OMDIA MARKET RADAR

Publication date:
23 Apr 2025

Author(s):
Alexander Harrowell, Principal Analyst, Advanced Computing for AI

Omdia Market Radar: AI Processors for the Edge 2024

OMDIA

Summary

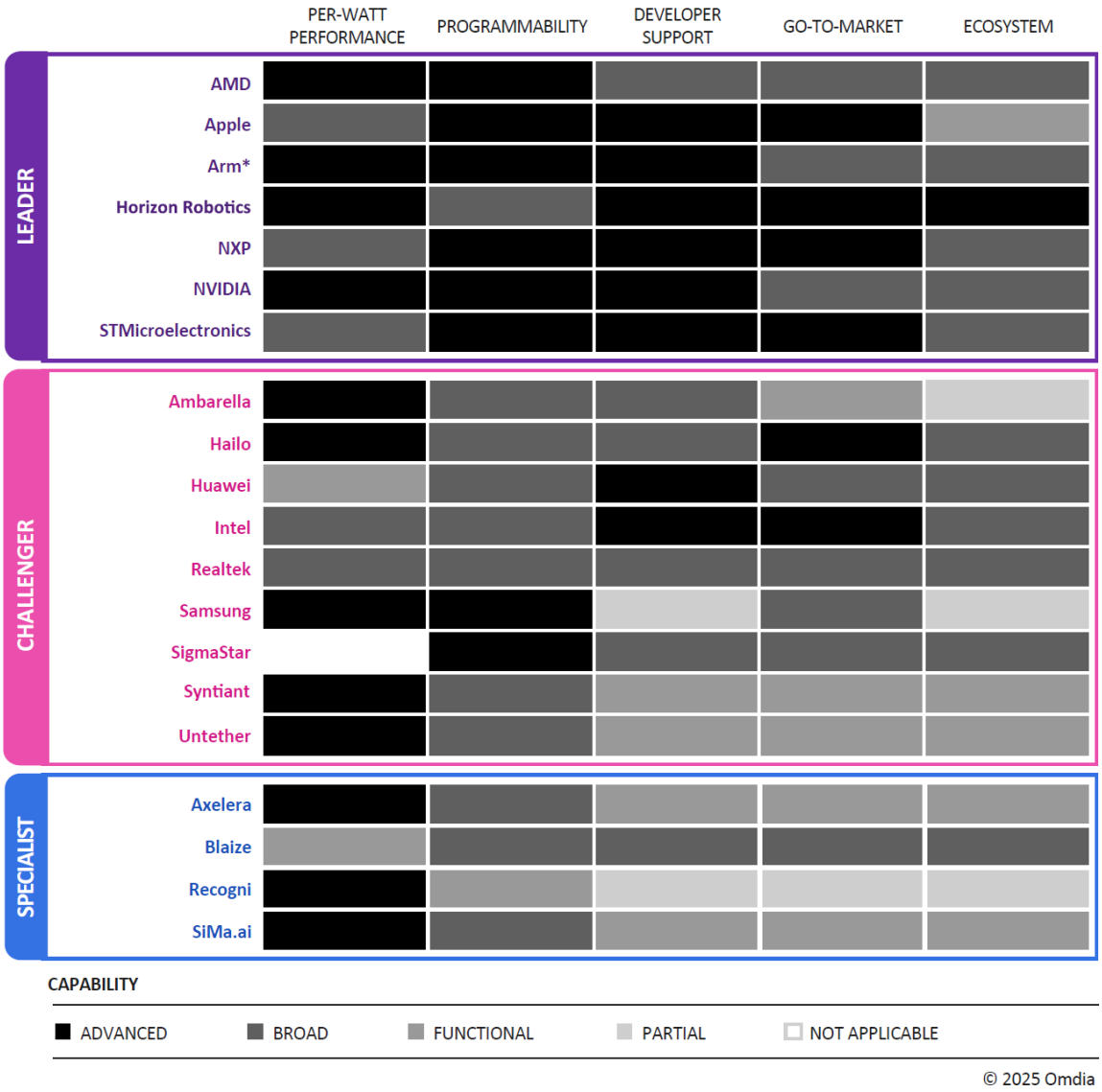
Catalyst

Omdia projected at the end of 2024 that the market for artificial intelligence (AI) acceleration hardware at the edge, defined to include all compute above the microcontroller class and within 20ms network round-trip time from the user, would grow from $43bn at year-end to $89.7bn by 2029. The market is being driven by the rapid adoption of AI and the clash between expanding AI model sizes and the price, power, and area constraints of edge computing, as well as the development of a new category of AI PCs. Omdia further forecasts a dramatic shift in emphasis away from GPUs as the primary accelerator chip type and toward a more diverse ecosystem of different architectures that include not only custom application-specific integrated circuits (ASICs) and field programmable gate arrays (FPGAs) but, most of all, application-specific standard products (ASSPs) such as Qualcomm's Snapdragons and Intel's Meteor Lake and Panther Lake CPUs. This Omdia Market Radar compares the market's key players and their products. It aims to guide enterprises and other AI implementers in choosing AI accelerators at the edge, as well as to help technology vendors in shaping their strategies and products.

Market snapshot

Figure 1 illustrates the solutions that Omdia investigated as part of this research, in addition to highlighting the capability categories that Omdia analyzed. The definitions, assessment process, and vendor information are described in more detail later in this report.

Figure 1: Omdia heatmap for edge AI processors



Source: Omdia * Note that business model change is a special risk for Arm.

Key messages

- **Technology from the mobile space has taken over.** The new generation of AI PCs is overwhelmingly built around large heterogeneous systems-on-chips or systems-on-packages incorporating CPU, GPU, and neural processing unit (NPU) cores, NPU being the emerging term for AI ASICs/ASSPs, just as you would find in a smartphone. Apple has led the way, but both Intel and AMD have adopted this trend, and Qualcomm and MediaTek are historic mobile vendors. With Intel's plans to reuse Core Ultra CPUs in other edge formats, this approach is now radiating back from the PC toward the far edge.
- **The key driver of this trend is performance per watt.** In mobile and embedded applications, it is impossible to overstate the importance of energy, thermal efficiency, and on-chip integration. Yet, the focus on these features has some counterintuitive consequences. Although there is

usually an efficiency boost when moving from a GPU to an NPU, GPUs have emerged as the primary developer platform for AI PCs. But for long-running tasks, NPUs provide offload to a less energy-hungry accelerator.

