartphones that creates a mix of CPU, GPU, and AI-ASIC (sometimes "NPU") cores in a system-on-chip.

The third, perhaps less well known, is Makimoto's wave, named for the former Sony CTO who described it in 1991. Tsugo Makimoto observed that the industry tended to cycle between specialized and standardized products; at the technology frontier, it becomes necessary to optimize whole systems rather than just chips, and consequently to develop custom products, while once the other two strategies catch up, it becomes more economic to use standard products, gaining economies of scale and making disaggregation and competition possible.

AI is definitely the workload that defines the technology frontier today, so it's not surprising in this framework that since around 2019, the surf has been up. The hyperscale cloud providers have all developed custom AI ASICs, as well as custom CPUs on Amazon Web Services' part. Apple has progressively moved its entire product line onto its own silicon, which prominently features both AI-related extensions to the CPU instruction set and a hard ASIC accelerator. IBM created an AI ASIC for its mainframe processors and then re-used the design on a PCIe card as a general-purpose AI inference accelerator. In China, Baidu and Huawei have both developed their own AI silicon, while Alibaba has in-house CPU designs. In automotive, Tesla Motors has replaced NVIDIA GPUs in its vehicles with in-house ASICs and moved on to create an AI training system based on the same design, while several other automakers are openly experimenting. Microsoft was a holdout, preferring to work with FPGAs, until the autumn of 2023 when it joined the push for customization.

The whole purpose of training an AI model is to run inference against it as part of some application. Training is a development task while inference is production.

Interestingly, the wave is also washing through the incumbent. At the outset of 2022, NVIDIA engineers issued a paper arguing that future GPUs must perforce become multi-chip modules that would be "composable" in order to permit customization into graphical, AI, or HPC applications. The company suited the action to the word. In the Ampere microarchitecture, it introduced a distinction between "graphical" and "AI" GPUs and then began to add dedicated silicon features to accelerate Transformer models. In the Hopper architecture, it transitioned to multi-chip. In the latest, Lovelace architecture, though, NVIDIA appears to have cycled back to standardization, marketing its L4 and L40S as "universal" GPUs, although it's a reasonable guess that the next flagship B100 will be a dedicated AI part. Faced with a constraint of a different kind, the US sanctions on China, the company has again chosen to make custom devices.

Another sign of the wave is the explosion of AI chip startups, who by definition are creating new AI-specific chip architectures. Between 2019 and 2022, the top 25 startups by funding raised over $6bn in venture capital, the great majority of it to develop various takes on coarse-grained reconfigurable architecture (CGRA) accelerators. Here, though, the wave is crashing into a mighty barrier; the lack of strong software support and the challenge of persuading developers to devote their time to learning a new architecture when the existing one offers a unique range of opportunities.

Three Key Challenges

Any future AI hardware solution has to address three key challenges that arise from the issues we've just discussed – the three Ds.

- **Delivering inference at scale –** The whole purpose of training an AI model is to run inference against it as part of some application. Training is a development task while inference is production. Depending on the application, the model may be inferenced anywhere from a few times to millions of times a day, but there will always be some multiplier from training to inference, and the increasingly interactive and iterative nature of applications such as Microsoft's GitHub Copilot tends to drive up inference usage. As a result, it is on the inference side where the power consumption issue will really bite. A twist on this is that once AI goes into production, model inference will often be on a critical path for the application's overall performance, so both throughput in tokens or samples/second and latency will count.

