Bryan, Garnier & Co

Edge Computing

TAKING AI OUTSIDE OF THE CLOUD

TECHNOLOGY WHITE PAPER JUNE 2019

Contents

EDGE COMPUTING: TAKING AI OUTSIDE OF THE CLOUD

Artificial intelligence technologies have become a powerful way of creating value. Many AI applications have been widely adopted, for example predictive analysis in recommendation engines for social networks and streaming platforms, or face and voice recognition in smartphones and voice assistants. In 2017, 99% of AI-related semiconductor hardware was centralised in the cloud, controlled by only a handful of players including Intel and Nvidia, and represented around USD5.5bn market value.

However, as sensors and compute technologies continue to evolve and become more affordable, AI is spreading to industries including automotive, industrial automation, and healthcare. These applications require low latency, reinforced security and data privacy, none of which can be provided by the current cloud computing architecture. As a consequence, we are moving to a more distributed landscape, where the AI computing capabilities increase at the edge of the IoT node, opening a new paradigm: edge computing.

In this paper, we explore the fundamental differences between cloud computing and edge computing architectures. Looking at different use cases across several industries, we highlight why edge computing is gaining so much attention. We also focus on the related semiconductor market, which is expected to soar from around USD100m to USD5.5bn by 2025. Unlike the cloud, it appears that edge computing will benefit a much broader set of players, from start-ups to well-established microcontroller players and outsourced embedded computing companies.

1. From cloud computing to edge computing

Computing power is a moving target

Centralised versus decentralised architecture

Throughout the history of computing, there have been several cycles alternating between centralised and decentralised architectures. These architectures are not mutually exclusive but complementary, because they address different needs and applications. Within the cycles, there is an emphasis on one of the two architectures to meet new challenges and create value.

175 ZETTABYTES

AMOUNT OF DATA CREATED GLOBALLY TO GROW FIVEFOLD BY 2025

FIG.1: MANY SHIFTS FROM CENTRALISATION TO DECENTRALISATION

1950s	COMMERCIAL COMPUTING SYSTEMS BEGAN WITH A CENTRALISED, LARGE, AND EXPENSIVE COMPUTER CALLED A MAINFRAME
1970s	DISTRIBUTED DATA PROCESSING STRATEGY WITH MINICOMPUTERS
1990s	FURTHER DECENTRALISATION WITH CORPORATE DATA CENTRES
EARLY 2000	CLOUD ARCHITECTURE
MID 2010	EDGE COMPUTING

FIG. 2: PROS AND CONS OF CENTRALISED AND DECENTRALISED ARCHITECTURE

CENTRALISED

Advantages	Disadvantages
Less capital intensive	Single point of failure
Efficiency	Data privacy (on open systems)
Scalability	Latency
Affordability	Less flexibility
Faster technology upgrades	

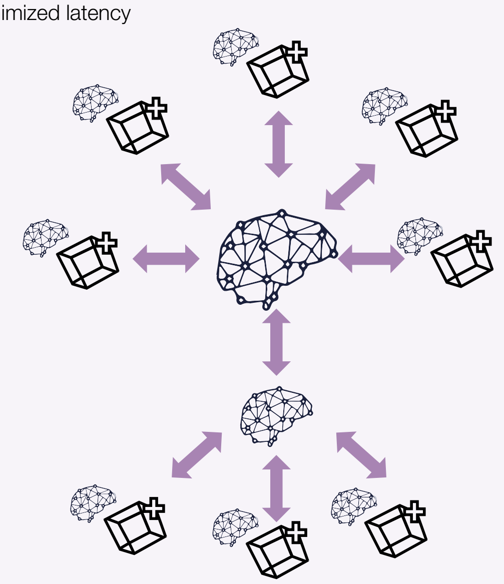
DECENTRALISED

Advantages	Disadvantages
Responsiveness	Maintenance
Reliability	Capital intensive
Flexibility	Form factor and power constraints
Privacy	

Source: Bryan, Garnier & Co

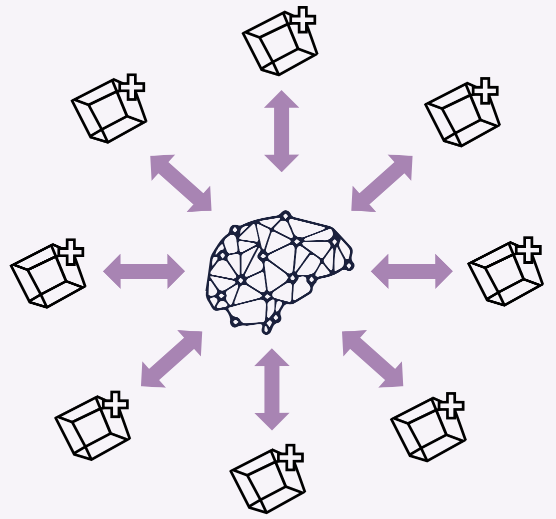
EDGE AI ARCHITECTURE

- Reduced Bandwidth
- Lower Centralized computing power
- Minimized latency



CENTRALISED AI ARCHITECTURE

- High Bandwidth
- HIGH Centralized computing power
- Potentially high latency

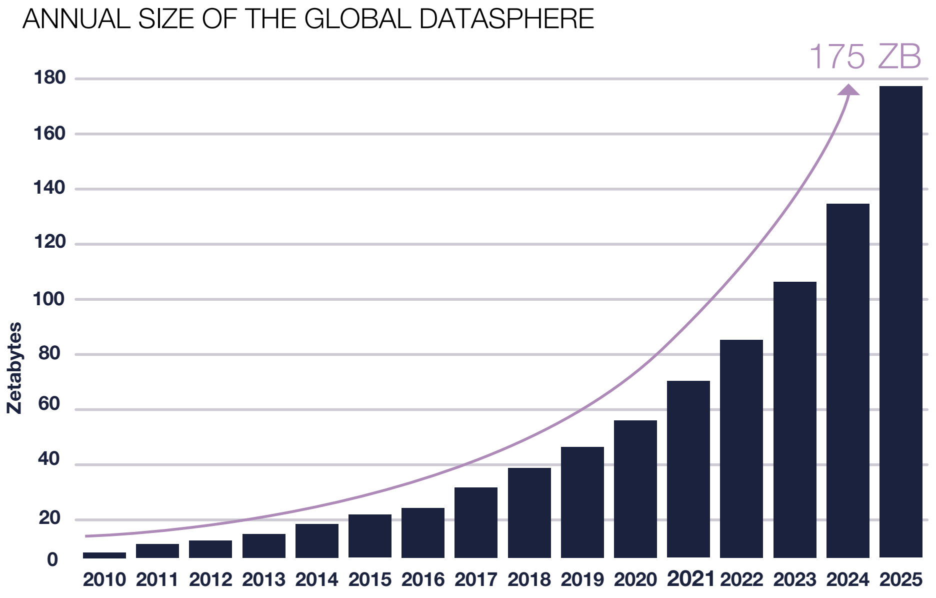


Source: STMicroelectronics

Cloud computing, the last phase of centralisation

Over the past 10 years, computing infrastructure has been centralised around the cloud-based data centre. Driven at first by companies' IT organisations moving data and applications to the cloud for scalability and to optimise on-premise hardware and software investments, cloud data centres have subsequently been fed by the massive growth in data volumes from social media platforms, smartphones and the internet of things (IoT). IDC predicts that by 2025, the amount of data created globally will grow fivefold and reach 175 zettabytes (ZB, 175 trillion GB) to be compared with 33ZB in 2018.

FIG.3: EXPONENTIAL GROWTH OF DATA CREATION



Source: IDC

Cloud computing became even more important in 2015. This was the year when the market achieved significant breakthrough in multicore application processors and graphics processing units (GPUs) to accelerate the processing of large datasets and complex algorithms, and deploy even more powerful AI applications. This has supported a significant growth in public cloud services, representing a market of USD176bn, growth of 21% in 2018 and a projected increase of 17% in 2019, according to Gartner.

For a long time, leading-edge technology like AI has been limited to the IT markets. However, as sensors and compute technologies continue to evolve and become more affordable for companies in the operational technology (OT) market (e.g. automotive, industrial, infrastructure and healthcare), these companies have been seeking to use AI to create value through better manufacturing efficiencies and new business models. However, cloud computing architecture cannot meet all the requirements of these new applications: this is why there is a need to implement a network that is more distributed toward the edge of the node.

Before explaining edge computing, it is important to understand AI.

Basics of artificial intelligence

AI is an umbrella term referring to hardware and software that mimics the cognitive functions of humans and make decisions based on information from the surrounding environment.
AI serves applications where traditional mathematical modelling is inefficient. Modern applications of AI have led to the development of natural language processing, which gives a machine the ability to understand and interact with human speech, and computer vision, which could ultimately lead to autonomous vehicles.

Machine learning

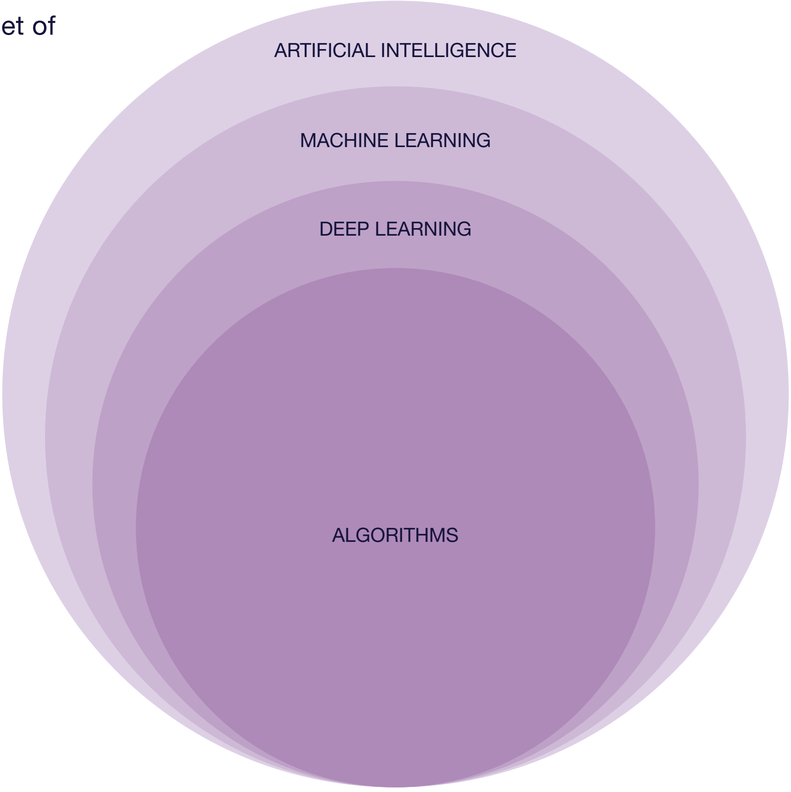
AI relies on machine learning (ML) and deep learning algorithms. Like humans that interact with their environment based on acquired knowledge and past experiences, ML-enabled computers use algorithms and large datasets to identify specific patterns, make classifications, and predict future outcomes. ML enables the computer to learn how to do a task itself and self-improve its accuracy over time.

Deep Learning is an advanced form of ML, using multi-layered neural networks to simulate human thought processes. These networks are made up of small computer nodes that act like the synapses in a human brain. Speech and image recognition, as well as natural language processing, are example of deep learning applications.

For the purpose of this paper, we will not distinguish ML from deep learning and will mostly use the term of ML.

FIG.4: THE MACHINE LEARNING UMBRELLA

Machine learning is a subset of artificial intelligence.



Source: ARM

ML is a two-stage process

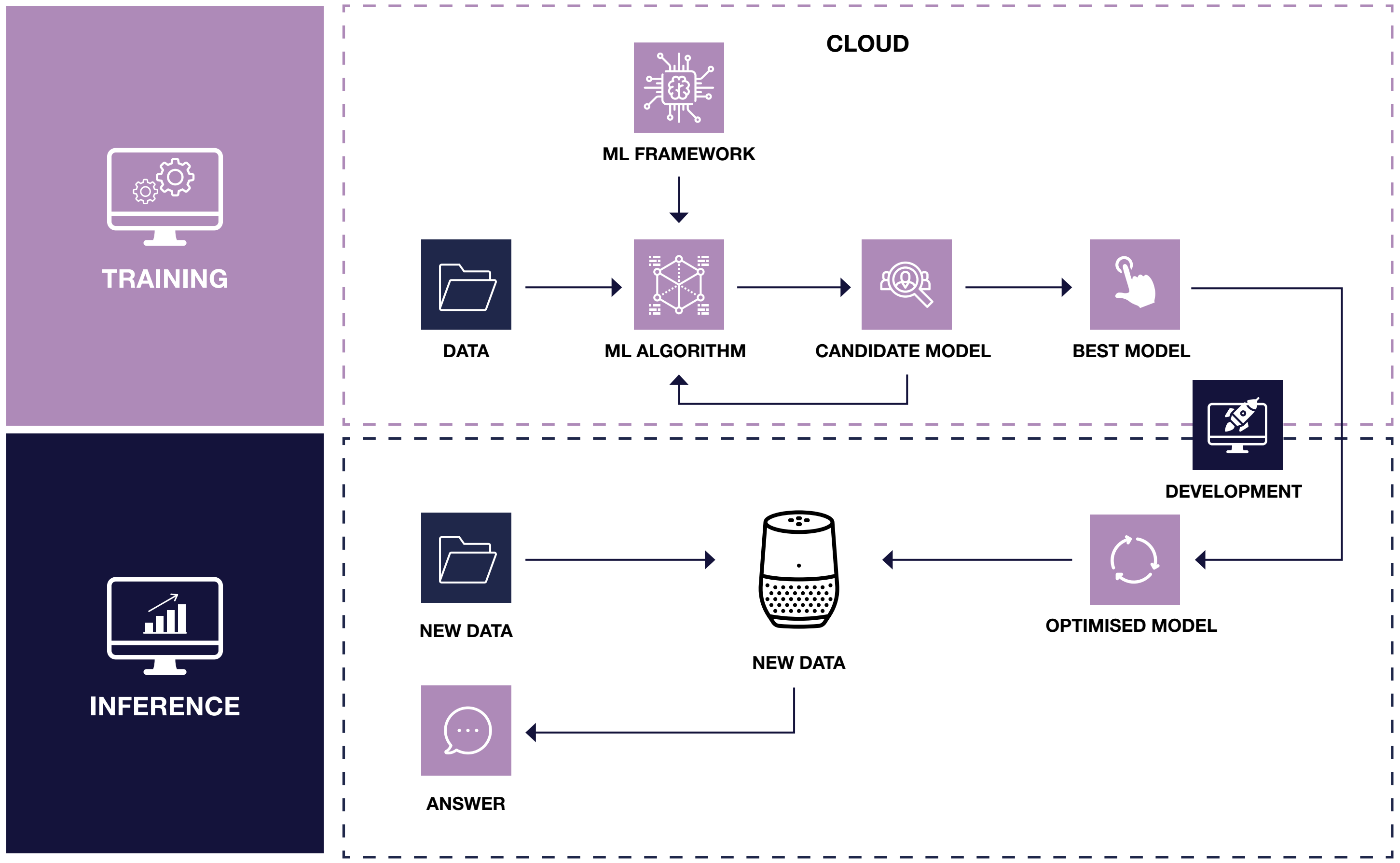
The first is the training stage, in which the cloud is provided with a large set of data to be classified. The algorithm is then fine-tuned until the desired outcomes are reached. The training stage is a long-duration task that requires high computing power. Once this step is finished, the ML system can run the final, trained model as an application to analyse new data, categorise it and infer a result. Known as the inference, this step requires much less processing power.

Where does ML take place?

Most of the inference and training steps today are performed in the cloud. For example, in the case of a voice assistant, the request provided by the user is sent to a data centre where the inference will take place and sent back to the device with the relevant response. To date, the cloud has been the most efficient way to take advantage of powerful and up-to-date hardware and software.