- **The key trade-off is between performance per watt and developer productivity.** Software remains the biggest challenge for any new entrant to this space. Many companies Omdia has covered have successfully taped out their chip but are still far from getting developers' buy-in because their software support and tools are sub-par. The edge situation is rather different than that of the data center because the x86 and Arm developer ecosystems are very deep, and the power/price/area constraints for truly niche devices are so severe that nothing like the dominance of NVIDIA's CUDA has come to be.
- **Central AI acceleration has won out over distributed.** When Omdia prepared the previous edition of this report, there was an open question as to whether smartphones and other devices would integrate accelerator cores into their separate functions (e.g., radio, audio, sensors) or pull them into a central AI hub. This issue has been resolved. There is now usually either a single large accelerator block or a high-low split with an AI engine and an additional couple of cores in a low power sensor hub.

Omdia view

The edge environment remains very different than that of the data center. A system-on-chip (SoC) or system-in-package approach is almost a requirement at the edge due to price/power/area constraints. The mix of cores included in the SoC is gradually converging from the different architectures: NVIDIA's Jetson line of GPU-based ASSPs includes dedicated silicon for specific AI tasks as well as its GPU Tensor cores; AMD's Xilinx Versal FPGAs include Xilinx's dedicated deep learning accelerator cores alongside the reconfigurable logic and Arm CPUs; and mobile SoCs from Apple, Qualcomm, Samsung, or MediaTek usually include and use both GPU cores and AI accelerator cores for AI workloads. New entrants, which include both startups and major vendors such as STMicroelectronics, are continuing to field their own AI accelerator designs, usually designated as NPUs or something similar. Nothing like NVIDIA's hegemony in the data center exists at the edge.

In general, everything is coming to look rather like a smartphone SoC, with the same being true even in some data center products. Amazon Web Services' (AWS's) infrastructure chief, senior vice president, and Distinguished Engineer James Hamilton said as long ago as 2009 that "what happens in mobile eventually happens in server." The success of the Arm architecture, the mobile vendors' expertise with SoC design, heterogeneous integration and packaging, and the development of the custom silicon ecosystem are all taking the industry in this direction.

Since the beginning of 2024, the integration of AI in the PC has changed the game. Although smartphones remain the biggest segment of this market, AI accelerator penetration is only growing slowly beyond 67%, and PCs are the major source of growth. The PC market is very large, and the PC form factor gives OEMs and developers a lot of scope. An interesting consequence of this has been to partly reverse the trend observed in 2023 to de-emphasize the GPU. GPUs' combination of programmability and power means GPUs are the primary AI developer platform for the PC, whereas the NPUs usually packaged with them are more of a power-saving optimization.

The degree to which the devices are supported by software and developer tools differs dramatically. As in the data center, one of the biggest competitive challenges is to convince developers to invest their time and effort in mastering your technology. Unlike in the data center space, where NVIDIA's software tools dominate, the edge is still much more open. Still, every serious participant is investing in their software tools, and being a serious participant is likely to be defined by the quality of these tools.

The mobile-first players have thoroughly battle-tested software development kits and substantial developer ecosystems, and both Intel and AMD are investing heavily in developer relations and software enhancements—but deeper into the market, moving closer toward the microcontroller class of devices or getting into more exotic technology, the variety of SDKs developers need to master increases; their quality gets more variable. That said, AI chip SDKs are commonly built on a relatively small and well-understood set of open source components in very active development, such as Apache TVM, ONNX, and MLIR. STMicroelectronics has gone so far as to package its tools as a Linux distribution, providing the entire stack needed to deploy an AI model into production inference.

Recommendations

Recommendations for enterprises

- **Take advantage of the emergence of AI PCs.** The adoption of AI acceleration in the PC means that edge AI chips will be manufactured on an even greater scale than they already were and that technology developed for PCs will spread into other edge contexts. This also means that the very strong developer ecosystem for PCs will be available for edge server/appliance/industrial PC contexts.
- **Think software first.** The edge ecosystem is much more diverse and competitive than that of the data center, with many different available hardware options. That very diversity threatens to make the software environment much more complicated. Although NVIDIA's hegemony with CUDA makes competition very difficult, it also (ironically) makes decisions about software development easy. Unless your application faces extreme performance demands, developer productivity will usually trump most other considerations.
- **Provide for future growth.** Paradoxically, AI research is discovering the possibilities of small Transformer-type models while computer vision models continue to grow. Although the size of general-purpose language models in the data center context has peaked and the excitement has moved on to the new class of small models, this may mean that the typical edge model is going to get bigger, as small large language models (small LLMs) get deployed to the edge instead of, say, YOLO or ResNet-50. As a result, there is value in having compute and memory available for future developments.

Recommendations for technology vendors

- **An NPU does not make an AI strategy.** Many, many vendors have discovered AI as a topic and responded by adding some form of NPU core to their products, especially in the far edge, Internet of Things (IoT), and microcontroller spaces. In quite a few cases, these represent third-party intellectual property. How much value they are adding is questionable, as the software

support is often lacking, the documentation is minimal, and the product is generic. If you are not offering something unique, it would be best to either skip it or go for a licensing option that developers are likely to have encountered and that has good software support. Arm's Ethos line would be an example or, alternatively, a GPU.