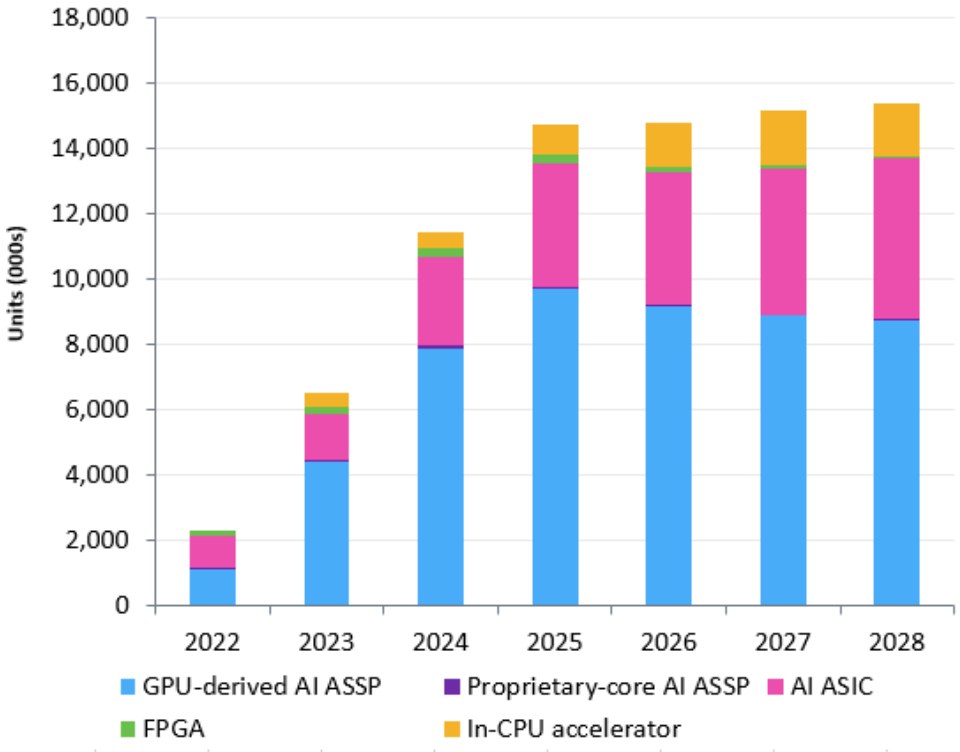
- **Democratizing AI model development –** Like any innovation, the ability to adapt AI to users' purposes is crucial to its adoption. Further, the ability for a much wider variety of actors to develop or at least fine-tune their own generative AIs is a crucial answer to many of the field's regulatory issues. And as we saw with Makimoto's wave, specialization is an important strategy to make generative AI more tractable. The rise of the open-source models in 2023 has rendered it much more important that future hardware supports efficient fine-tuning.

- **Developers –** The entire digital industry has always relied critically on applications developers to adapt general-purpose technologies to specific tasks. Nobody wants an AI model, rather, they want to run inference against it, and they want to do so to make some application do useful work. Without applications, the entire enterprise is useless. This principle rarely fails – it would have served you well with regard to IBM System/360 in the 1970s, the proprietary Unix systems or Apple Macintoshes of the 1980s, the Windows/Intel PCs of the 1990s and 2000s, the Linux servers of the 2000s, or the iOS and Android mobile devices of the 2010s. NVIDIA's success in AI has been largely down to its consistent investment in developer tools and software; any competitor must answer how it can serve the developer.

The years ahead

Omdia's *AI Processors for Cloud and the Data Center* forecast estimates that we're in for another 18 months to two years of wild boom conditions for data center AI accelerators before an inflection point is reached. After this point is reached, Omdia expects growth to slow down markedly as we move from a deployment phase dominated by technology adoption as a buyer motivation and AI model training as a use case into a steady-state phase where inference and fine-tuning will become progressively more important.

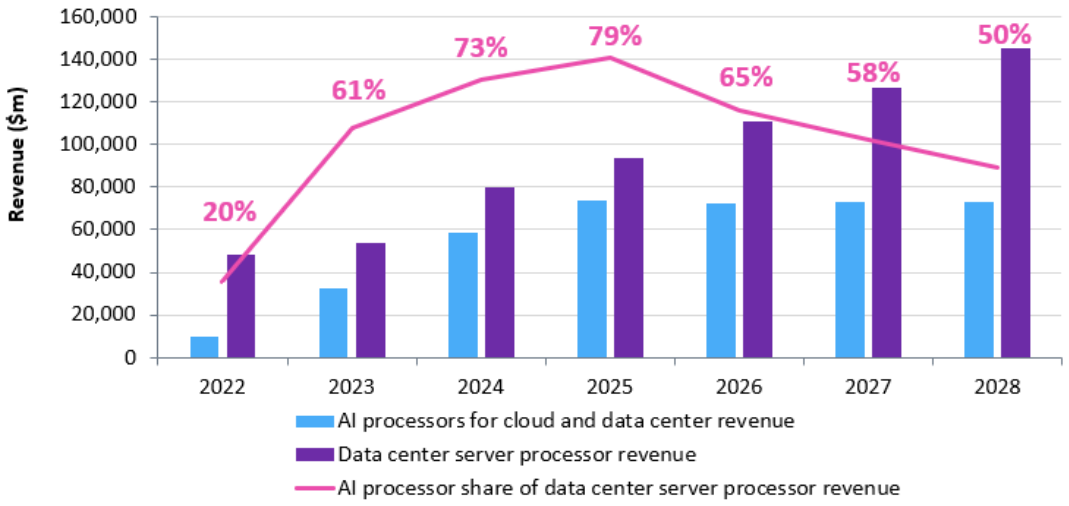
Figure 4: AI processor volume for cloud & data center, forecast, 2022-2028