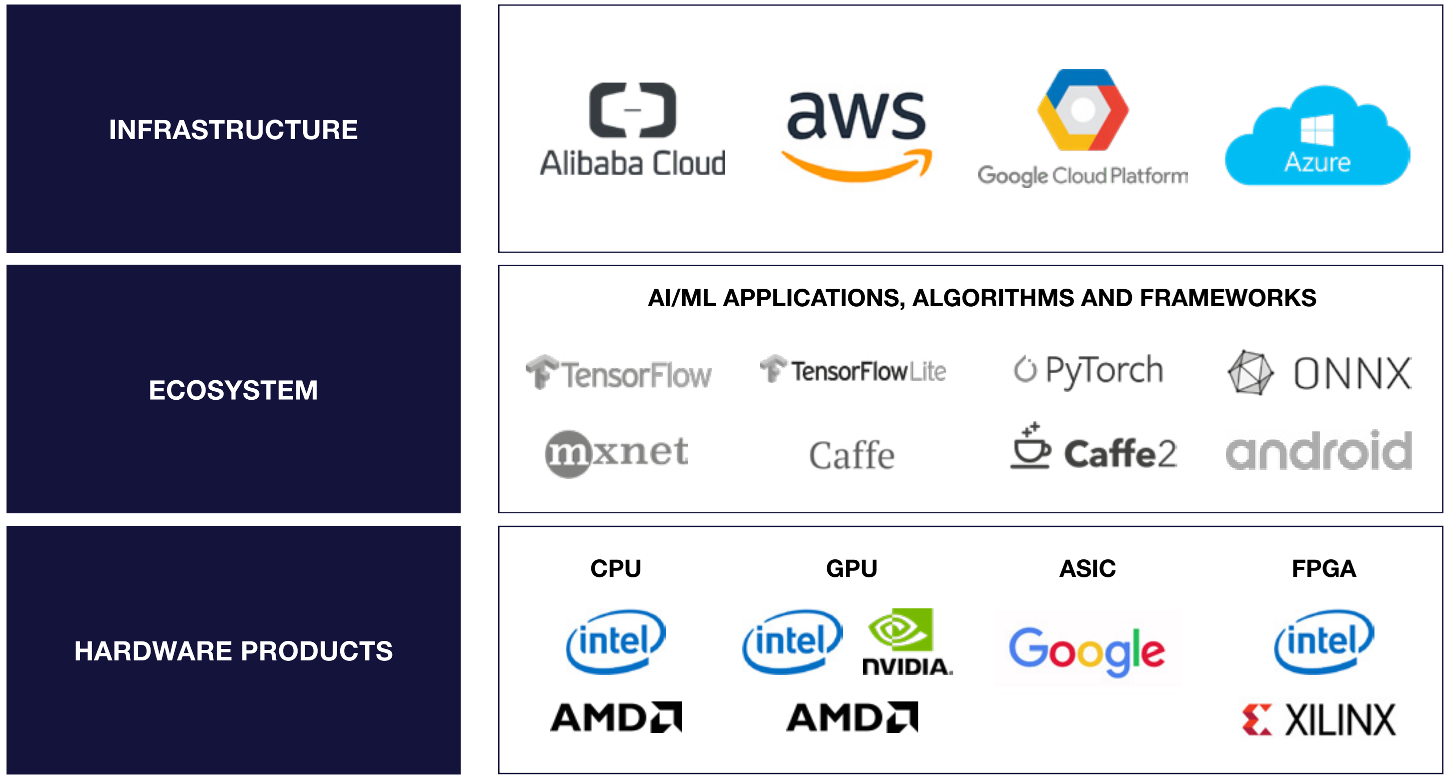
However, together with the need for AI-based technologies to be more responsive and private, the emergence of new applications in various industries is driving the need for a switch from a centralised to a distributed infrastructure, with more and more computing power transferred to the edge of the nodes.

FIG.5: MACHINE LEARNING WORKFLOW



Source: NXP

FIG.6: CLOUD COMPUTING IN THE HANDS OF VERY FEW COMPANIES



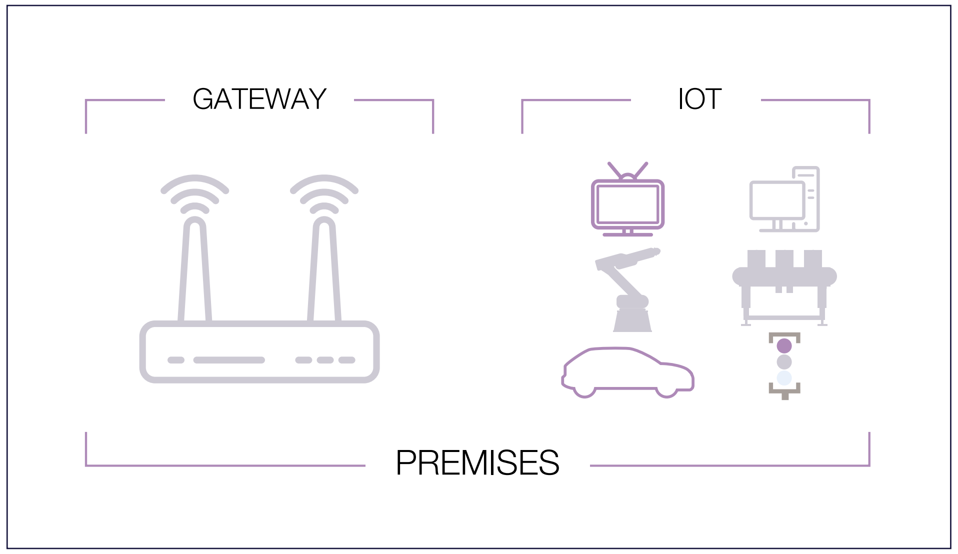
Source: Bryan, Garnier & Co

The emergence of edge computing

Edge computing refers to the practice of bringing compute and storage resources from cloud data centres to an IoT device or a small on-premise data centre or gateway, in order to process data near to where it is generated.

In the context of IoT, an edge device can be a smartphone, security camera, thermostat, a robot – or a self-driving car. Size is not important: anything can be considered as an edge device as long as a portion of the AI computation happened locally.

FIG.7: EDGE COMPUTING IS MULTIFORM



Source: NXP

Edge computing topologies

Edge computing is not meant to displace cloud computing but to help enhance AI computation processes currently carried out in the cloud.

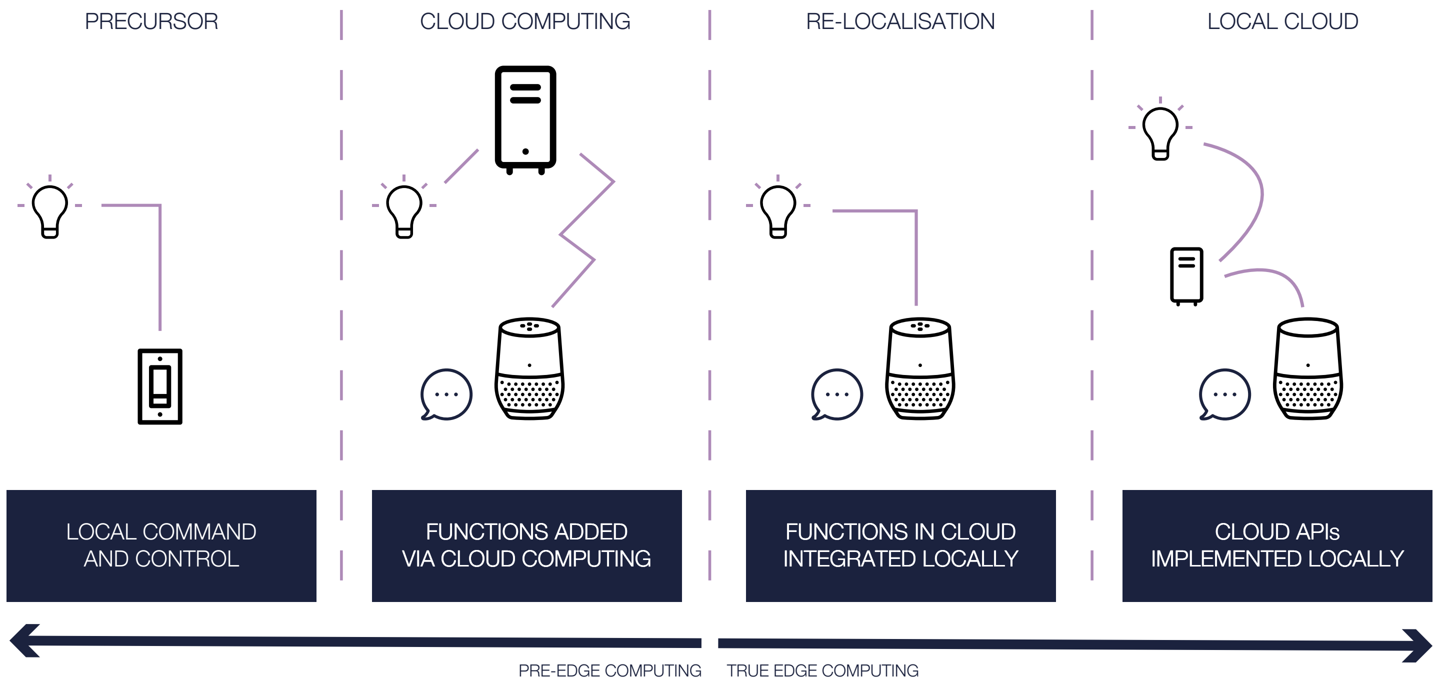
In a traditional cloud ML application, the training and the inference steps are processed at the data centre level. In this architecture, IoT devices have to constantly offload data to the cloud for the inference to be handled through the training model, get a response back to the device and ultimately trigger the desired action. This is not ideal for applications that require real-time responses.

Edge computing devices cannot offer the same computing power and memory capacity as cloud servers, because the form factor of IoT devices and the limited space available for on-premise hardware constrains thermal performance and energy consumption.

Therefore, the idea is to keep the cutting-edge computing devices in the cloud to manage the training phase, while more and more edge devices will be able to do all or some of the inference layers.

There are different cloud computing/ edge computing topologies depending on the applications, but in our view, hierarchical edge computing is the topology most likely to prevail in the future.

FIG.8: EDGE COMPUTING EVOLUTIONARY STAGES

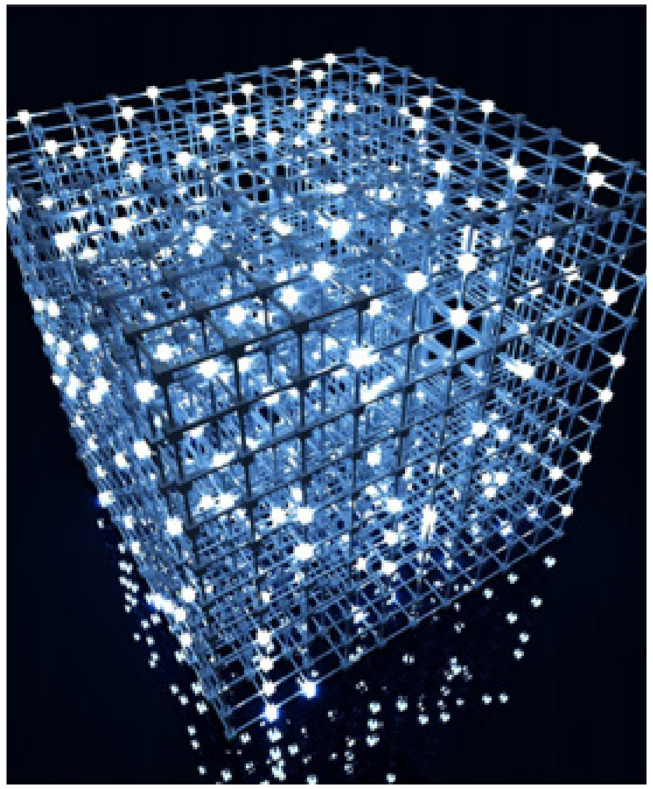

PRECURSOR
CLOUD COMPUTING
RE-LOCALISATION
LOCAL CLOUD
LOCAL COMMAND AND CONTROL
FUNCTIONS ADDED VIA CLOUD COMPUTING
FUNCTIONS IN CLOUD INTEGRATED LOCALLY
CLOUD APIs IMPLEMENTED LOCALLY
PRE-EDGE COMPUTING
TRUE EDGE COMPUTING

Source: NXP

FIG.9: DIFFERENT EDGE COMPUTING TOPOLOGIES

- **Self-contained:** Edge node does all computation for a specific machine or IoT endpoint
- **Hub and spoke:** One edge node services multiple machines/endpoint
- **Peer-to-peer:** Loads migrate among nodes with free capacity or the cloud
- **Hierarchical:** Edge node shares computation, e.g.:

- Endpoint classifies observations (e.g., extracts region of interest, recognizes class of object)
- Edge node
  - Performs next-level classification (e.g., uniquely identifies object within a class)
  - Predicts/decides next steps
- Cloud performs longitudinal analysis

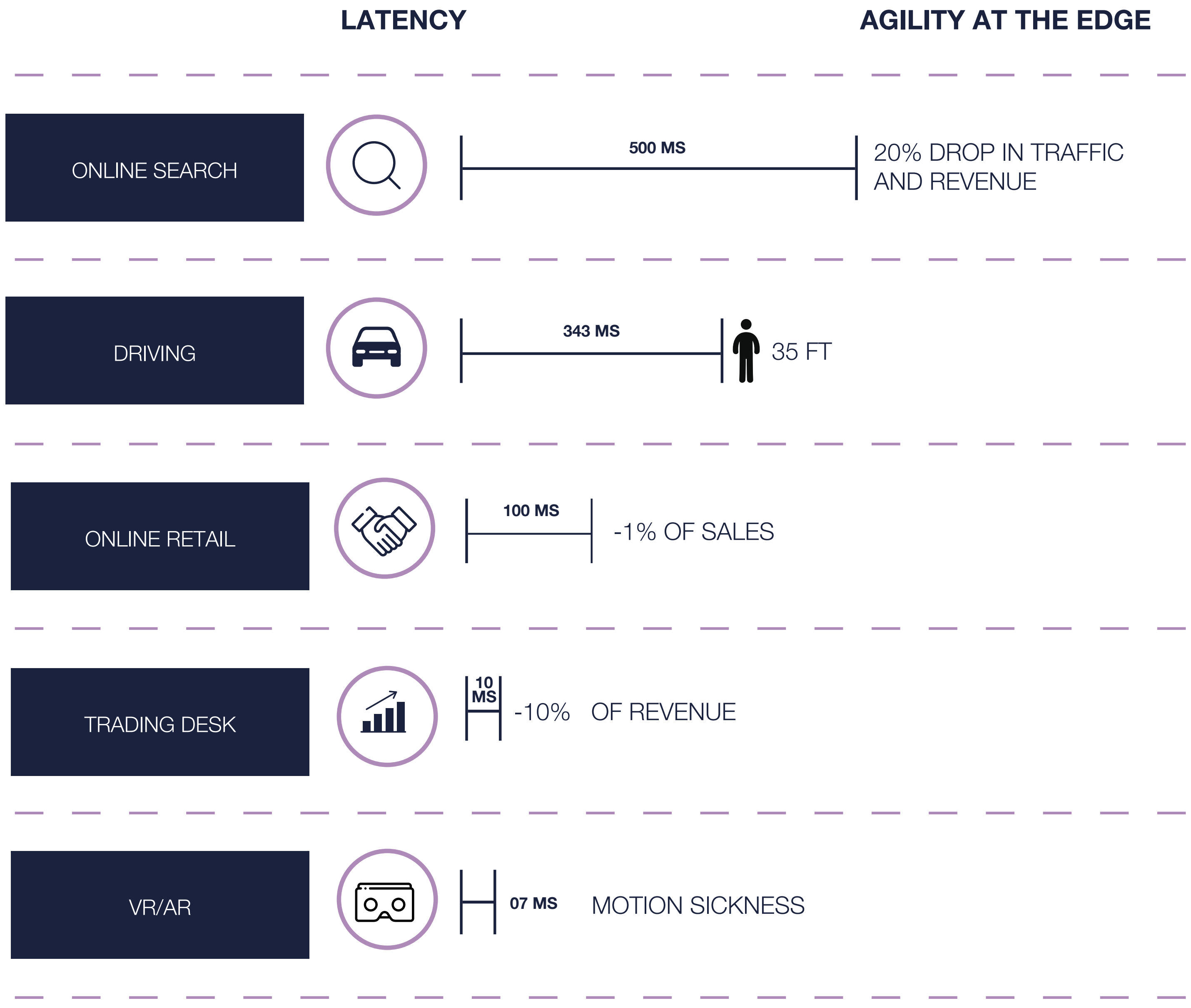


Source: NXP

Reduced latency

The number one goal of taking computing power to the edge is to minimise input-output latency by reducing the time taken for data to travel between the device and the cloud-based data centre. Many time/mission-critical applications, such as automotive, healthcare, or some industrial systems cannot rely on connectivity to the cloud due to safety considerations.