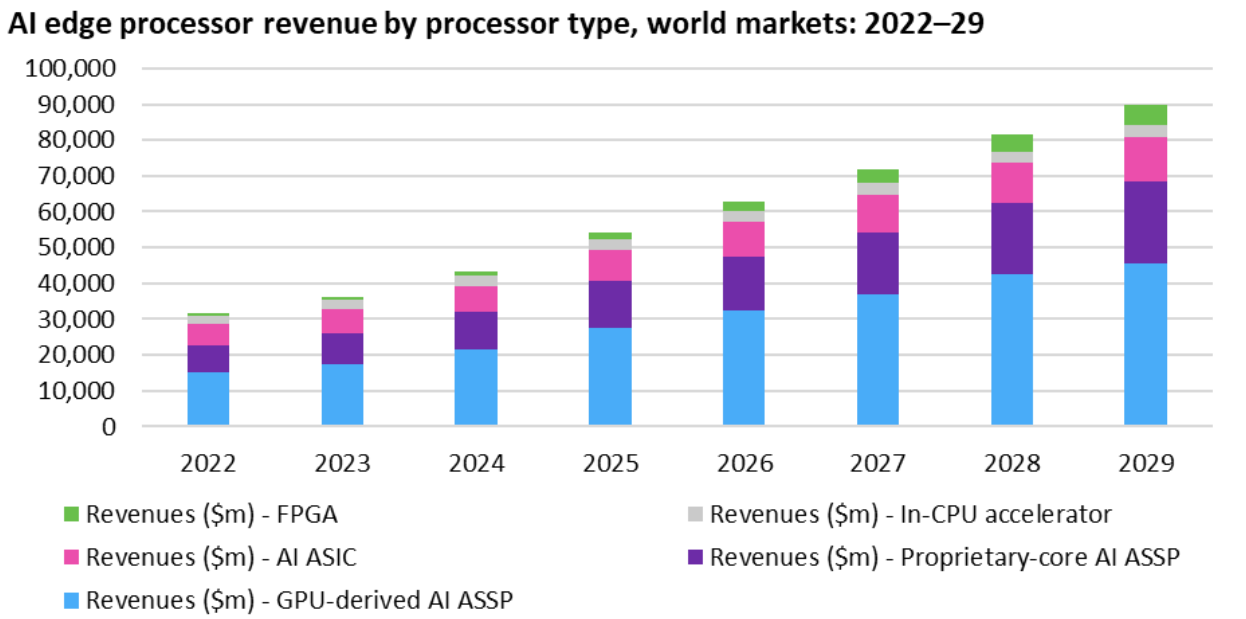
- **The diversity of edge applications, devices, and customers is proving difficult for custom silicon.** Custom AI ASICs and CPUs are a substantial force in the data center. Revenue passed through $11bn in 2024, with iconic projects such as Google's Cloud TPUs. Changes to intellectual property licensing, electronic design automation, and silicon packaging have made custom and semi-custom silicon a much more accessible option, but the minimum economic unit size of a custom project is still quite a barrier. That said, edge projects above a certain size tend to evolve toward custom chips; DJI's drones are an example.
- **Support for software developers remains crucial.** Developer issues are a key barrier to the adoption of any new hardware architecture.

Defining the edge AI processor market

Definition and characteristics

Omdia defines the edge as including all compute within 20ms network round-trip time of the user. For the purposes of this report, Omdia includes any products that fulfill this definition above the microcontroller/IoT product category. "AI processors" means processors that include dedicated hardware acceleration for AI/machine learning (ML) workloads, as long as it is implemented either as a discrete chip or as an identifiable functional block within a SoC, system-in-package (SiP), or system-on-module (SoM). **Figure 2** summarizes findings from Omdia's *AI Processors for the Edge Forecast Report – 2024 Database*, showing the market growing at a 16% CAGR from 2022 to 2029.

Figure 2: The edge AI silicon market



Source: Omdia

Technology

In practice, almost all AI workloads today are based on so-called deep learning, an ML methodology that uses multiple layers (hence "deep") of neural networks. The devices this report discusses are primarily accelerators for deep learning neural networks. Each "neuron" within the network is, in fact, a mathematical function that will "activate" and return output to the next neuron, depending on the inputs from the ones above it in the network. The neural network is trained by adjusting the constants—known as weights or biases—in these functions, depending on how well the network performs on its training data. To execute each layer in the network, whether in training or in inference, it is necessary to evaluate each of the functions before updating the next layer. As such, each iteration can be reduced to a matrix-multiply operation, and each evaluation of the whole network can be reduced to a succession of matrix-multiply and accumulate operations, plus some general-purpose compute operations at the end of each pass. Because model training is only rarely carried out at the edge, this report concentrates on inference, the process of generating output from the model.

Because the number of neurons and their parameters can be very large (more than one trillion, although networks of this size are not practically deployable at the edge), it is necessary to carry these operations out with the greatest possible degree of parallelism to get acceptable performance. As a result, GPUs have become the most common accelerators, having many thousands of parallel processing cores. The hardware this report covers is defined largely by the need for matrix-multiply-accumulate (MAC) units. This, however, is an oversimplification. Supercomputer pioneer Seymour Cray remarked in the early 1980s that anyone could design a fast chip, but the problem was designing a fast system. Stephen Jones, NVIDIA's chief GPU architect, has described the GPU in general-purpose computing as a device that transforms compute-bound workloads into input/output(I/O)-bound workloads.

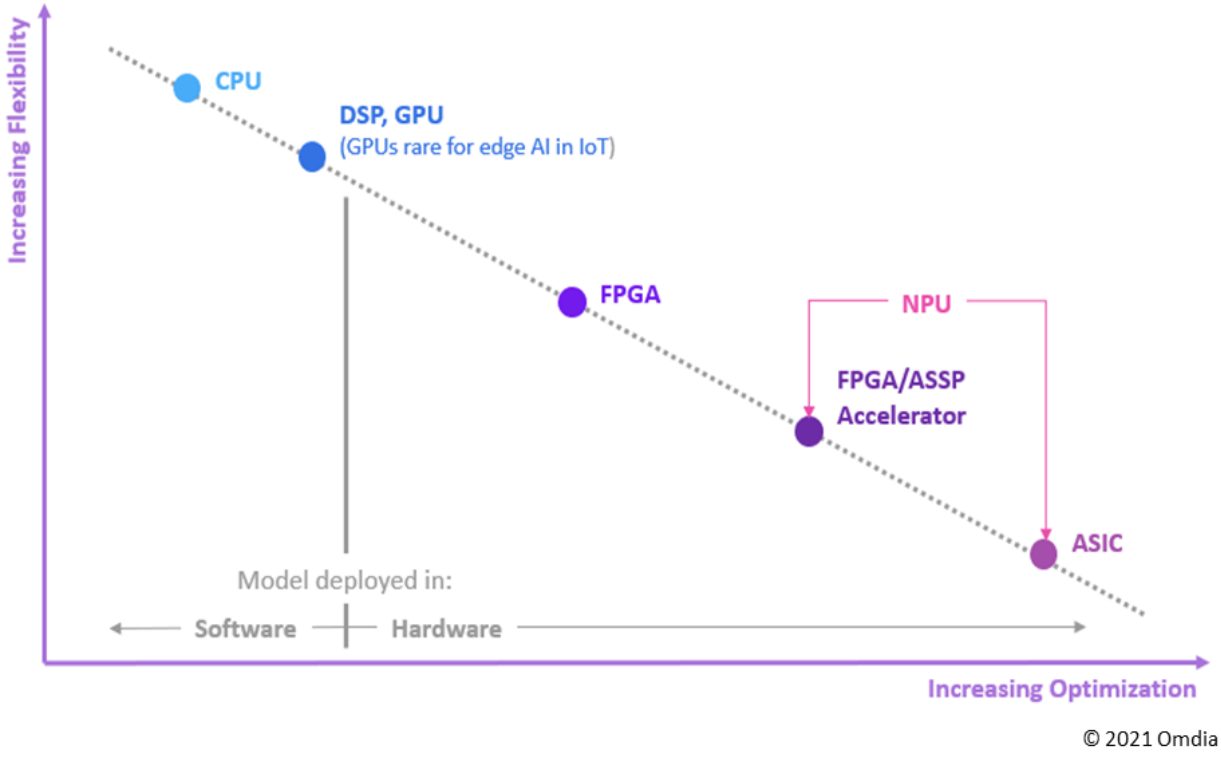
Much of the technical challenge revolves around I/O and how to move both the model parameters determined during training and the example data the model is meant to classify or optimize in and out of the accelerator efficiently. In the classic GPU-based approach, the whole model needs to be loaded into on-chip memory, which drives a huge requirement for RAM and power. At the same time, it is essential to