Source: AI Processors for Cloud & Data Center 2023

That said, the industry will not go back to how it was before the AI boom, which will be transitory in the sense that a revolution is transitory. Omdia expects AI acceleration to rise from 20% of data center processor revenue in 2022, or $9.7bn, to 79% at the peak in 2025, and then fall back to around 50%, around $73bn; accelerated computing in general and AI in particular will be a much bigger share of the industry than it historically has been.

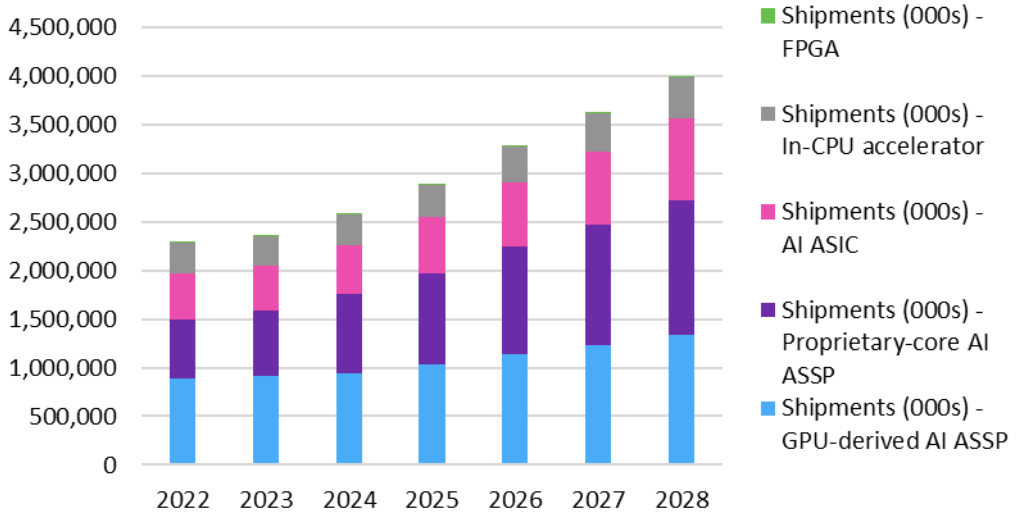
Figure 5: AI processor revenue for cloud & data center as % of total data center processors



Source: AI Processors for Cloud & Data Center 2023

In parallel with this, AI acceleration is marching steadily through the edge computing space, defined in Omdia's terms as including everything within 20ms network roundtrip time of the user. Omdia expects this segment, which is mostly smartphones and PCs but also includes markets such as cameras, industrial machine vision, robotics, and automotive, to grow from around $31bn in 2022 to $60bn in 2028.

Figure 6: AI accelerators for the edge having specified accelerator types, forecast, 2022-2028



Source: AI Processors for the Edge 2023

The data center space is overwhelmingly dominated by GPUs, with Google's TPUs coming in next, while the edge remains very different, with GPUs struggling to break through and most growth coming from either proprietary AI-ASICs (such as Apple A-series) or merchant AI-ASSPs with integrated NPUs (such as Ambarella CV72, Ambarella N1, or the Qualcomm Snapdragon range). The relationship between the two is changing – the relatively rapid uptake of AI acceleration in smartphones (the attach rate is currently around 65%) means that the edge market is substantially bigger today, but the spectacular GPU boom will change that, although the combination of getting past the training roll-out and the arrival of AI PCs will limit how much bigger than edge the data center will get.