FIG.10: TIME IS MONEY AND SAFETY



	LATENCY	AGILITY AT THE EDGE
ONLINE SEARCH	500 MS	20% DROP IN TRAFFIC AND REVENUE
DRIVING	343 MS	35 FT
ONLINE RETAIL	100 MS	-1% OF SALES
TRADING DESK	10 MS	-10% OF REVENUE
VR/AR	07 MS	MOTION SICKNESS

Source: Gartner

Other factors driving computing at the edge :

Lower operational costs

ML tends to be data-intensive, consuming large amounts of bandwidth and storage. By selecting which data should be stored in the cloud and which should be kept on the edge and processed on the go, companies can optimise costs related to transmission and storage.

Better efficiency and personalisation

In addition to lower latency, edge computing is no longer dependent on internet communication with the data centre and is therefore less prone to network glitches. Also, with ML hardware and software on-premise, companies have better control of the computing design they need to optimise their business.

Enhanced security and data privacy

A single attack on a fully centralised architecture in the cloud can jeopardise the entire workflow of a company, whereas edge computing's distributed system is more resilient. In many cases, sensitive data is transmitted back and forth between the cloud and the edge, making it vulnerable to cyber-attacks. Edge computing can maintain and analyse data locally, avoiding this risk. When processing is executed at the edge, it can also avoid legislative issues around the storing or transmission of data and helps compliance with privacy regulations such as EU GDPR.

Limitations of edge computing

Overheating and power consumption are prominent constraints considering the relatively small size of edge devices. **Hardware choice is key to balancing computing performance and power consumption.** In addition, while in theory the network cannot be compromised with a single cyber threat, multiple edge computing solutions can become overwhelming to manage and maintain in terms of hardware, software or security.

If not properly executed, the cost of deploying and managing an edge computing environment can exceed that of centralised solution, negating many of its benefits.

Given this complexity many companies would not be able to create edge solutions on their own and may need outsourced computing design partners.

2. New business model opportunities

Edge computing to bring bright economic prospects

Gartner anticipates that AI will be pervasive in almost every new software product and service by 2020. By leveraging AI technologies, companies are creating value through cost reduction, customer experience and new revenue streams.

Gartner predicts that the global AI-derived business value would grow from USD1.2trn to USD5trn by 2025.

But, for Gartner's assumptions to be realised, companies first have to take AI to the edge. Accordingly, Gartner believes companies will move progressively closer to the edge, with the share of enterprise-generated data created and processed outside a traditional centralised data centre or cloud increasing from 10% to 75% by 2025. **The edge computing market could be worth USD13 billion worldwide by that time.**

Personal electronics applications were among the first to embrace edge computing topology, for immersive technologies such as personal assistants controlled by voice, image processing, or augmented reality/ virtual reality (AR/VR). These features require high processing power and low power consumption, but above all else need low latency in order to make the experience responsive and natural. They cannot rely solely on the cloud.

As the different technologies are becoming more affordable, AI and edge computing technologies can enhance operational efficiencies and create new business models in a broad range of sectors including automotive, industrial, and healthcare.

By 2030, Gartner believes the industries that will benefit the most from AI are heavy industry, where predictive maintenance could save operators around USD1trn a year; communications, media and services sector, where AI helps decision support and automation; and natural resources and materials, which can use AI to enhance detection and extraction.

There are use cases in almost every sector where edge computing appears to be relevant due to latency, security and privacy requirements. We have decided to exhibit four edge computing applications that we have found convincing in their use and their potential value creation.

USD13 BILLION

VALUE OF EDGE COMPUTING MARKET BY 2025

Autonomous driving

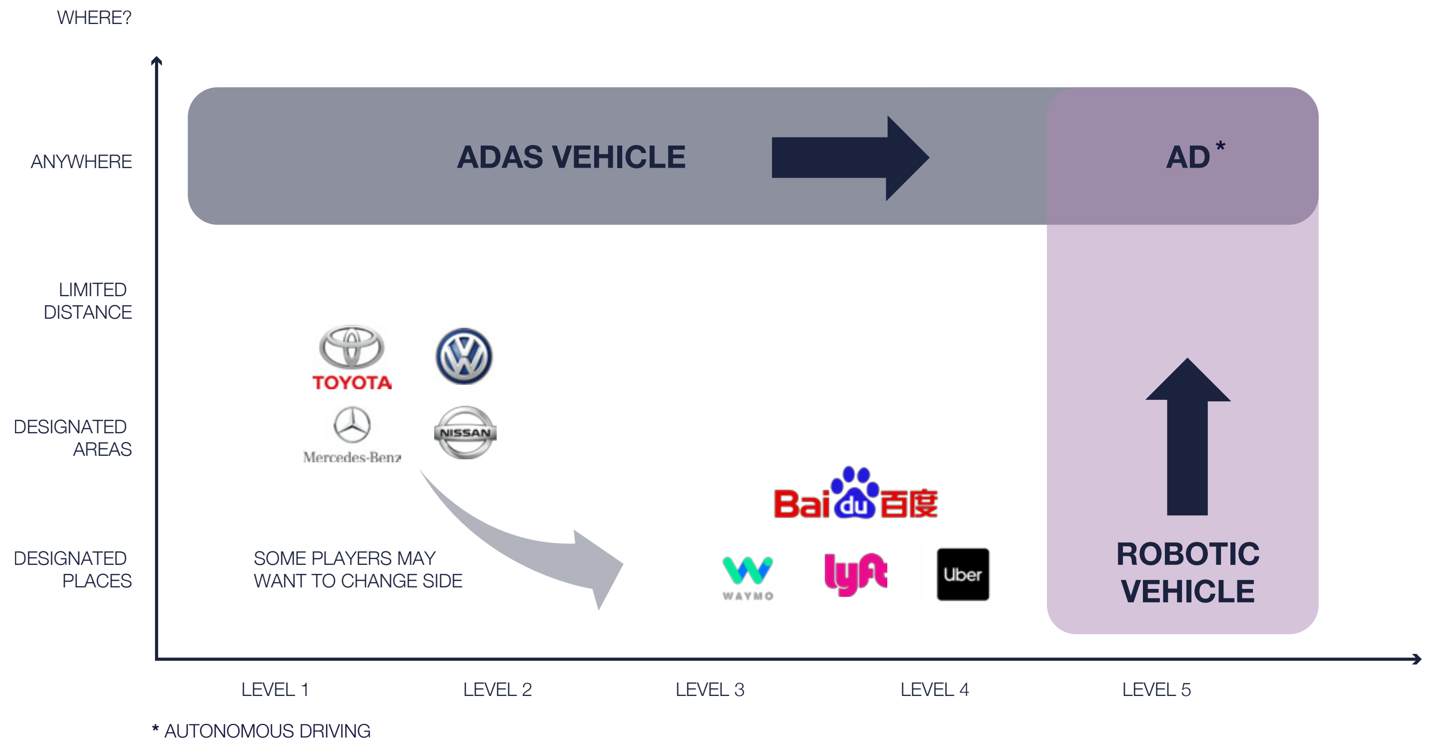
ADAS versus autonomous driving

Car manufacturers may invest billions of dollars to provide a safer driving experience, more connectivity and infotainment, with the ultimate objective for some to create a driverless car. According to IHS Markit, the automotive semiconductor market was worth USD41.8bn in 2018, a growth of 9.5% y/y. It considered the fastest-growing semiconductor application in the future, with the average semiconductor content per car expected to double to around $650 by 2025. New features are made possible by the addition of computing power, sensors, cameras, and radar that collect surrounding data to be analyzed to trigger desired actions.

Let's dismiss from the outset the idea that all new vehicles put into service in the next 20 or 30 years from now will be driverless. The complexity and cost of both the components – an extra USD1,200 worth of semiconductors in a vehicle with level 5 versus level 0 autonomy – and the ecosystem of self-driving cars will limit the deployment of these vehicles in volume.

Whether for advanced driver assistance systems (ADAS) or autonomous driving systems, AI in vehicles is a growing reality, and its applications cannot rely solely on remote servers to make decisions about, for example, whether or not a car should brake in front of an obstacle. **Automotive is one of the mission-critical applications that require the lowest levels of latency and therefore greater use of edge computing.**

FIG.12: NOT ALL OEMS WILL GO TO AUTONOMOUS CARS



Source: Yole

Healthtech

Patient-generated health data

IoT solutions give the healthcare sector new opportunities for patient care delivery and medical data collection. This leads to the collection of large amounts of patient-generated health data related to the use of connected wearables such as blood glucose monitors. It makes possible better diagnosis of diseases and real-time monitoring of patient health, and can also play a preventive role by sending alerts when patients show anomalies. However, it also brings challenges in terms of data management and security. Edge computing can solve this by avoiding the use of the cloud and keeping the backup and analysis processes in local areas. It also favours fast real-time analysis, which is crucial in health emergencies.

Edge machine learning solution for asthma patients

According to the World Health Organisation, there are an estimated 340 million people worldwide who suffer from asthma; in 2018 alone, over 360,000 died from the illness.

Amiko's Respiro smart inhaler technology uses a sensor to collect inhaler use data and sends it to an application on the patient's smartphone to improve asthma treatment. The data transfer uses Bluetooth low-power connectivity, which favours safety without the need to connect to the cloud.

FIG.13: MOBILITY-AS-A-SERVICE: DOMINO'S PIZZA AND FORD

In 2017, Domino's, the world leader in pizza delivery, and Ford Motor launched a delivery service using self-driving vehicles. Ford provided its self-driving Ford Fusion Hybrid Autonomous Research Vehicle.

Once the pizza is ordered, it is distributed by the vehicle and the client receives a message with a unique password to open the car to retrieve their delivery. This first "experience" took place under limited and supervised conditions. In 2018, after the first positive results, Ford and Domino's Pizza conducted a two-month test in Miami. This second study added the difficulty associated with the urban environment (heavy traffic, more parameters to be taken into account for AI decisions). Ford's objective is to have a fleet of autonomous-drive vehicles delivering goods in the US by 2021. Critical in allowing this to occur is edge computing technology.

USD 20 MILLION

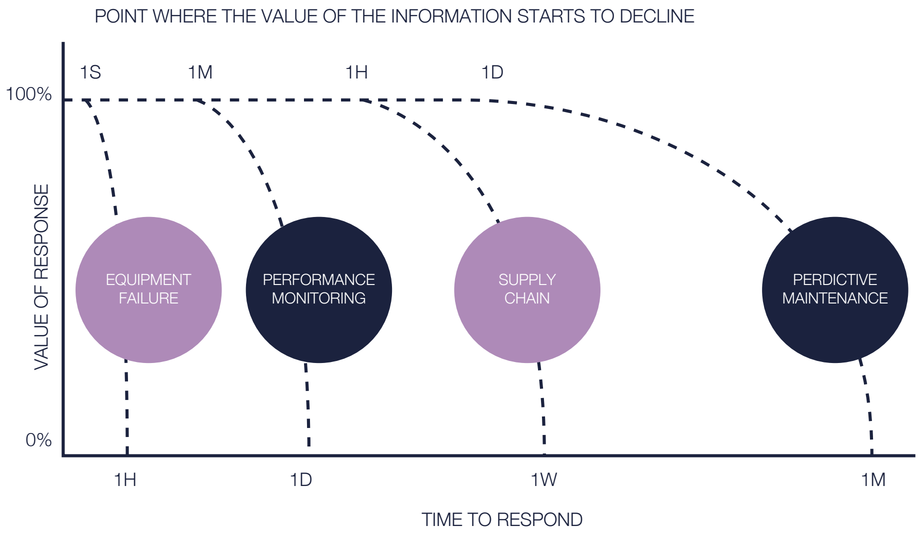
COST SAVINGS BY LG DISPLAY USING GOOGLE-BASED EDGE COMPUTING.

Industrial IoT (IIoT)

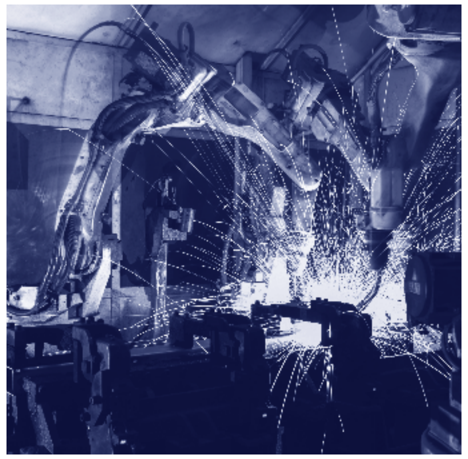
Manufacturing production line

Industrial companies can use edge computing to increase operating efficiency. For example, real-time analytics from sensors in the plant area can be processed on an edge device and provide alerts in case of malfunction. Any delay in information transfer caused by the cloud can cause significant economic losses. Edge computing can help not only in the monitoring and detection of a point of failure, it can also be used to optimise a supply chain and predict potential failures that need immediate attention. As an example, an LG Display manufacturing line improved its anomaly detection from 60% to 99.9% by using a Google-based edge computing system, saving USD20m per year.

FIG.14: TIME-VALUE RELATIONSHIP IN IIOT



Source: Industrial internet consortium



Mining industry

In mining, sensors and AI help to significantly reduce maintenance costs and downtime while increasing output. A metal mine that uses industrial IoT sensors associated with edge computing on a gateway and/or directly in the machines can collect and analyse real-time data in order to constantly run at an optimal state. As a consequence, miners can increase mineral recovery by 1 to 3% and raise throughput by 4 to 8%, while reducing energy consumption. Together, these gains can enhance productivity by 5 to 10%, the equivalent of opening of a new mine.

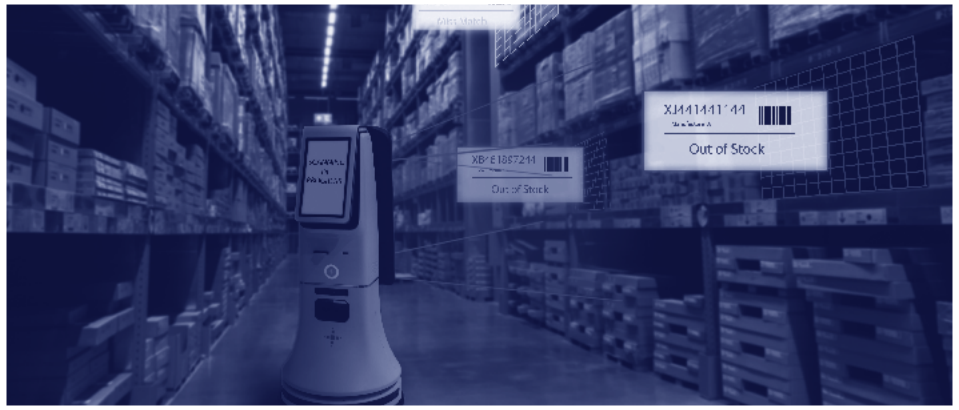
According to McKinsey, Western Australian miners operating with autonomous haulage technology report a productivity improvement of 20% compared to only 2.8% average gain in productivity for the overall mining industry between 2014 and 2016.