achieve high memory bandwidth so as not to leave the large, hot, and power-hungry accelerator idling for want of input data (or waiting to write out the output). There are also complex system-level interactions between the GPU and the CPU, as the CPU is responsible for scheduling the GPU, and the GPU is limited by the need to avoid congesting the CPU.

As a result, there is increasing interest in architectures that create a data path independent of the CPU optimized for the specific neural network. Interest is especially high in architectures that either permit processing to happen as data and model parameters flow through the accelerator (so-called dataflow computing) or else carry out the processing in-memory or immediately adjacent to memory (often called at-memory or Harvard architecture computing). A crucial issue is how to accommodate the need for programmability. GPU systems, coarse-grained reconfigurable architectures (CGRAs), and architectures based on very long instruction word (VLIW) technology are generally fully software-programmable, whereas FPGAs are reconfigurable at a hardware level. Some ASICs or ASSPs, on the contrary, are dedicated to one particular neural network or simply act as a pool of highly optimized MACs.

Trade-offs

Figure 3: The flexibility–efficiency trade-off



Source: Omdia adaptation of "Computer vision algorithms and hardware implementations: A Survey," by Xin Fang, 2019

Source: Omdia, Xin Fang

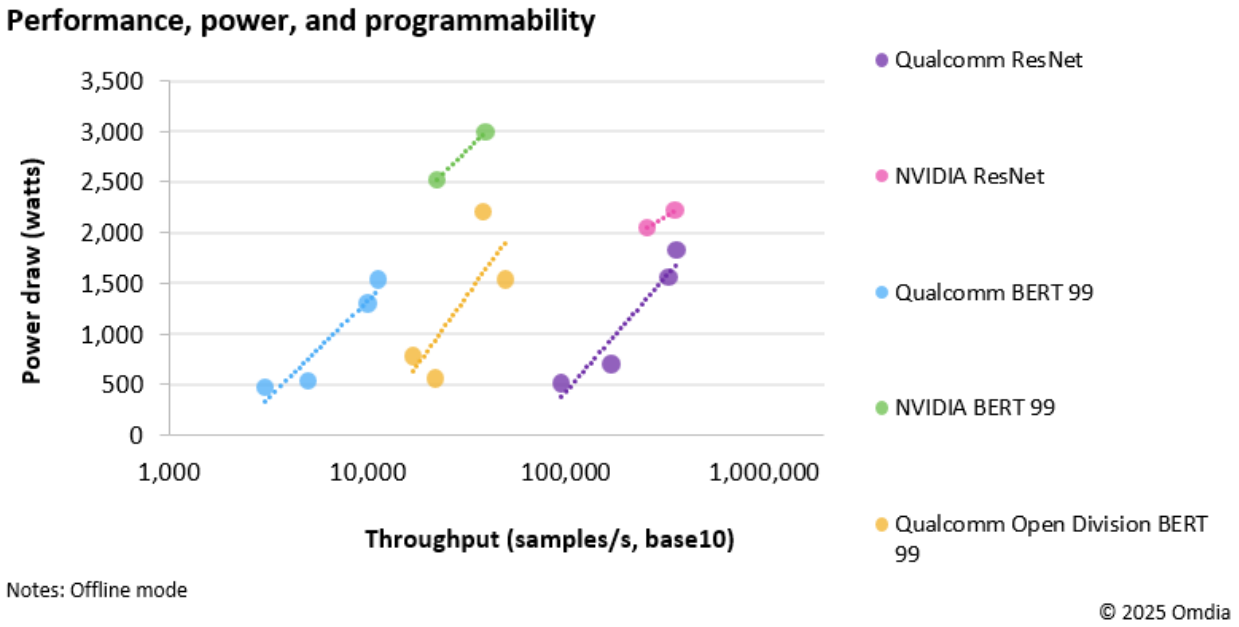
In general, there is a trade-off between optimization for performance and for programmability. An ASIC or an FPGA circuit design could theoretically be optimal for the neural network in question, whereas a software implementation using a GPU incurs various kinds of overhead, meaning a GPU needs to be significantly bigger and faster to achieve the same performance. However, this overhead comes with the benefit of being able to program any model that will fit into RAM and do so with well-known interfaces and

programming constructs. Omdia presents a case study of this last problem in "What's happening to Microsoft's FPGA strategy for AI hardware?"

The Omdia-wide edge definition includes devices destined for smartphones, tablets, and laptops as well as quite a large range of embedded or vertical market products that will be used in a mobile context (e.g., in an unmanned aerial vehicle, or UAV; a drone). These battery-powered systems face very stringent constraints on their power supply, as well as serious thermal constraints. As such, the key metric of performance here is not so much floating-point operations per second, inferences per second, or time-to-complete model training—the usual figures-of-merit for AI processors—but rather these variables divided by power consumption. Omdia expects that given the centrality of power and cooling to data center cost models and the increasing salience of climate change, this will also become true for edge servers and in the data center. As such, AI processor designers are essentially trying to maximize teraoperations per second (TOPS) per watt while limiting how much flexibility, in the sense of general-purpose programmability, they must give up to achieve this goal.