The data center space is overwhelmingly dominated by GPUs, with Google's TPUs coming in next, while the edge remains very different, with GPUs struggling to break through and most growth coming from either proprietary AI-ASICs (such as Apple A-series) or merchant AI-ASSPs with integrated NPUs (such as Ambarella CV72, Ambarella N1, or the Qualcomm Snapdragon range).

Future shape of the market beyond 2025

Figures 3 and 4 suggest the rush for flagship GPUs will be over relatively soon as the market saturates. Beyond the inflection point, however, AI will not go away. In fact, rather than a plateau as in the model output, it might be more likely to transition to linear growth, more like the edge scenario in Figures 5 and 6. Once the installation phase is over, the future demand for AI processors is likely to be a function of the following variables. On the training side these are:

- **Adoption of new models within existing systems** – enterprises that have already deployed an AI training capability can both train new models when new technology appears, and train more instances of existing ones to address more use cases within the enterprise.

- **Frequency of retraining** – even if the same model remains in use, it is likely to generate some training demand through the practice of periodically retraining to incorporate new data and tuning to adjust for new features or to respond to observations from model monitoring as part of the MLOps process. Retraining is a multiplier for training demand in general.

On the inference side there is of course:

- **Demand for inference** – the more users AI applications have, and the more intensively they use the application, the more processors will be required to scale out the services that provide AI inference to the applications. There is a complex engineering judgment to make between using techniques such as multi-GPU virtualization to split up flagship processors and using more, smaller accelerators. In general, it is good practice to scale out with the smallest possible increments of capacity rather than scale up with bigger monolithic systems.

All three are subject to the multiplier of model size, which is consequently very important:

- **Model size growth –** evidently, bigger models demand more compute and more memory both in training and inference. Future developments in fundamental AI research will determine what the typical requirement for on-chip memory will be, which will in its turn influence power consumption and price.

What can we say about this, here in 2023? For a start, the trajectory of model size growth appears to be very different from what might have been predicted in 2022. The success of the open-source AI models has filled in what used to be called the "missing middle" between 5 billion and 50 billion model parameters, while the model platform HuggingFace reports that the most heavily-used size band is that between 500 million and 7 billion parameters. The big change in 2023 was that the open-source projects showed it was possible to match the giants, or even beat them on domain-specific applications, with a much less prodigal use of computing resources.

Crucial to this was the development of better techniques for fine-tuning and otherwise refining existing models against task-specific data. Probably the most important paper of the year, *QLORA: Efficient Finetuning of Quantized LLMs*, demonstrated that it was possible to effectively fine-tune a LLaMa-derived model against tens of thousands of new examples within a few hours on a flagship GPU or within a day on a good desktop GPU. This has itself already brought about an impressive wave of new models, notably the multi-modal ones that operate in multiple media, and at least one startup, Lamini, which promises to create proprietary models based on enterprises' data as a service. The future of AI models, looking forward from 2023, is specialized, multi-modal, and fine-tuned on the user's own data.

The future of AI models, looking forward from 2023, is specialized, multi-modal, and fine-tuned on the user's own data.

At the same time, inference-grade accelerators have been growing, both to cope with bigger models' inference demands and also to grow into the role of fine-tuning or training smaller ones. NVIDIA's L40S is not that far behind the A100's performance parameters. Both operational efficiency and responsible AI practices, meanwhile, require the ability to do regular tuning. Omdia therefore expects that over time, the inference/training distinction will become less salient.

AI inference, and even fine-tuning if not training in the true sense, is also moving edgewards to some extent. 2024 is expected to see the arrival of credible AI accelerators in the PC market, with products such as Intel Core Ultra (aka "Meteor Lake"), AMD Ryzen 8040, and Qualcomm Snapdragon Compute Elite X joining Apple's well proven M-series SoCs. A vigorous community of local AI developers and enthusiasts has sprouted around the LLaMa family of models and either Apple machines or gaming/workstation PCs with NVIDIA RTX-series GPUs. Interestingly, the Intel, AMD, Apple, and Qualcomm products all display the CPU/GPU/NPU architecture now familiar from smartphones.