Autonomous mobile robots

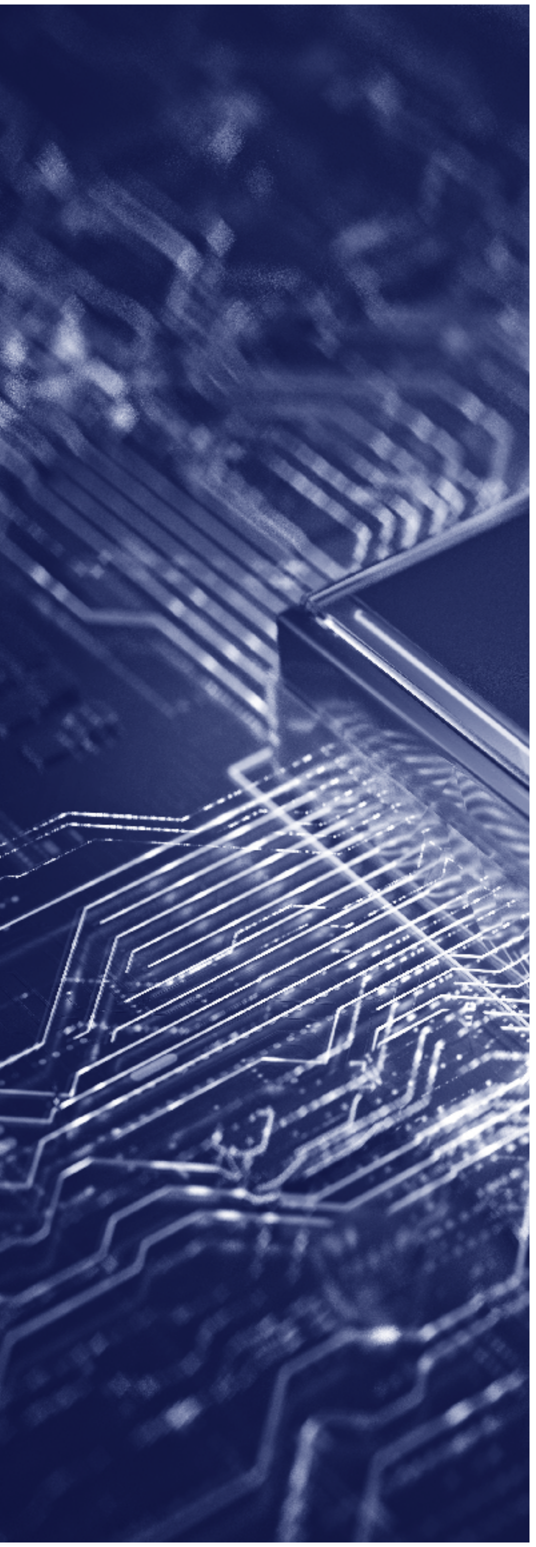
Robots in industry have evolved into a new generation of autonomous mobile robots (AMR) using visual processing unit (VPU)-accelerated AI computation. AMRs are a form of automated guided vehicle but differ by their degree of autonomy. Mainly implemented in logistics, they can move throughout warehouses and logistic facilities using AI at the edge that helps them find the best path to take and avoid collision risk.

With fleet management software, robots can be controlled globally and assigned according to their location and availability, increasing their efficiency. Amazon recently bought warehouse automation company Canvas Technology, which builds autonomous robots with edge computing allowing them to move and put away boxes on shelves while avoiding people and obstacles, even in a crowded space.

FIG.15: REPRESENTATION OF A 3D IMAGING PROCESSING IN A WAREHOUSE



Source: TechCrunch



Control energy consumption of motors and drives

According to a study from congatec a Germany-based company with a leading position in embedded systems for edge computing, more than 45% of the global electrical energy is currently used by motors and drives. For all manufacturers, one of the key questions is: where are the energy losses in the system and how big are they? To maximise productivity, companies need to implement applications that can continuously adjust machines in the most efficient way possible. Edge computing can measure different performance variables and then transfer orders directly to the production machinery in order to control energy consumption and improve productivity.

For this congatec is, for instance, actively involved in developping advanced server-on-module embedded systems to help manage IIoT and limit the latency associated with the use of the cloud.

Smart retailing

Advanced shopping experience

A new type of store has appeared using advanced technology to renew the classic retail industry model. Using computer vision, it increases convenience for shoppers thanks to time-saving and user-friendly infrastructures. The concept allows customers to make purchases without having to wait in a queue at the checkout.

In the US, Amazon, via Amazon Go, initiated this new approach. Amazon Go stores use the Just Walk Out technology. This includes the use of computer vision, deep learning algorithms and sensor fusion. The use of edge computing is necessary to reduce compute latency and create a smooth and pleasant customer experience, especially in case of high traffic in the store. In order to enter the store, the customer has to use the free Amazon Go mobile application to then browse and choose items. The vision processing tracks customer activity and detects when and which items they pick up. Upon leaving the store, the Just Walk Out technology automatically calculates the invoice and withdraws the total from the customers' Amazon account. This process avoids the potentially long checkout queue that is often experienced in traditional retail stores. There are currently, 12 stores using this technology that are already operational in the US.

In China, JD.com, one of the largest e-commerce retailers, has also seized on the opportunity with its JD.id X-Mart concept. Like Amazon Go, JD.id X-Mart uses AI, facial recognition and RFID (Radio Frequency Identification). To enter the store, the customer has to login with his personal QR code, and his identity is confirmed by facial recognition. Then, the shopper is free to select items on sale (consumer goods, clothes, HPC). Currently, 20 stores have already been opened in mainland China and the concept is beginning to be exported to East Asia.

Drone delivery

For supply chains, drone delivery could reduce costs, environmental footprint and congestion. To increase safety and avoid collisions, drones use AI-enabled embedded computing systems that can scan the airspace and navigate in real time.

As with autonomous vehicles, this is an application where edge computing should be prominent due to low latency requirements. It makes drones safer, faster, more efficient and environment-friendly, especially in remote areas.

Alphabet in Australia

In April 2019, Alphabet, through its subsidiary Wing Aviation, received approval from regulators in Australia to launch a drone delivery service. Approval followed 18 months of testing and over 3,000 deliveries.

Around 100 houses in Canberra and its immediate suburb can now receive orders (small items only) delivered by drones.

Alphabet in the US

At the end of April 2019, Alphabet received approval as an airline in the US, with subsidiary Wing Aviation permitted by the Federal Aviation Administration (FAA) and the Department of Transportation to deliver goods (small consumer items first) in Virginia. Alphabet can now invoice customers for drone delivery. This is the first time that regulators have granted a technology company the same status as a charter company or air cargo carrier and therefore represents a giant leap for the drone industry as it reinforces its position and legality.

3. Great opportunities for a new set of players

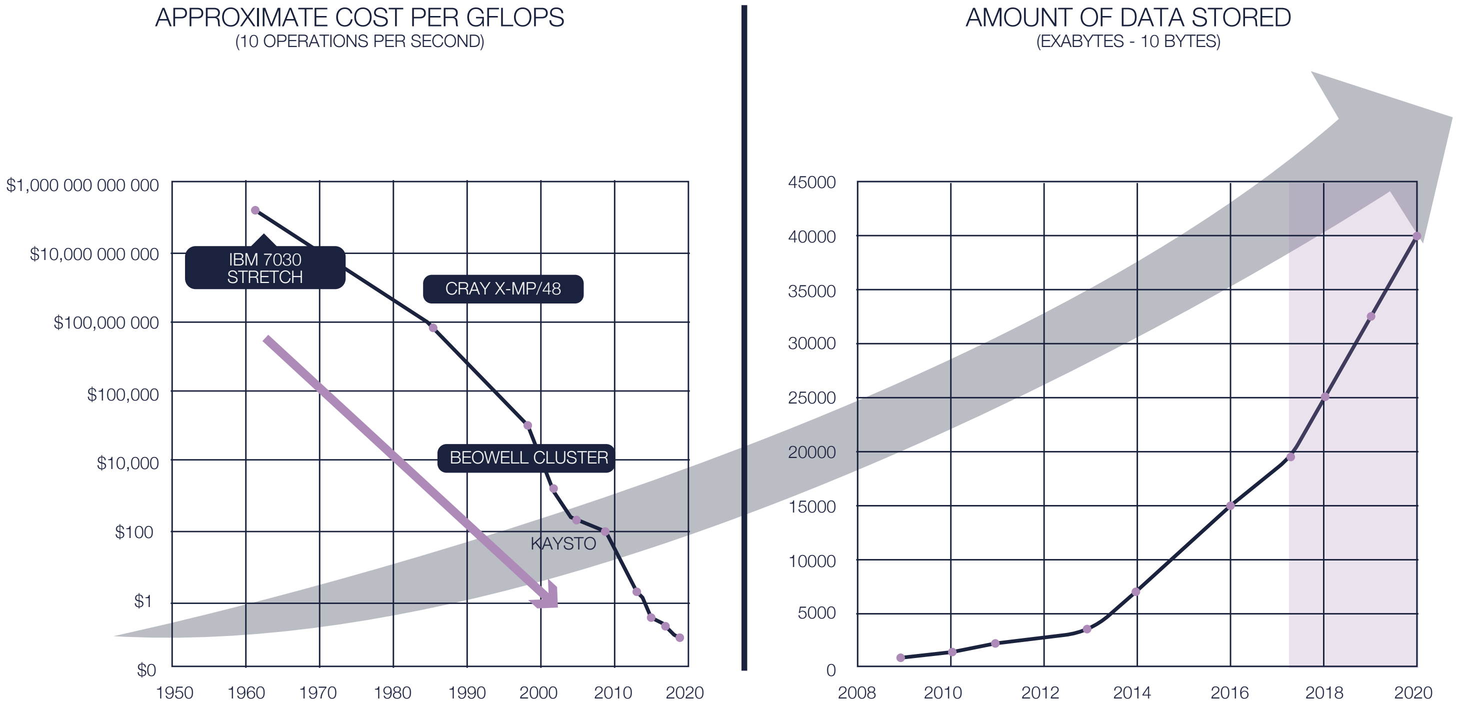
Semiconductors at the forefront of AI innovation

The evolution of semiconductor technologies has made possible the current IT infrastructure built around fast and large-capacity data centres. Raw performance of semiconductor chips keeps improving as engineers strive to keep Moore's Law a reality (Moore's Law defines the pace of innovation in processing speed and indirectly, the price curve of computing technologies).

The current trend around IoT and AI creates even more value for semiconductors, as it relates not only to logic (computing) and memory chips, but also to more advanced sensors and connectivity.

McKinsey estimates that semiconductor companies will capture between 40 to 50 percent of the total technology value created by AI technologies. AI-related semiconductors are expected to grow at 18% CAGR over the next years to reach USD67bn revenue by 2025, a rate that is approximately 5 times higher than non-AI semiconductor business.

FIG.16: DECREASING COST OF COMPUTING POWER



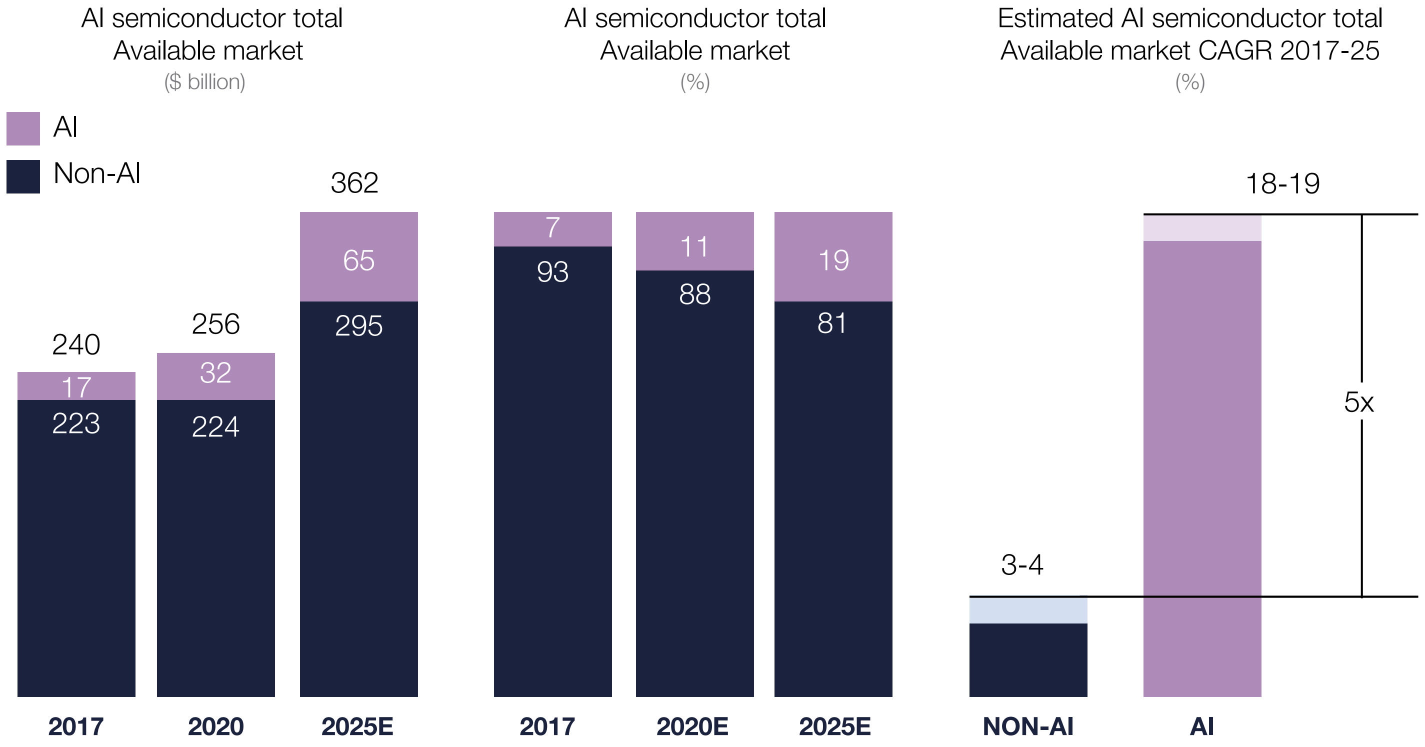
Source: Yole

No one size fits all

Due to the deceleration in Moore's Law and the need to go to the edge with new constraints in form factor, power consumption, and thermal issues, the industry has developed new AI-related hardware designs to meet specific requirements.

ML workloads can be processed using different types of computing hardware, such as central processing units (CPUs), graphics-processing units (GPUs), field programmable gate arrays (FPGAs), or application specific integrated circuits (ASICs). The optimal choice between these computing architectures will depend on the nature and complexity of the AI application and whether it would be a cloud- or edge-located system. Depending on this information, the processing requirements and the power and heat constraints, the accuracy and speed of the computation differ significantly and influence the hardware design.