Figure 4 shows the relationship between performance, power consumption, and programmability from the results of MLPerf Inference 3.0.

Figure 4: The relationship between power, performance, and programmability



Source: MLCommons, Omdia

Figure 4 displays the relationship between power consumption (vertical axis) and throughput in samples per second (horizontal axis) for the MLPerf Data center/Power category on the ResNet and BERT-99 tasks. It is important to note that although neither model is cutting-edge, BERT is at least a large Transformer architecture language model and consequently much more representative of the state-of-the-art than the relatively simple ResNet. In the MLPerf Closed division, in which entrants must run the benchmark application as MLCommons provides it without any optimizations of their own, Qualcomm's CloudAI 100 accelerators showed a distinct lead on the ResNet task in terms of per-watt performance over NVIDIA's A100 and H100 GPUs. However, this lead largely evaporated on the BERT task, showing that some of that boost came from specialization; the Qualcomm device was especially good at vision-centric tasks such as ResNet or RetinaNet, and the NVIDIA GPUs performed very well on both tasks.

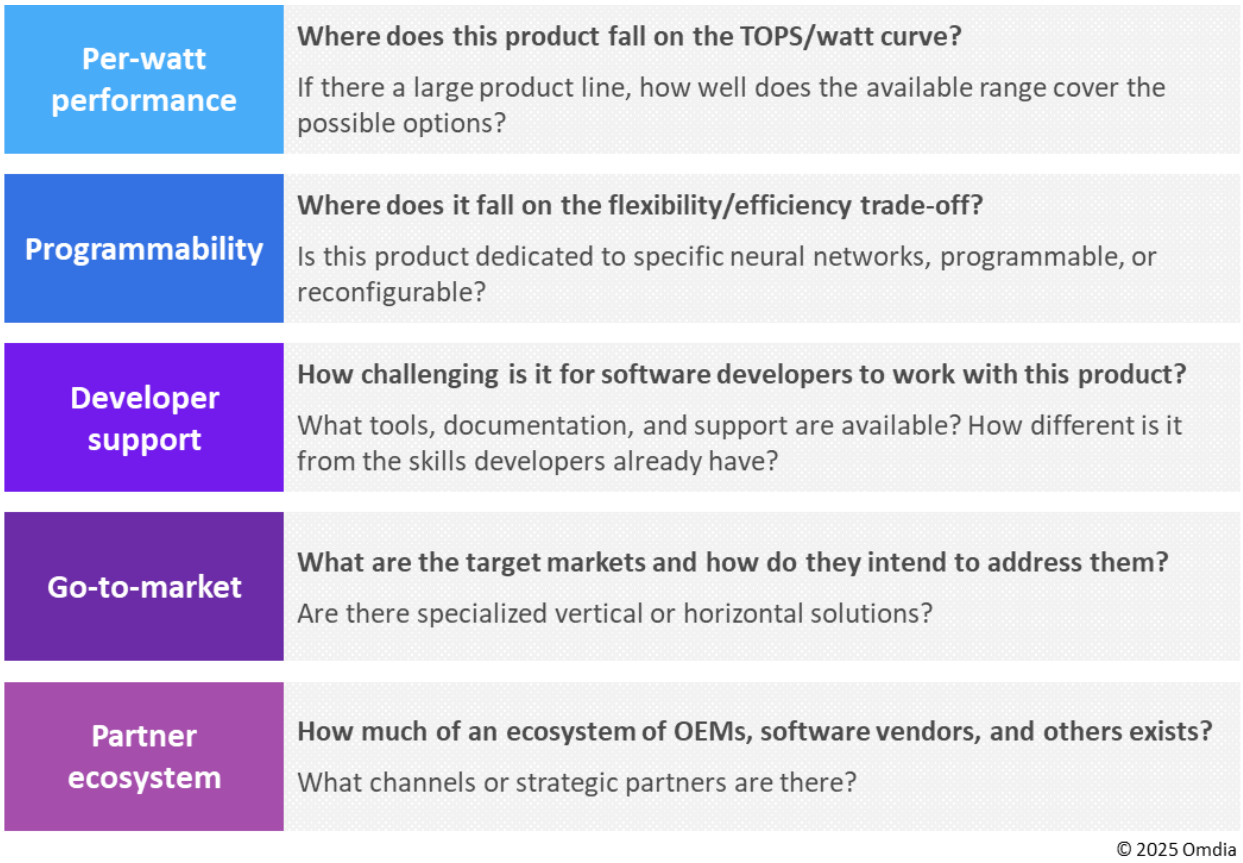
A twist on this comes when considering the MLPerf Open division. In the Open division, entrants are permitted to make hardware-specific optimizations to the models used in the benchmark. Some vendors, notably NVIDIA, only enter the Closed division on the grounds that the requirement to run the benchmark as provided makes it a more rigorous evaluation. However, enterprises deploying AI in practice will very likely use model libraries provided by either the hardware vendor, a cloud service provider, or an AI specialist, which will be optimized for the hardware in use. There is therefore a trade-off between rigor and realism in interpreting MLPerf results.

Given the ability to optimize the software, Qualcomm's part took a very definite lead over its competitor on the Open division BERT task—although, of course, it is not known how much better the NVIDIA results would be using their own code. The lesson here is that although there is definitely a trade-off between programmable, general-purpose capability, and high performance, software programmability and tools can also help to transcend it.

Key capabilities and vendor landscape

Omdia interviewed a list of semiconductor vendors taking part in this market on a semi-structured basis and further reviewed their product data sheets for relevant parts. The resulting assessment is based on this information as well as on broader sources such as published papers and journal articles, conference presentations, and third-party benchmarking exercises such as MLCommons' MLPerf. The criteria used by Omdia are summarized in **Figure 5**.