"Small LLMs" will drive this development forwards. Replacing models such as YOLO or ResNet (or indeed no AI at all) with ones in the 1-10B range means a substantial new requirement for compute at the edge, and especially memory and memory bandwidth, as Transformer-type models' limiting factor in inference is memory bandwidth. Having a unified memory architecture across the CPU/GPU/NPU domains will be an important advantage, and we may be seeing this already with the Apple Silicon Macs.

Ambarella: An Edge AI Solution

Ambarella's vision and edge AI system-on-chip products have been adding machine sensing to cameras in a wide range of industries for more than two decades, now – including consumer, video security, safety-critical automotive, robotics, and industrial applications. The company began as a provider of video SoCs before introducing computer vision and machine learning, and is currently on its 3rd generation of AI products, with an installed base of 20+ million inference processors deployed. As such, there is a significant base of applications and developers currently active for its products.

Ambarella AI SoCs incorporate the company's proprietary NPU, the CVflow® AI inference engine, for efficient machine learning inference against Transformer architecture neural networks. These SoCs adopt the now-classic AI processor format, incorporating ARM-architecture CPUs, integrated in a single chip with its proprietary NPUs and several additional application-specific compute engines, something the company has substantial expertise in. For example, Ambarella builds on its video-processing heritage by integrating the latest generation of its proprietary advanced image signal processor (ISP). This is particularly useful for multi-modal LLM applications, for example real-time text based analysis of camera feeds, wherein the ISP processes all the captured camera data over a wide dynamic range and then passes it to the SoC's CVflow AI engines for analysis by the model.

As SoCs with high performance host CPUs and other specialized compute engines, Ambarella's devices are complete solutions for the edge device rather than add-on AI accelerators; and as edge-first devices designed from the beginning to operate far from the support of a data center, they are optimized for power efficiency. In the latest Ambarella N1 SoC, the integrated, third-generation CVflow NPU dominates the total TOPS available across all of these integrated processor types.

On the latest Ambarella N1, key multi-modal LLMs such as LISA and LLaVA run with 50W, while star open-source AI models such as Code-LLaMa 34B run with low enough latency to be an effective coding co-pilot with well under 50W of power. This makes the N1 a contender to provide the boost to computing performance on the edge that's necessary to add the creativity and flexibility of generative AI to the key edge devices laid out in Table 1.

In addition to the N1 itself, Ambarella launched its Cooper™ developer platform during CES, which includes a comprehensive stack of software tools starting with the Linux-based OS, compiler, SDKs and profiler, and working up through implementations of the top 10 edge AI models to include sample solutions for common sensor-based use cases.

Appendix

About Ambarella

Ambarella's products are used in a wide variety of human vision and edge AI applications, including video security, advanced driver assistance systems (ADAS), electronic mirror, drive recorder, driver/cabin monitoring, autonomous driving and robotics applications. Ambarella's low-power systems-on-chip (SoCs) offer high-resolution video compression, advanced image and radar processing, and powerful deep neural network processing to enable intelligent perception, fusion and planning. For more information, please visit http://www.ambarella.com.

Further reading

www.ambarella.com/cooper

Author

Alexander Harrowell

Principal Analyst, Advanced Computing for AI

Applied Intelligence

Alexander.Harrowell@Omdia.com

Get in touch

www.omdia.com
askananalyst@omdia.com

Omdia consulting

Omdia is a market-leading data, research, and consulting business focused on helping digital service providers, technology companies, and enterprise decision-makers thrive in the connected digital economy. Through our global base of analysts, we offer expert analysis and strategic insight across the IT, telecoms, and media industries.

We create business advantage for our customers by providing actionable insight to support business planning, product development, and go-to-market initiatives.

Our unique combination of authoritative data, market analysis, and vertical industry expertise is designed to empower decision-making, helping our clients profit from new technologies and capitalize on evolving business models.

Omdia is part of Informa Tech, a B2B information services business serving the technology, media, and telecoms sector. The Informa group is listed on the London Stock Exchange.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help your company identify future trends and opportunities.

Copyright notice and disclaimer

The Omdia research, data and information referenced herein (the “Omdia Materials”) are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together “Informa Tech”) or its third party data providers and represent data, research, opinions, or viewpoints published by Informa Tech, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees, agents, and third party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.