FIG.17: GROWTH FOR SEMICONDUCTORS RELATED TO AI IS EXPECTED TO BE FIVE TIMES GREATER THAN GROWTH IN THE REST OF THE MARKET

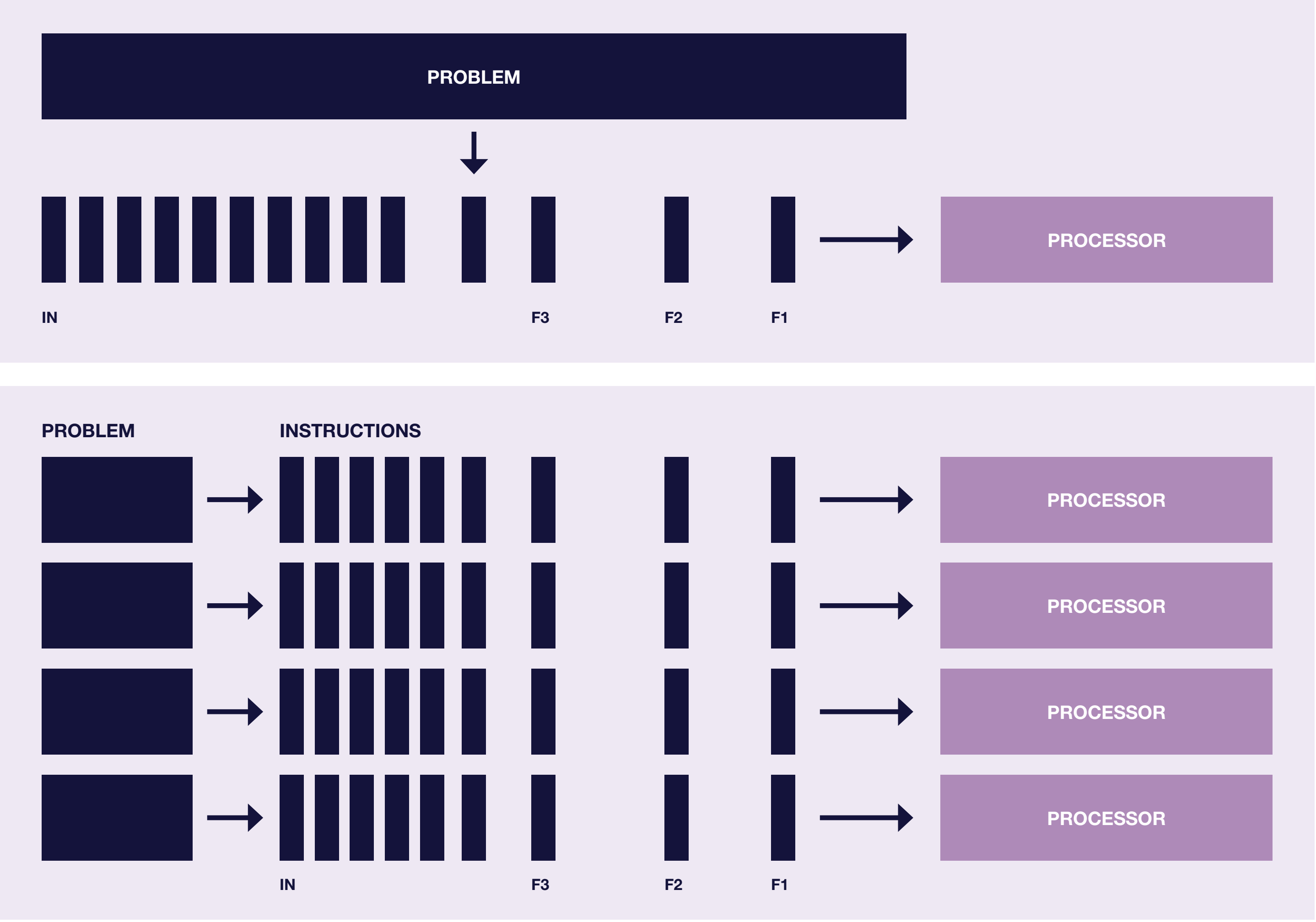


Source: McKinsey

Serial versus parallel computing

In a serial computation, computing hardware, usually a CPU, is used to read a series of instructions that are executed sequentially. Only one instruction may execute at any moment in time. Parallel computing refers to a computational problem that need simultaneous use of multiple compute resources in order to be solved. Applications such as 3D image processing, machine learning or bitcoin mining rely on parallel processing. Compared to CPUs that have a couple of cores, GPU architecture can go up to hundreds of cores and can handle multiple tasks simultaneously, which explains, for example, the success of Nvidia's GPUs in solving machine learning algorithms during the training phase. Conventional multicore CPU systems are not suitable especially for the training phase of machine learning and AI-dedicated GPUs or other specialised hardware are generally used to accelerate the training phase. In parallel, CPUs will generally handle the inference phase in a data centre.

FIG.18: SERIAL VERSUS PARALLEL COMPUTING



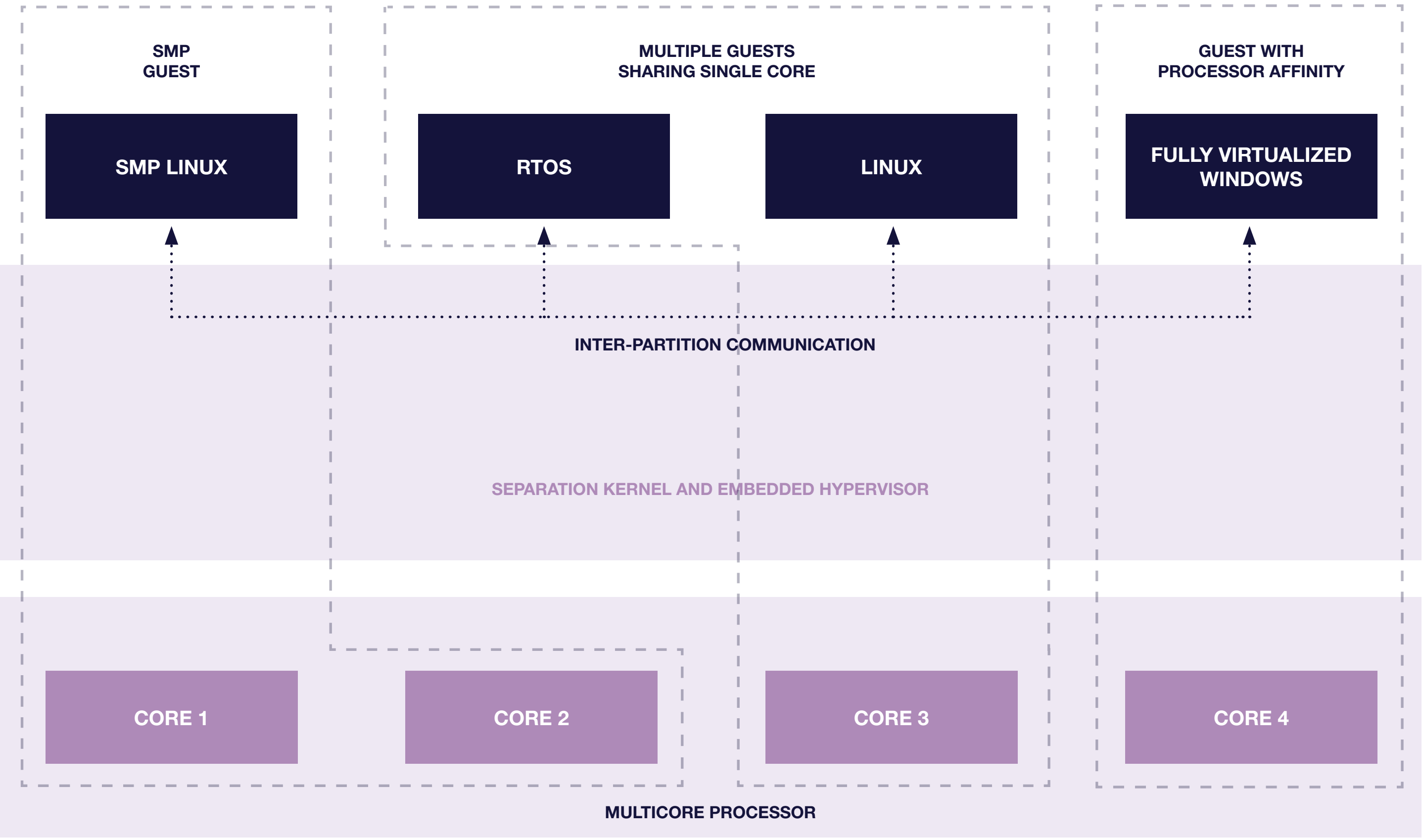
Source: Lawrence Livermore National Laboratory

Embedded virtualisation brings more flexibility and efficiency to hardware design choices

Virtualisation is a software layer managed through a hypervisor, also known as a Virtual Machine Monitor (VMM), that allows a single hardware device to run different operating systems (OS). As embedded computing becomes more complex and requires, in many cases, different types of hardware and operating systems, embedded virtualisation becomes a critical software tool to efficiently optimise the system cost and workloads.

For example, in industrial IoT, an embedded system typically runs on a real-time OS (RTOS) that handles basic real-time operations. On the other hand, there are numerous enterprise-level applications for user interfaces, 3D graphics or AI that require the use of more advanced OS such as Linux or Windows, and therefore would need the support of additional hardware. Instead of multiplying the number of components with cost, form factor and power consumption constraints, a hypervisor allows to run several operating systems in parallel by dedicating one or more processor cores to each individual operating systems. This ultimately decreases the overall total cost of ownership for the end-customer and enables workload consolidation.

FIG.19: VIRTUALISATION ENABLES TO RUN DIFFERENT OS AND APPLICATIONS ON SPECIFIC CORES


SMP GUEST
MULTIPLE GUESTS SHARING SINGLE CORE
GUEST WITH PROCESSOR AFFINITY
SMP LINUX
RTOS
LINUX
FULLY VIRTUALIZED WINDOWS
INTER-PARTITION COMMUNICATION
SEPARATION KERNEL AND EMBEDDED HYPERVISOR
CORE 1
CORE 2
CORE 3
CORE 4
MULTICORE PROCESSOR

Source: RTC Magazine

ASICs are gaining momentum

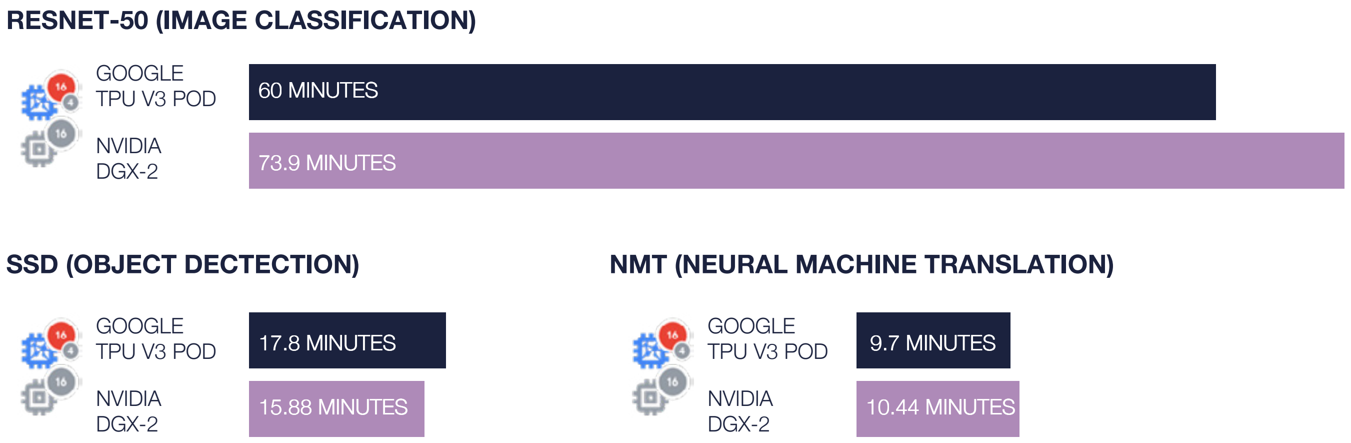
GPUs have proven to be far better than CPUs for training neural networks, but the latter should keep its dominance for inference tasks. Nonetheless, ASIC chips are becoming more attractive and are expected to gain market share both in the cloud and at the edge. For long, GPUs have been preferred to ASICs due to their greater flexibility and lower cost, but the need for cloud-computing providers to improve their computing efficiencies and create differentiated solutions from the competition will increase the use of ASICs. ASICs stand for Application Specific ICs and are designed for a specific customer and a very specific task. In the context of AI, a dedicated ASIC can address a specific layer of the ML algorithms and significantly improve the process speed and the results.

For example, to accelerate learning and inference for products like Google Translate, Google Photos and Google Assistant, Google built a custom-designed machine learning ASIC called a Tensor Processing Unit (TPU).

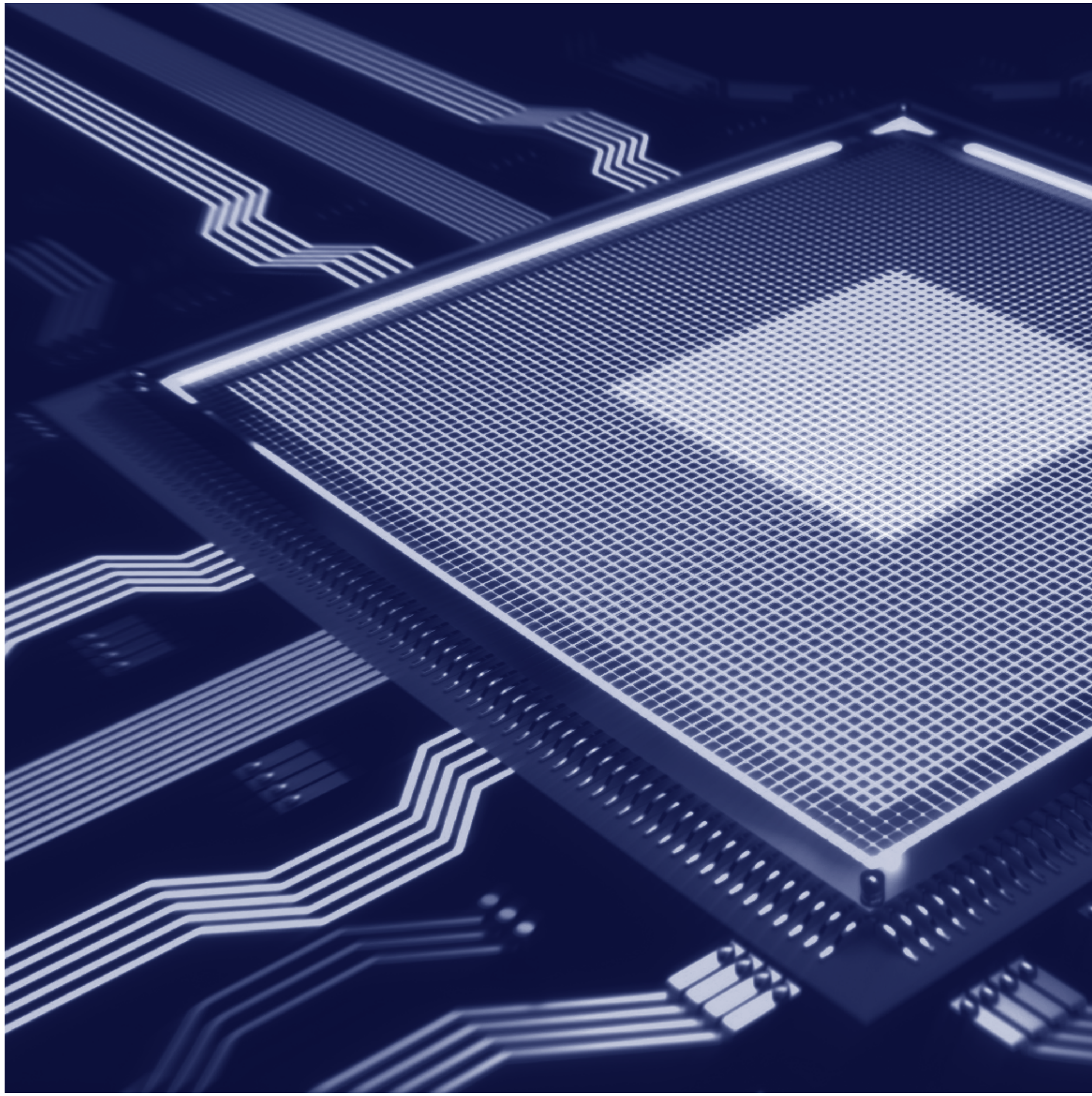
Google also provides a development toolkit for its TPU to allow third parties to use its chip for edge computing (example LG Display factory, for more details see chapter “Industrial IoT”, page 17). Based on MLPerf (a recognised ML benchmark solution) results, Google claims its TPU achieves a 19% speed improvement compared to a competing NVIDIA solution.

GPU IS NO LONGER SUITABLE FOR MOST APPLICATIONS AT THE EDGE DUE TO ITS HIGH ENERGY CONSUMPTION

FIG.20: ABSOLUTE TRAINING TIMES BETWEEN GOOGLE’S TPU AND NVIDIA’S DGX-2



Source: MLPerf, Google



Energy consumption is key at the edge

Processing power is added at the edge as companies look to increase operational efficiency, reduce latency, and reinforce security and privacy. We previously noted that edge computing, especially at the device level, will focus on inference computation. Compared to the learning phase, inference needs less processing power, However, energy consumption and thermal limitations become major factors at the edge of the node. GPU-based computing architecture is no longer suitable for most applications at the edge due to its high energy consumption. As a consequence, CPUs and MCUs will be preferred for edge computing, and ASIC chips will be used in some cases to process very specific applications in the most efficient way.

For example, leading MCUs providers like NXP, Renesas, or STMicroelectronics have been engaged in AI R&D for several years and now are capable of providing MCUs that are able to run deep learning algorithm to do specific tasks. These MCUs have pre-trained neural network models and can infer outputs for different applications. MCUs are optimised to run with limited memory and processing power. Alternatively, CPUs will be used where the is higher computing needs and less emphasis in energy efficiency.