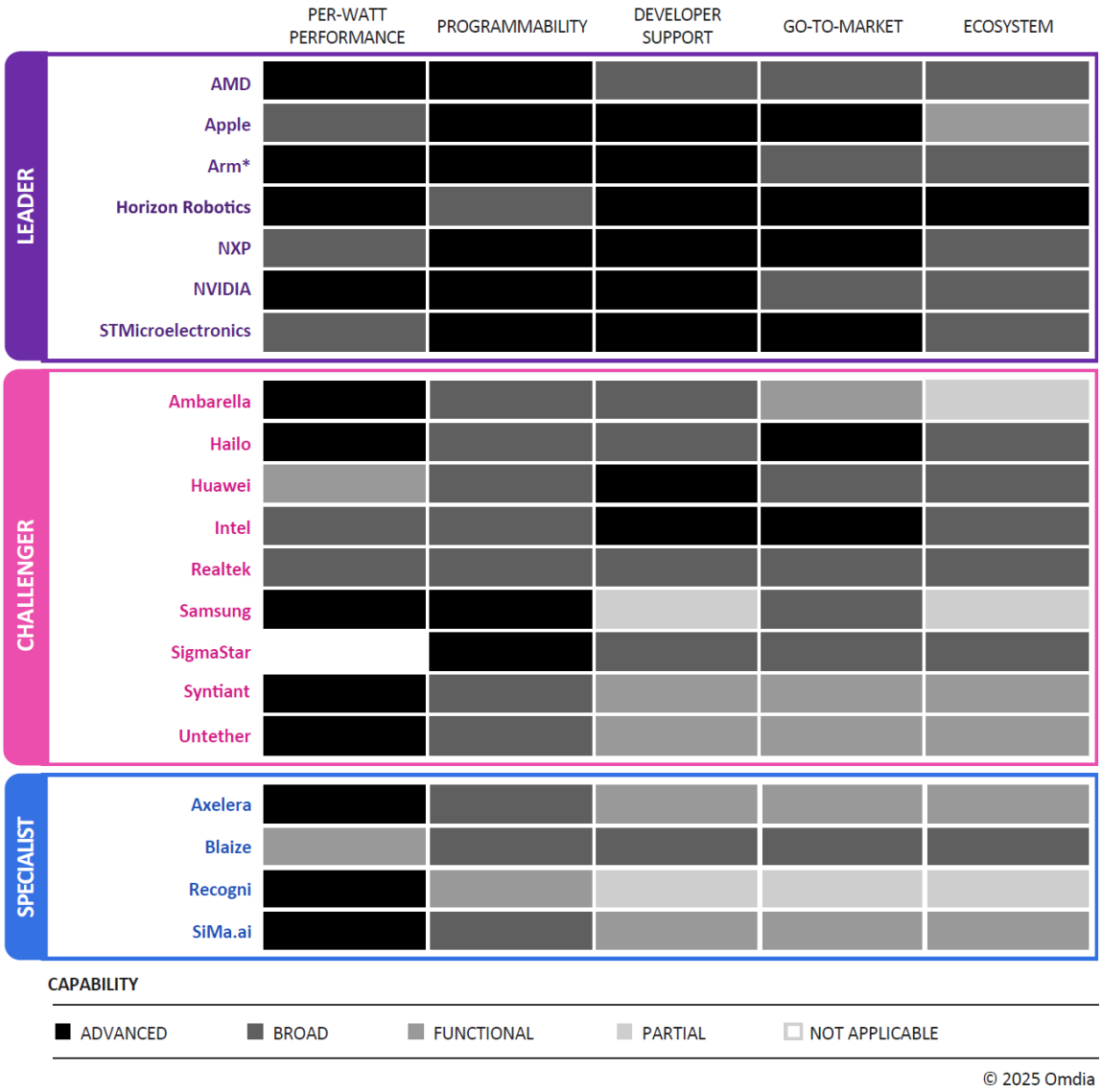
Figure 5: Summary of assessment criteria



Source: Omdia

Figure 6 summarizes the results of Omdia's assessment against the five points in **Figure 5**. It is important to note that the vendors involved have often targeted very specific markets or point solutions, and as a result, not all of them have set out to cover all the criteria.

Figure 6: Omdia heatmap for edge AI processors



Source: Omdia * Note that business model change is a special risk for Arm.

The Omdia Heatmap is colored as follows:

- **Advanced capability:** The vendor demonstrates very strong capabilities and/or capabilities in alignment with what Omdia explored as part of this research.
- **Broad capability:** The vendor offers better-than-expected capabilities that are well-suited to the needs of most businesses.
- **Partial capability:** The vendor provides the expected capability but lacks some of the advanced capabilities assessed as part of this research.
- **Limited capability:** The vendor provides limited or none of the expected capabilities explored as part of this category.

The categorization of each vendor is as follows:

- **Market leader:** This category represents the leading solutions that provide advanced capabilities across six or more areas explored and which Omdia believes are worthy of a place on most technology selection shortlists.
- **Market challenger:** The solutions in this category offer some advanced or broad capabilities, have appropriate functionality across other areas, and should be considered as part of a technology selection process.
- **Market specialist:** Solutions in this category offer a good set of capabilities but lack some of the more advanced features and/or capabilities offered by some competitive offerings.

Market landscape and participants

Market origin and dynamics

The modern market for AI acceleration hardware originated with the mid-2000s insight that GPUs could be used to accelerate general-purpose computing tasks that required massive parallelization. One of the first applications of this insight was the training of neural networks for reasons already discussed. Concurrently, the first very large public datasets for AI training were introduced (e.g., ImageNet from 2004 onward). These developments launched the current AI boom.

The market for GPUs in the PC and server space was and is overwhelmingly dominated by two players—NVIDIA and AMD—yet CPU market leader Intel has long struggled to make an impact, at least with dedicated graphics silicon rather than low end integrated graphics. Before AI played a significant role, the gaming market was the key driver of competition, forcing the development of more and more powerful GPUs that were then reused as AI accelerators. Synergy emerged through the use of bigger neural networks, the creation of bigger training datasets, and the development of more powerful server GPUs and their integration into more powerful machines, resulting in substantial convergence between high performance computing (HPC) and AI training.

Since 2018, there has been an explosive increase in the size of state-of-the-art AI models (a factor of 50x is often quoted), with the result that the demand for compute from AI has also dramatically increased. However, since late 2021, there has been no material increase in the size of the biggest AI models; instead, starting with the leak of the LLaMa models in early 2023, there has been an enormous proliferation of small and medium-sized models, often specialized. This shift is changing the requirements for AI computing and the balance between data center and edge.