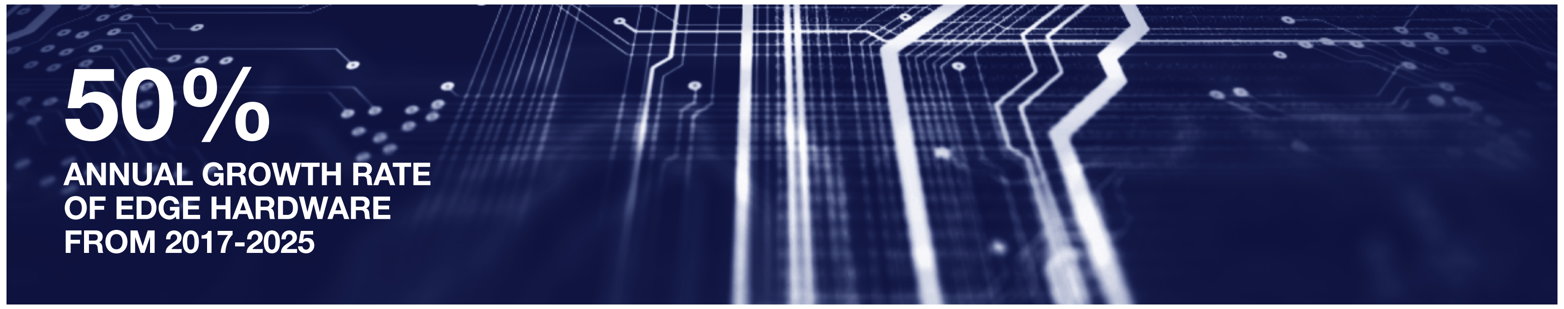
In some cases, edge computing designs can combine CPUs or MCUs with more advanced and customised chips like ASICs and FPGAs to improve the process speed and results accuracy.

KALRAY

Kalray is a publicly traded French company created in 2008 that has developed a FPGA-like family of programmable silicon devices called MPPA (Multi-Processor Array) to address AI inference. The company expects its sales to ramp up in 2019 (EUR77m in 2018) driven essentially by data centres. While sales from automobile applications are not expected to be significant for 3 to 5 years, the company achieved important milestones as attested by strategic alliances with NXP, Renault, Baidu, and Autoware, and believes its addressable market would be worth EUR1.5bn by 2025. The microcontroller portfolio of NXP, the world's second-largest automotive processor supplier, will combine Kalray MPPA processors to address advanced driver-assistant system (ADAS).

GRAPHCORE

Graphcore is a UK-based company that has designed a new type of processor for AI acceleration called the intelligence processing unit (IPU). The IPU is capable of processing complex parallel computation and is said to be suitable for both training and inference. Similar to Kalray, Graphcore's modules will be first used into servers for cloud computing, and potentially go into autonomous vehicles as well in the future. Graphcore raised USD200m at its last funding round in December 2018, valuing the company at USD1.7bn. Graphcore investors include Dell, Bosch, BMW, Microsoft and Samsung.



50%

ANNUAL GROWTH RATE OF EDGE HARDWARE FROM 2017-2025

Edge computing hardware is a fast-growing market that will benefit more players

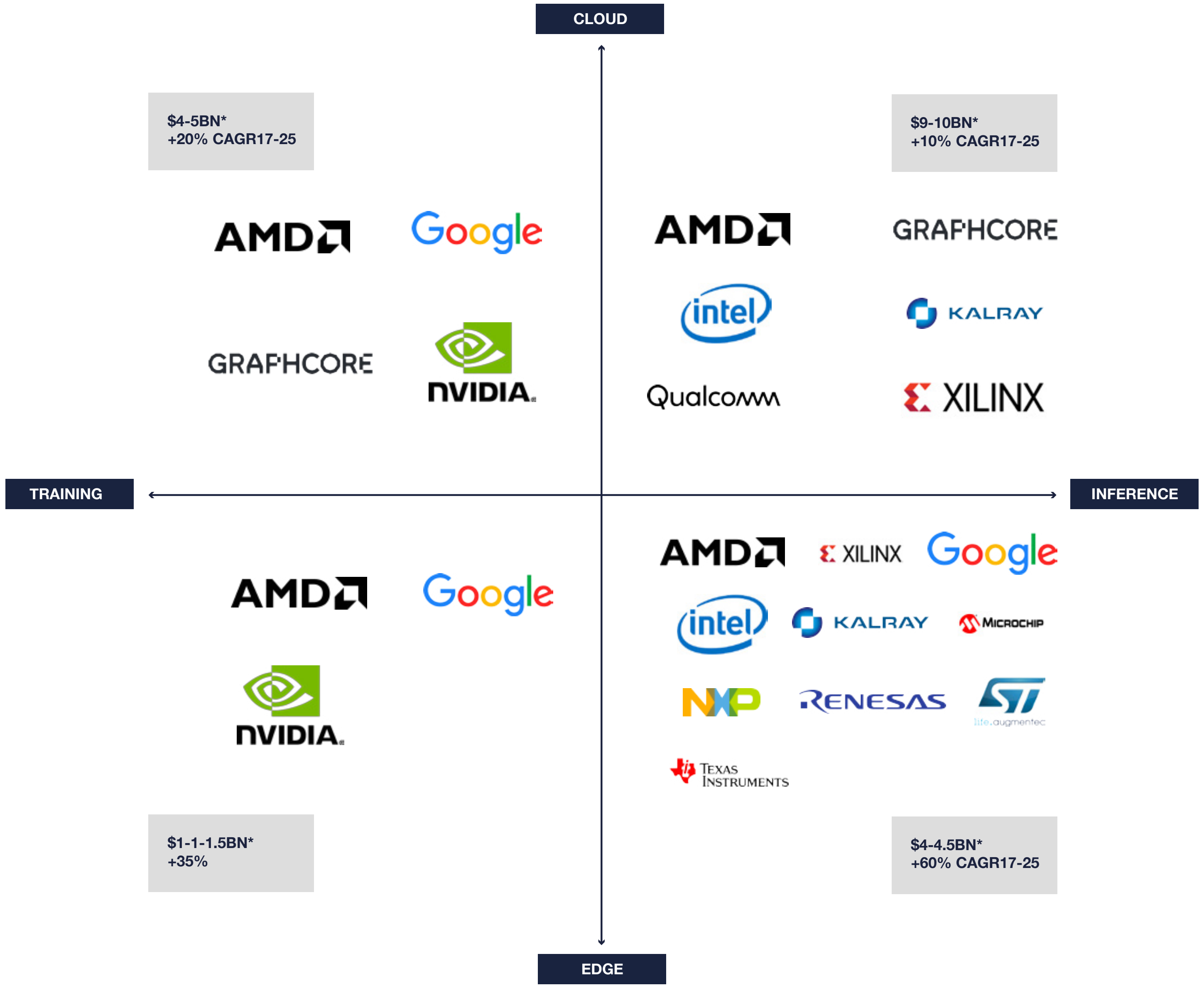
Intel controls nearly 90% of the cloud inference processor market today, while Nvidia has the lion's share for the training-related applications. Nonetheless, both are expected to be challenged by well-established players like AMD, as well as start-ups with innovative hardware such as Kalray or Graphcore.

According to a McKinsey study, AI-related computing hardware will increase by about 18% annually, to hit nearly USD20bn by 2025. Cloud computing hardware will account for 75% of all demand, growing at approximately 12% CAGR. It will be dominated by a handful of companies that either can bear the costs of expensive leading-edge chip designs or have achieved breakthrough innovations to offer completely different architectures.

Looking at the edge, AI computing hardware is expected to grow from around USD100m in 2017 to USD5.5bn in 2025, or a staggering annual growth rate of more than 50%. Unlike the cloud, the market will be more fragmented, especially at the inference level, because the computing power requirement is lower compared to the cloud, and leading microcontrollers Industrial Design Manufacturers (IDMs) can address this market.

FIG.21: COMPUTING HARDWARE VENDORS THAT WILL BENEFIT FROM FAST-GROWING DEMAND FOR AI APPLICATIONS

*AI computing hardware market data estimated by 2025



CLOUD

TRAINING

INFERENCE

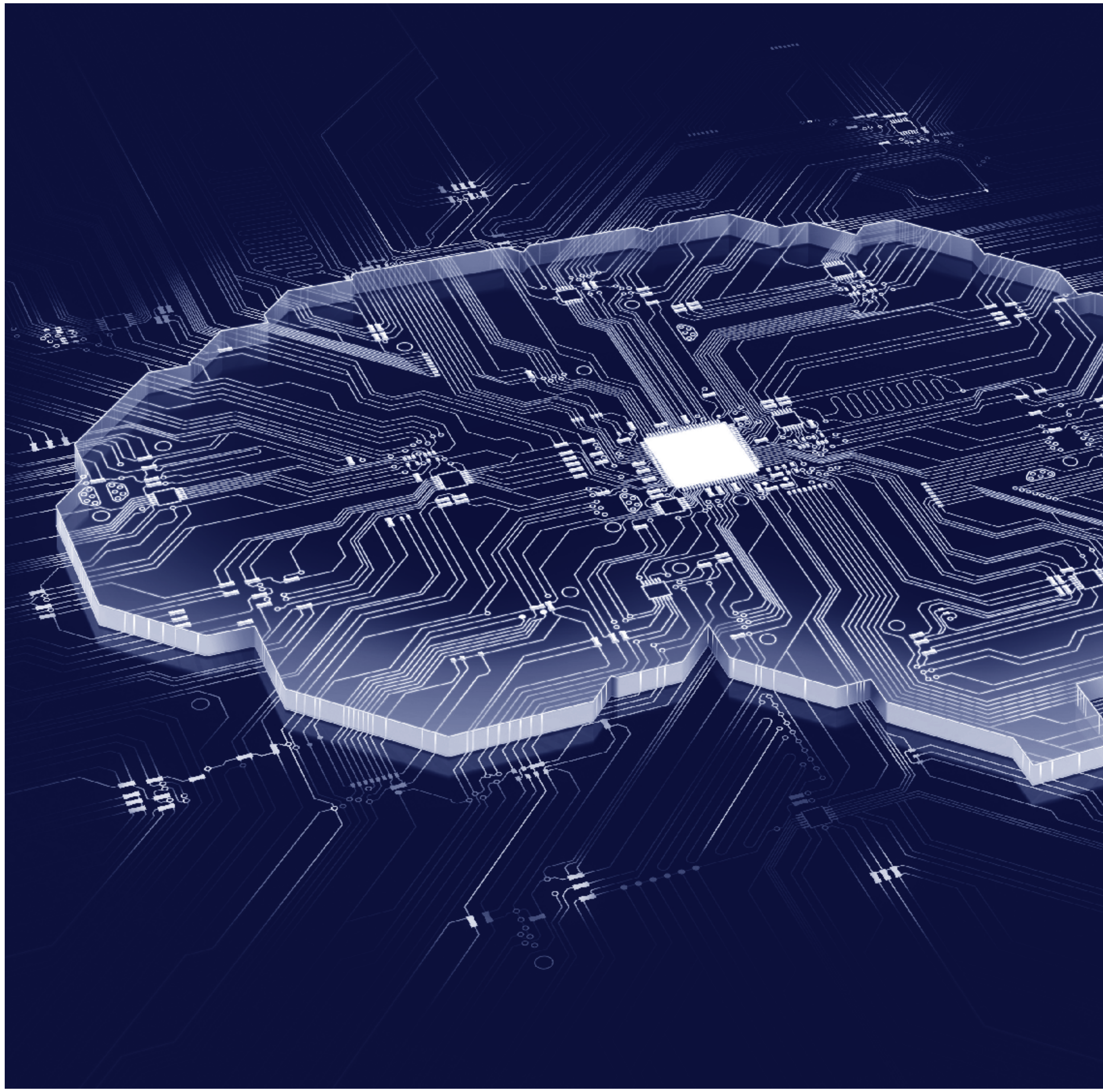
EDGE

$4-5BN*
+20% CAGR17-25

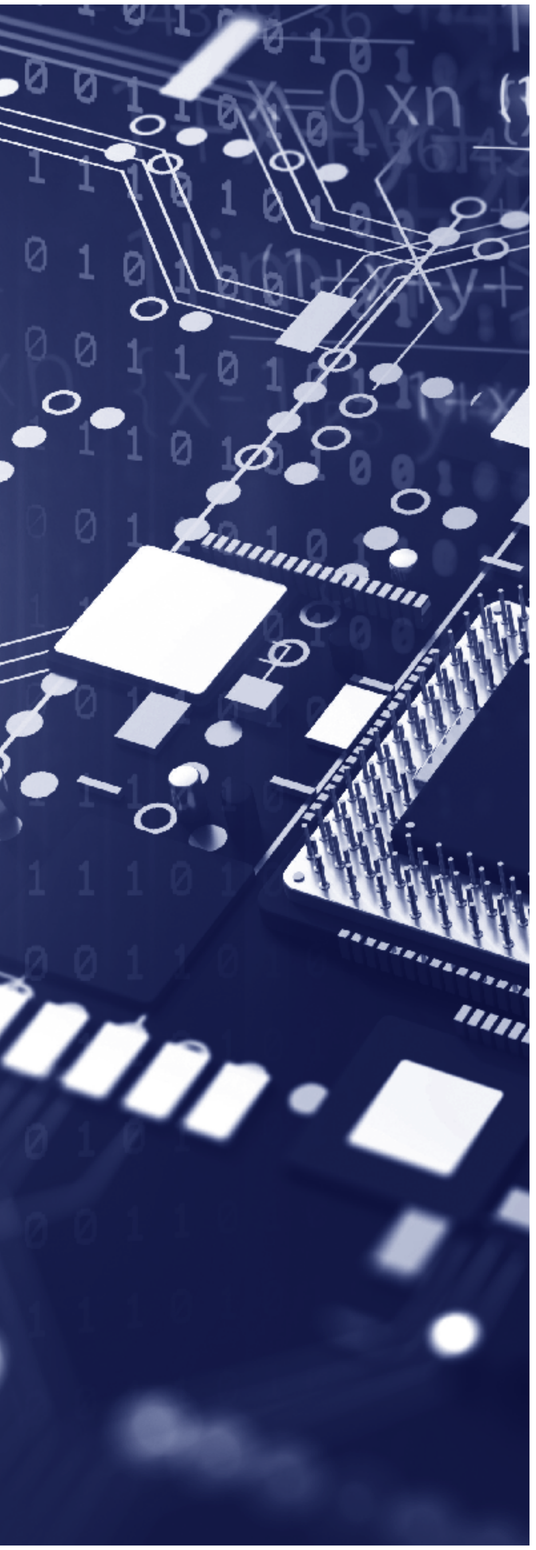
$9-10BN*
+10% CAGR17-25

$1-1-1.5BN*
+35%

$4-4.5BN*
+60% CAGR17-25



Source: McKinsey; Bryan, Garnier & Co



Industrials are looking for the right embedded computing partners

What is an embedded computing system?

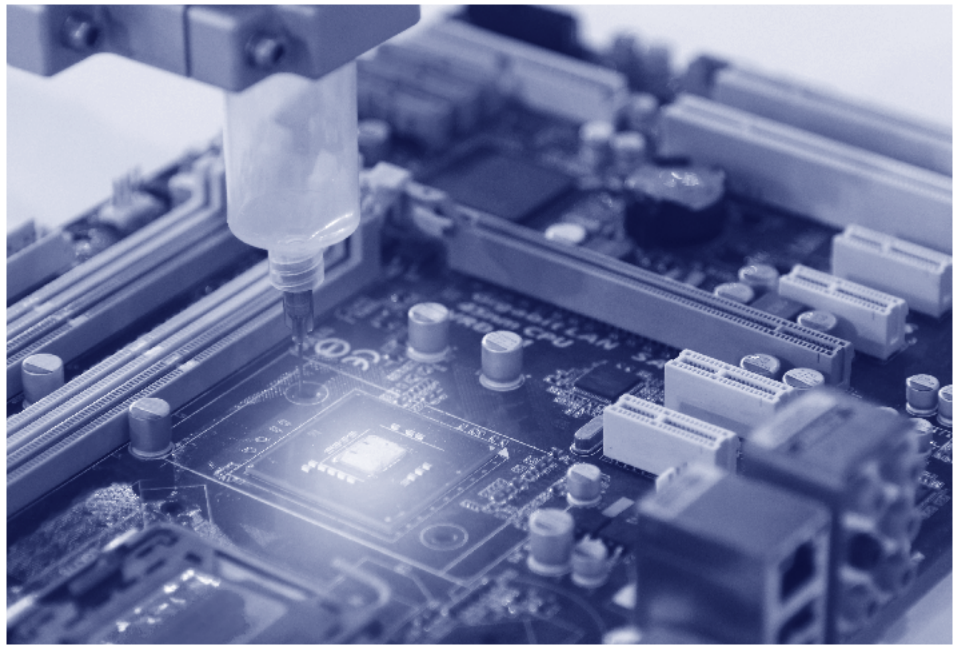
Embedded computing or embedded system refers to a controller designed to automate the mechanical or electrical systems that are commonly found in industrial, consumer, automotive, medical or military applications.

An embedded system consists of different semiconductor hardware including computing chips, sensors, actuators, connectivity, peripherals etc., as well as a real-time operating system (RTOS).

In general, embedded systems are built around MCUs or CPUs and dedicated software with the purpose of executing a pre-installed set of instructions to perform a very specific task.

However, as we said in the previous chapter, edge computing brings more diversity to the hardware choice, and this has also been reflected in embedded computing solutions.

FIG.23: EXAMPLE OF AN EMBEDDED SYSTEM

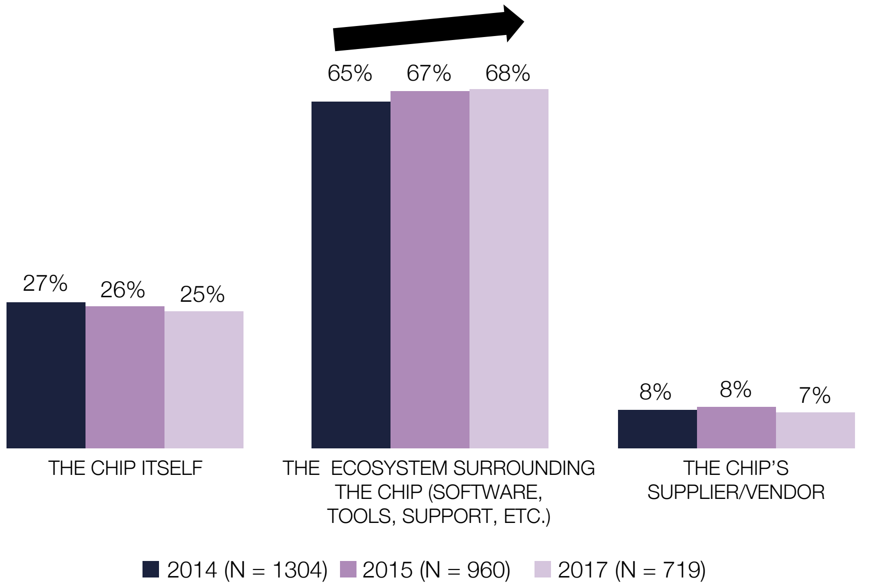


Source: Advantech

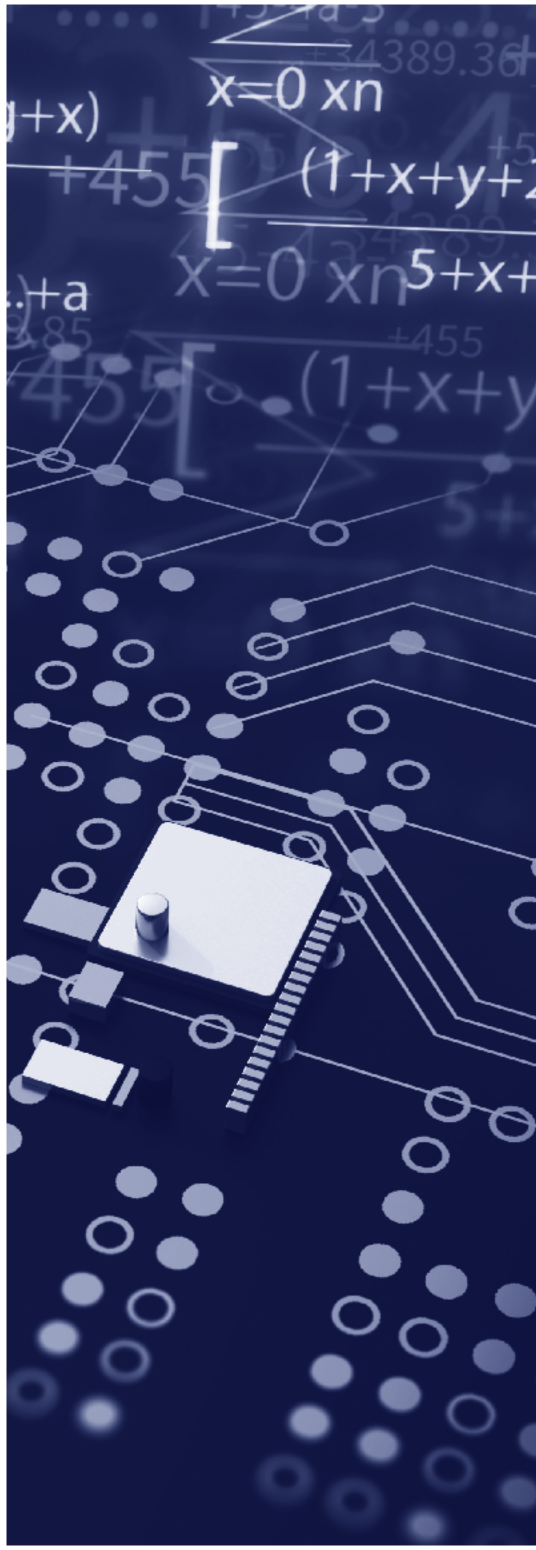
The ecosystem is a key consideration

Embedded systems solutions can be classified according to the supplier of the processor. Well-known suppliers include Intel, Microchip, NXP, Renesas, STMicroelectronics, and Texas Instruments. Each processor is designed based on an instruction set architecture (ISA) which is either licensed from Intel's x86 or ARM, or based on an open source ISA such as RISC-V, or even a proprietary architecture. The choice of the processor is central, as it will define not only the raw performance and power consumption but also the entire ecosystem of the embedded computing hardware, including the roadmap of new technologies, software and hardware support, compatibility and connectivity with other systems. In the case of industrial, medical or automotive applications where the operating environment involves multiple interactions and end-points, and for which the average lifespan is long, the choice of the ecosystem cannot be understated. An EETimes study reveals that more than two-thirds of engineers surveyed consider the ecosystem surrounding the chip as the most important factor in choosing an embedded system.

FIG.24: THE ECOSYTEM IS THE MOST IMPORTANT FACTOR WHEN CHOOSING A MICROPROCESSOR



Source: EEtimes embedded system study 2017

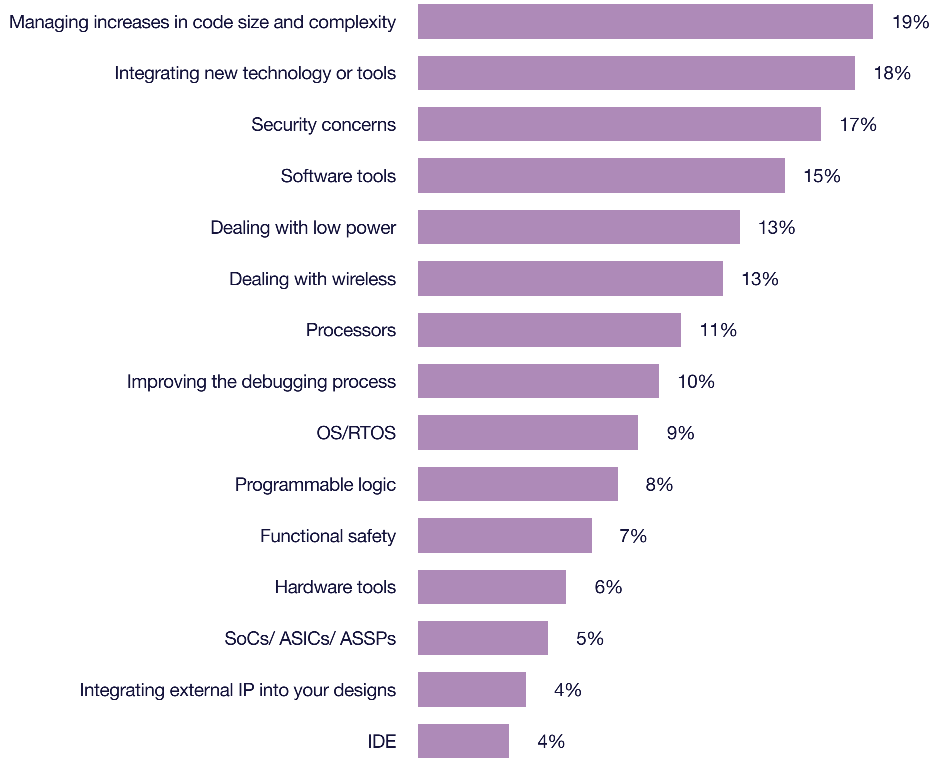


AI to bring more challenges for system integration

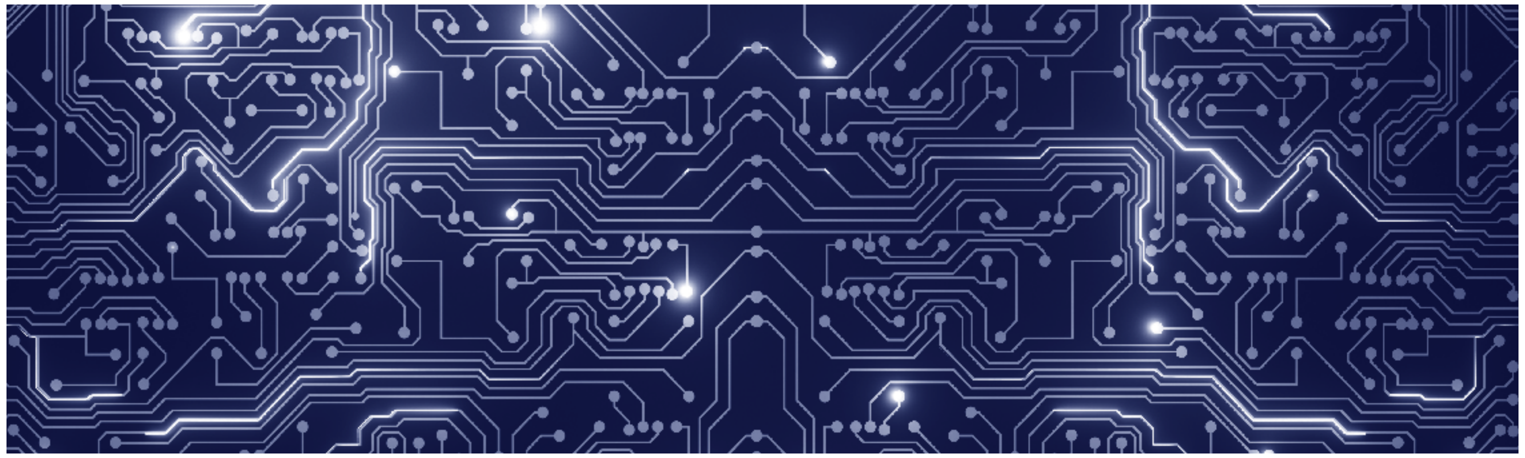
IoT has brought additional technology layers, especially on the wireless side, with more communication interfaces and protocols (Wi-Fi, Bluetooth, LoRA, SigFox, Zigbee, Z-Wave…)
as well as new concerns about power consumption and security. Additionally, the integration complexity of CPUs and MCUs have developed expotentially, now requiring highly specialised equipment and experienced engineering staff.

On top of that, companies that are pursuing the integration of edge computing to enhance their productivity will face even more challenges with new AI-related hardware, framework and software. For example, there are several deep learning frameworks to take into consideration in AI such as Google's TensorFlow, Caffe, Microsoft's CNTK, MXNet or PyTorch.

FIG.25: MAIN CHALLENGES THAT COMPANIES FACE

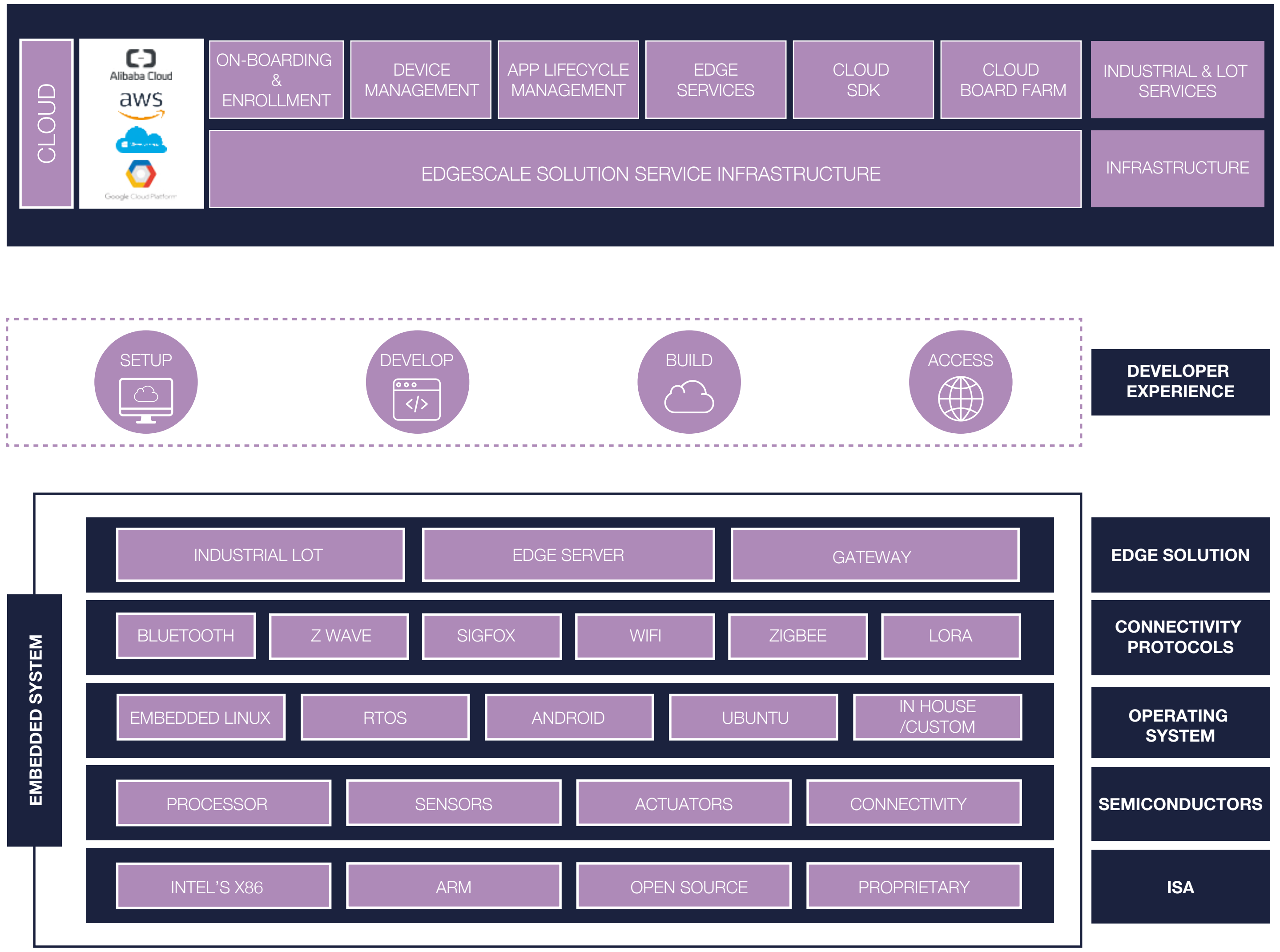


Source: EE Times



In industrial IoT, the location of the edge is also more complex than for consumer devices. In an industrial zone with multiple plants and equipment, edge computing can be either in an on-premise server, a gateway or the device itself. Therefore, depending on the form factor, the requirement of the embedded systems could differ dramatically.