In parallel with this, in the mobile ecosystem, Arm, Qualcomm, Apple, and others competed to field GPU cores for mobile SoCs. Here, the importance of photography as a competitive differentiator drove the search for higher performance, although gaming also played a role. Increasingly, smartphone platform vendors integrated ML into their camera applications and introduced voice-based virtual assistants, creating a demand for on-device AI/ML processing and, hence, acceleration.

Above all, the mobile players had to focus on performance per watt, given the extremely strict power and thermal constraints of the form factor. They also, however, had some interesting advantages. To work at all, they had to work at an extremely high level of integration, building much of the device, including

processors, accelerators, memory, and radios, into a single large SoC. They also worked with technologies from the digital signal processing (DSP) domain that were perhaps less familiar in the PC/server space.

Until recently, the markets for server GPUs, mobile SoCs, DSPs, and FPGAs were largely distinct from one other. NVIDIA and AMD fought their tournament largely independently from the parallel competition among Arm, Qualcomm, MediaTek, and Apple, while Intel's ex-Altera FPGA business competed with Xilinx and Lattice and the separate struggle between Intel and AMD for hegemony in the CPU market proceeded along its own lines. Since Apple's seminal decision to move the Mac product line onto its own M1 SoCs, derived from the A-series mobile SoCs, though, the situation has changed, and the market has been redefined around the application—AI/ML acceleration—rather than around the form factor. Vendors attempt to sell their products based on their performance on AI benchmarks, and buyers choose not just among vendors but also architectures. As a result of this shift from a set of distinct duopolies to a more genuinely competitive market structure, there has been a wave of AI accelerator startups and much greater interest in semiconductors from the venture capital (VC) industry.

At the same time, however, many major customers have begun to internalize some of the innovation involved in AI by developing custom silicon for their own needs. Apple's Neural Engine is one example, and Google, AWS, Microsoft, Tesla, and others have all brought in-house silicon into production. Semiconductor designers and foundries such as SiFive, Taiwan Semiconductor Manufacturing Company (TSMC), AMD, Marvell, and Samsung have all developed semi-custom lines of business to support customizers. Meanwhile, both NVIDIA and AMD have begun a major effort to scale down GPU technology for the edge, mobile, and embedded markets, and Intel's multi-front effort has begun to concentrate on the emerging AI PC market, itself in part a consequence of the rise of the small language models.

Key trends in the market

- Although we are living through the rise of small—2 billion to 13 billion parameters—language models, this shift means that the typical model at the edge is still getting bigger. Although smaller models may supplant bigger ones in the data center, now that it is possible to deploy small LLMs to the edge, they are likely to supplant less capable models.
- It was thought that 2021 was a banner year for AI chip startups, but in fact, the industry has not seen anything yet. The downturn in 2022 was only a pause, even with the failure of Silicon Valley Bank, and 2024 crushed the $3bn record from 2021 with total raisings of $4.7bn. VC appetite for AI startups is evidently still keen.
- GPUs are back; the trend toward AI PCs especially is encouraging vendors to develop powerful integrated GPUs, often a kind of Cinderella product. These also provide a relatively good developer experience across device platforms. How workloads are routed among the CPU, GPU, and NPU is increasingly a question of power management and is often somewhat opaque. Mobile-first vendors such as Apple, Qualcomm, or MediaTek have been doing this for years and likely know a lot about it.

Future market development

Growth will fundamentally be driven by AI adoption. The combination of the size of the PC market and the relatively fast adoption of AI within that market is currently the most important factor in this space, and it has already tilted the balance back toward the GPU. However, the tight power, price, and area constraints of the edge will not change; they will enforce a more diverse landscape than the data center. In the PC space, the competitive pressure on both Intel and AMD from the Arm architecture vendors Apple,

Qualcomm, and the MediaTek-NVIDIA partnership will only increase—although concerns about the Windows on Arm experience persist.

AI accelerator products at the edge, and even in the data center, have converged on a roughly similar overall plan, with an SoC architecture, including CPU and accelerator cores, high bandwidth memory, and reconfigurable networks, integrated through an advanced packaging process. As such, the industry is mostly achieving improvements in performance through the so-called more-than-Moore pathway of higher level integration rather than through Moore's Law, sometimes with an assist from shifting to custom products. This is especially true at the edge, where higher integration helps most as it conserves area and does not generate more heat, unlike greater transistor density or faster clock speed. The movement of new packaging options into edge devices will be very important, although held back by price.

The cohort of chip startups funded in 2021 is proving more robust than was expected. Despite two lean years for VC funding in 2022 and 2023, the mini-business cycle associated with the transition to higher interest rates, and the general struggle to get developer mindshare for another AI software development kit (SDK), there have been very few exits. These startups made it to the VC surge of 2024. The exception—and it is a big exception—is China, where there has been a major shakeout of AI chip startups.

AI chip startups in Asia are both more likely to target the edge and much more likely to develop a GPU, and it was precisely the Chinese GPU startups that encountered a sudden stop of funding in 2023□2024. Many of them started with an edge device, asserting it was a stepping stone toward developing a bigger flagship data center GPU; that was a mistake, as edge projects are not really easier in any way. Technology migration has tended to be from mobile-first toward the data center rather than vice versa. These projects faced a double disadvantage: Unlike those that offered a qualitatively new accelerator architecture, they lacked differentiation yet still needed to convince developers to adopt a new GPU.

By far, the biggest upcoming event in the edge AI space will be the release of the PC SoCs developed under the MediaTek-NVIDIA partnership, which will reportedly be in the second half of 2025. These devices will field the NVIDIA Project Grace Arm-based CPU in scale and bring some version of NVIDIA's GPU architecture to a modern PC system. NVIDIA has long offered SoC products for the edge in the Jetson product line, but with the partial exception of automotive, they have not been a major line of business. If these products break through, it will signal the moment at which PCs generally shift to being Arm-based and centered on accelerators rather than on CPUs. It will also seal MediaTek's catch-up development.

Vendor analysis

Why put NXP on your radar?

NXP was the biggest world supplier of MCUs with 20.1% of the market by revenue in 1H24, ahead of Infineon Technologies in second place and Renesas Electronics Corporation in third place. As such, it is a key company in the far edge and IoT markets as AI pushes into these spaces. In this very competitive and commoditized space (top 10 vendors = 86.3% revenue market share), NXP is strongly incentivized to adopt any technology that promises to deliver major design wins, and the key sectors are highly exposed to AI.

NXP manufactures both MCUs and more powerful SoCs for the edge, some of which are offered with one of several AI accelerator types. Most NXP products are Arm-based, and some of these processors include dedicated NPU AI accelerator cores such as the Arm Ethos-U65 microNPU in the i.MX 93 Applications Processor and a 2.3 TOPs NPU core licensed from VeriSilicon in the i.MX 8M Plus Applications Processor. In

2023, the VeriSilicon and Arm cores were replaced by NXP's own highly scalable AI accelerator architecture, the eIQ Neutron NPU, which is currently available on the i.MX 95 applications processor, the i.MX RT700 crossover MCU and the MCX N54/94 microcontrollers. NXP also integrates with several third-party accelerator cores at the system level, such as Kinara's Ara devices (see below), to offer AI processing performance expansion beyond the native capabilities of the NXP SoCs.

Notably, in February 2025, NXP announced its intention to acquire Kinara in an all-cash transaction valued at $307m, which is expected to close in the first half of 2025. As noted, NXP already has a strong system-level partnership with Kinara, and if the acquisition is approved, it will fill out NXP's product lineup with the Ara-1 and Ara-2 discrete NPUs as well as Kinara's AI software portfolio, which will be integrated into NXP's eIQ AI software development environment and includes extensive model libraries and model optimization tools. Briefly, the Ara-1 is designed for general edge inferencing, whereas the Ara-2, which Kinara says is capable of 40 TOPS equivalent performance, is optimized specifically for GenAI applications at the edge. As discrete parts, the Ara processors can be integrated quickly into embedded systems, including upgrading existing in-field systems.

The eIQ Neutron NPU supports a wide range of CNNs, such as YOLO, MobileNet, MobileNet-SSD, and also Transformer model architectures for GenAI. It is described as a flexible accelerator for matrix-multiply and other neural network layer operations and transforms. One reason for going with a design like this is that NXP offers a 15-year guarantee of availability, so it must be expected that new AI models and model architectures will be deployed to devices in the field during their lives, and therefore the devices cannot be too tailored to any particular model architecture. In theory, the design is scalable beyond 10s of TOPS (10,000 ops/cycle at 1GHz). The scale of the NPU needs to be aligned with the use cases, and data pipes and memory sizes must also be considered to keep the system optimal.

The eIQ AI development environment supports the importing of models developed with TensorFlow or PyTorch frameworks, as well as models in the ONNX exchange format. The eIQ Time Series Studio (eIQ TSS) is a newer tool suite that enables the automated creation of classification, regression, or anomaly detection ML models (AutoML) for time series sensor signal sources only from data, removing the need for any data science expertise. There are also a variety of conversion and optimization tools as well as built-in functionalities to provide model copy protection with watermarking and explainability for trustworthy AI features. All these developments are focused on running AI/ML on very resource-constrained devices, including porting models trained on GPUs or other accelerators to NPUs, DSPs, or embedded CPUs. Of note: Despite the name, the eIQ Auto ML extensions are not an AutoML tool but rather a tool for AI/ML development for automotive applications, specifically, developing ML applications while remaining in compliance with automotive standards.

Background

NXP originated as a 2006 spinoff from Philips' semiconductor business that later acquired Freescale Semiconductor in 2015. The company is one of the biggest of the second tier of semiconductor vendors, with revenue of $12.61bn in 2024. It operates its own fabs and specializes in industrial, IoT, automotive, and telecommunications applications.

Strategy and roadmap

NXP's strategy focuses on the automotive, industrial, and IoT markets. It hopes to grow with the increasing semiconductor content in the first two and move upmarket to more powerful embedded processors in the third. Omdia expects that automotive, which accounts for half of NXP revenue, will be the main growth category in semiconductors generally for the rest of the decade, so this makes a lot of sense.

NXP aims to have the biggest selection of Arm-based microcontrollers and applications processors in the markets it serves, and from this point of view, AI acceleration is just another feature. It is, however, an important feature, being both a growth category and one that tends to sell the highest-margin advanced products. NXP management also sees AI as a theme that its three target markets all have in common. This is the motivation for the decision to develop the eIQ Neutron NPU core and progressively replace other NPUs with it in the company's first-party products.

NXP benefits from specializing in the far edge and IoT, which permits it to get out of NVIDIA's shadow. Instead, it is strongly aligned with Arm from a technology point of view. The eIQ ML software platform offers an impressive range of tools for AI at the far edge and is not just a copycat, bare-bones TensorFlow integration but an expandable, fuller solution for ML at the edge, including on some very minimal devices, with some interesting features such as a watermarking model protection tool to identify model copying and IP theft. A clear motivation for the eIQ Neutron NPU project is that it helps to establish the eIQ ML software product as a platform.

Market impact

The company has a No. 3 position in automotive MCUs and a No. 1 position in computing, but the industrial segment struggles by comparison. While NXP was in third place in 2023, it risks falling out of the top five. According to the Omdia *Microcontroller (MCU) Market Tracker*, NXP tends to do better in the more advanced 32-bit MCU category and is most likely to be involved in TinyML AI applications and automotive and computing applications. These are edge AI opportunity areas where NXP is the second-placed vendor, behind Renesas and Infineon, respectively.

Appendix

Methodology

Omdia analysts reviewed product plans, roadmaps, technical documentation, and financial reporting for the companies involved. Companies involved were contacted for comment.

Further reading

AI Processors for the Edge Forecast – 2024 Analysis (January 2025)

Technology Analysis – AI Edge Software Platforms 2024 (January 2025)

AI Processors for the Edge Forecast Report – 2024 Database (December 2024)

Market Landscape: Top AI Hardware Startups, Asia & Oceania – Funding and Trends (October 2024)

Authors

Alexander Harrowell, Principal Analyst, Advanced Computing

Sam Lucero, Chief Analyst, Artificial Intelligence

askananalyst@omdia.com

Citation policy

Request external citation and usage of Omdia research and data via citations@omdia.com.

Omdia consulting

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 you. For more information about Omdia's consulting capabilities, please contact us directly at consulting@omdia.com.

Copyright notice and disclaimer

CONTACT US

omdia.com

askananalyst@omdia.com