FIG.26: MULTI-LAYER EDGE COMPUTING ARCHITECTURE



Source: NXP; Bryan, Garnier & Co

Partnering with embedded computing experts to accelerate edge computing

As most embedded systems are used in applications that require high precision and quality in their operations such as industrial, automotive or medical, companies tend to avoid off-the-shelf hardware. According to a market study run by EE Times in 2017, 80% of the companies surveyed primarily build or subcontract custom tailored embedded systems in order to have optimal designs and performance in accordance to specific needs.

We believe that this trend will increase over time and that the share of design that is subcontracted – around 25% in 2015 according to VDC Research – will also tend to increase.

Considering how complex the ecosystem around embedded systems for IoT and edge computing has become, OEMs and ODMs are likely to increase their outsourced embedded design share of wallet. Leading companies that can take advantage of this trend are Taiwan-based ADLINK and Advantech, or European players such as congatec or Kontron (owned by S&T Group).

These companies help to design and develop embedded computing solutions tailored for specific end-market, application, and customer requirements. Outsourced embedded computing companies have different partnerships with leading processor designers and help their customers to develop and design embedded computing systems to be integrated seamlessly to their ecosystem, minimising hardware and software engineering costs as well as maintenance. In addition, the accelerating pace of technology innovations requires companies to reduce the time to market for new products, which should drive demand for outsourcing.

According to IHS Markit, the overall market size for embedded boards was around USD2.5bn in 2016 and is expected to grow at 6.3% CAGR between 2015 and 2020 to USD3.4bn. While there is a contraction in legacy products addressing markets such as telecoms, embedded system designers with a strong exposure to IoT innovation and edge computing applications, especially in the industrial automation sector, will significantly outgrow the overall market.

FIG.27: PACE OF NEW TECHNOLOGIES DEMAND TO BE MORE REACTIVE

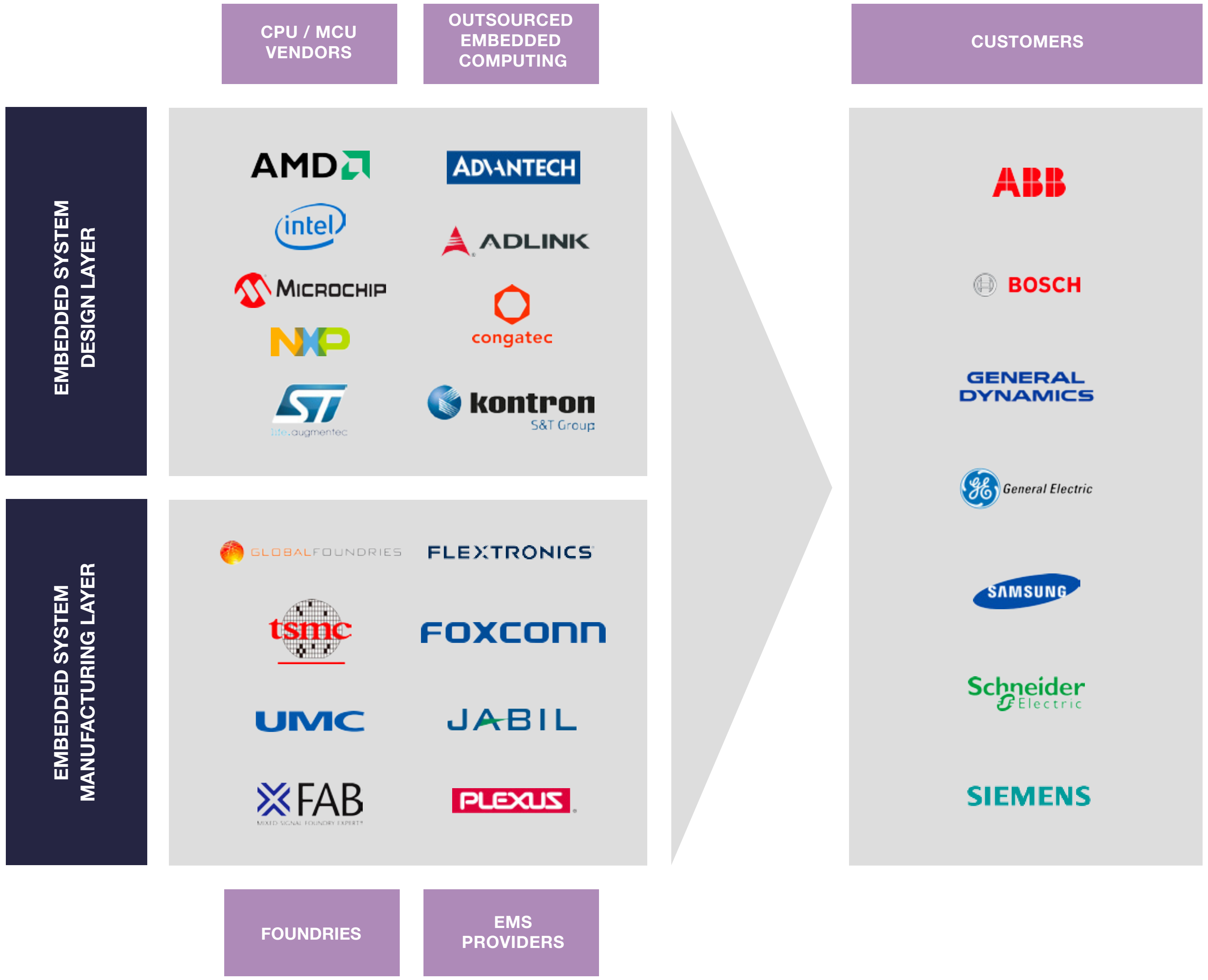
TECHNOLOGY COMPLEXITY IS INCREACING WHILE DESIGN CYCLES ARE SHORTENING

	WAS	NOW
TECHNOLOGY REFRESH CYCLE	18 MONTHS	3-6 MONTHS
CUSTOMER REFRESH CYCLE	8 MONTHS	1-3 MONTHS

TIME TO MARKET MORE IMPORTANT THAN EVER

Source: AVNET

FIG.28: EMBEDDED DESIGN SUPPLY CHAIN


CPU / MCU VENDORS

OUTSOURCED EMBEDDED COMPUTING

CUSTOMERS

EMBEDDED SYSTEM DESIGN LAYER

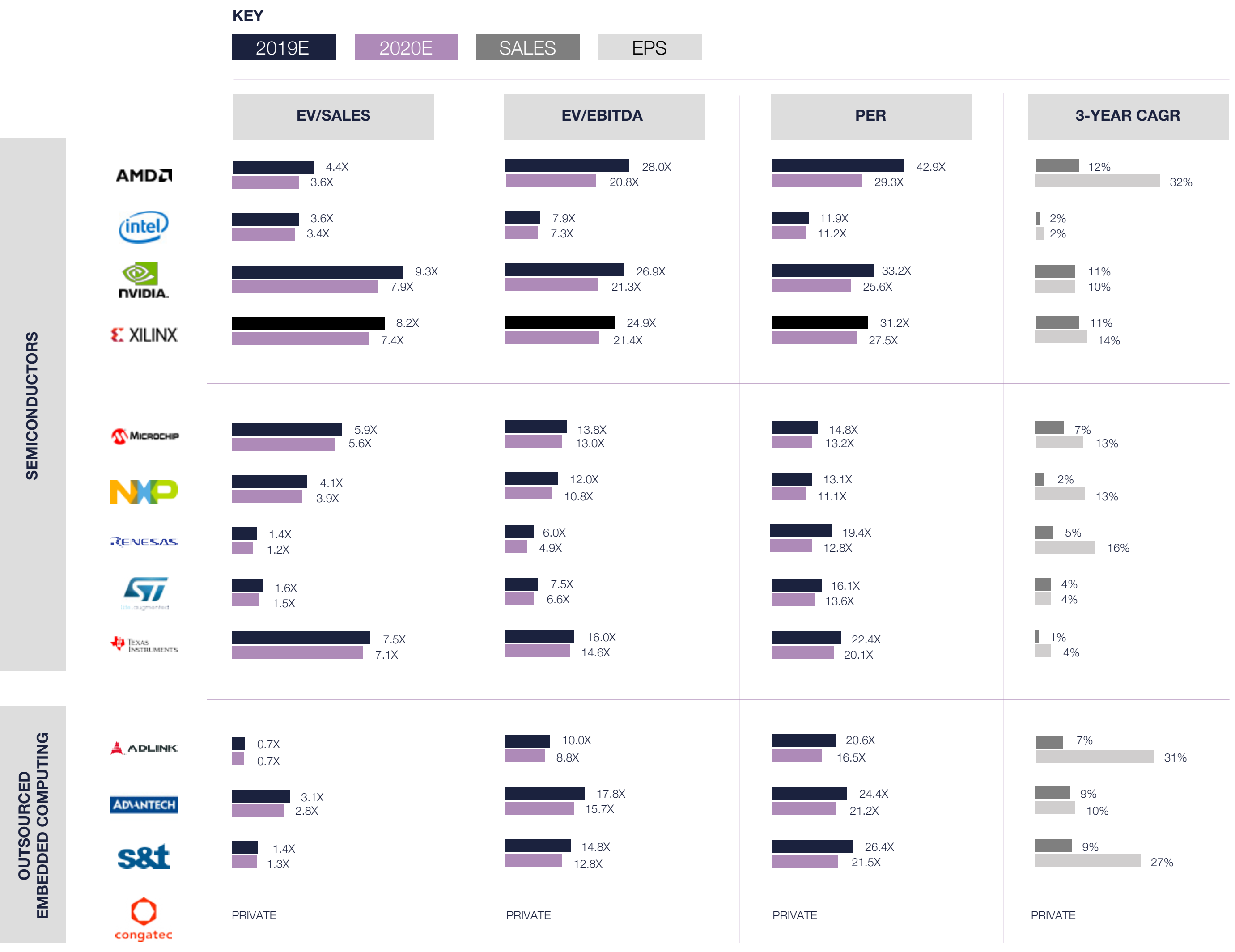
EMBEDDED SYSTEM MANUFACTURING LAYER

FOUNDRIES

EMS PROVIDERS

Source: Bryan, Garnier & Co

FIG.29: MULTIPLES TABLE


KEY

		EV/SALES		EV/EBITDA		PER		3-YEAR CAGR	
		2019E	2020E	2019E	2020E	2019E	2020E	SALES	EPS
SEMICONDUCTORS	AMD	4.4X	3.6X	28.0X	20.8X	42.9X	29.3X	12%	32%
	intel	3.6X	3.4X	7.9X	7.3X	11.9X	11.2X	2%	2%
	NVIDIA	9.3X	7.9X	26.9X	21.3X	33.2X	25.6X	11%	10%
	XILINX	8.2X	7.4X	24.9X	21.4X	31.2X	27.5X	11%	14%
	MICROCHIP	5.9X	5.6X	13.8X	13.0X	14.8X	13.2X	7%	13%
	NXP	4.1X	3.9X	12.0X	10.8X	13.1X	11.1X	2%	13%
	RENESAS	1.4X	1.2X	6.0X	4.9X	19.4X	12.8X	5%	16%
	ST life.augmented	1.6X	1.5X	7.5X	6.6X	16.1X	13.6X	4%	4%
	Texas Instruments	7.5X	7.1X	16.0X	14.6X	22.4X	20.1X	1%	4%
OUTSOURCED EMBEDDED COMPUTING	ADLINK	0.7X	0.7X	10.0X	8.8X	20.6X	16.5X	7%	31%
	ADVANTECH	3.1X	2.8X	17.8X	15.7X	24.4X	21.2X	9%	10%
	s&t	1.4X	1.3X	14.8X	12.8X	26.4X	21.5X	9%	27%
	congatec	PRIVATE		PRIVATE		PRIVATE		PRIVATE	

Source: Reuters, Bryan, Garnier & Co

4. Conclusion

In this report, we have demonstrated the need for a wide variety of sectors including automotive, industrial automation and healthcare to use AI to enhance operational efficiency and create new revenue streams. However, in order to cope with a continuing increase in data volume and future bandwidth constraints, as well as real-time operation requirements and cost efficiencies from these new applications, the network infrastructure needs to become more distributed toward the edge of the node.

Cloud computing is characterised by the need to have best-in-class processors using the most advanced semiconductor technology nodes. For this reason, only a handful of companies, including Intel and Nvidia, control this market. Unlike the cloud, edge computing requires much less computing performance and therefore the market structure should be more fragmented, with traditional microcontroller players taking the lion's share next to CPU providers.

We do not see one player standing out from the crowd at the edge. In the case of industrial, medical or automotive applications where the operating ecosystem involves multiple interactions and end-points, there is a need for tight integration both in terms of hardware and software, which means the hardware offer should be as diversified as the number of end markets.

Finally, our analysis has highlighted the more important role that outsourced embedded computing players will have to play to help industrials to develop and accelerate the time to market of edge computing technologies.

BRYAN, GARNIER & CO

White Paper Authors

Olivier Beaudouin
Partner, Investment Banking
Technology & Smart Industries
obeaudouin@bryangarnier.com

Frédéric Yoboué
Equity Research Analyst
Semiconductors
fyoboue@bryangarnier.com

Frans-Matthis Pleie
Director
Investment Banking
fpleie@bryangarnier.com

Technology Team

INVESTMENT BANKING

PARIS

Greg Revenu
Managing Partner
Technology
grevenu@bryangarnier.com

Olivier Beaudouin
Partner, Investment Banking
Technology & Smart Industries
obeaudouin@bryangarnier.com

Guillaume Nathan
Partner, Digital Media
& Business Services
gnathan@bryangarnier.com

Thibaut De Smedt
Partner, Application
Software & IT Services
tdesmedt@bryangarnier.com

Philippe Patricot
Managing Director,
Technology
ppatricot@bryangarnier.com

MUNICH

Falk Müller-Veerse
Partner
Technology
fmuellerveerse@bryangarnier.com

EQUITY RESEARCH ANALYST TEAM

Olivier Pauchaut
Managing Director
Financials & Fintech
opauchaut@bryangarnier.com

Thomas Coudry
Managing Director
Telecoms & Media
tcoudry@bryangarnier.com

Eric Lemarié
Smart Industries
elemarie@bryangarnier.com

Xavier Regnard
Smart Industries
xregnard@bryangarnier.com

Gregory Ramirez
Software & IT Services
gramirez@bryangarnier.com

David Vignon
Software
dvignon@bryangarnier.com

Fréderic Yoboué
Semiconductors
fyoboue@bryangarnier.com

EQUITY CAPITAL MARKETS

Pierre Kiecolt-Wahl
Partner, Head of ECM
pkiecoltwahl@bryangarnier.com

EQUITY DISTRIBUTION

Nicolas-Xavier de Montaigut
Partner, Head of Distribution
nxdemontaigut@bryangarnier.com

Nicolas d'Halluin
Partner, Head of US Distribution
ndhalluin@bryangarnier.com

Corporate Transactions

Bryan, Garnier & Co leverage in-depth sector expertise to create fruitful and long lasting relationships between investors and European growth companies.

metrologic group	xerox	SMART ME UP	AUTOFORM Forming Reality	REstore
Strategic Investment	Acquired by	Acquired by	Acquired by	Acquired by
Astorg	NAVER	MAGNETI MARELLI	Astorg	centrica
Undisclosed	Undisclosed	Undisclosed	Undisclosed	€70 000 000
Sole Advisor to the Buyer	Sole Advisor to the Seller	Sole Advisor to the Sellers	Advisor to the Buyer	Sole Advisor to the Sellers

About Bryan, Garnier & Co

Bryan, Garnier & Co is a European, full service growth-focused independent investment banking partnership founded in 1996. The firm provides equity research, sales and trading, private and public capital raising as well as M&A services to growth companies and their investors. It focuses on key growth sectors of the economy including Technology, Healthcare, Consumer and Business Services. Bryan, Garnier & Co is a fully registered broker dealer authorised and regulated by the FCA in Europe and the FINRA in the U.S. Bryan, Garnier & Co is headquartered in London, with additional offices in Paris, Munich, Zurich, Stockholm, Oslo and Reykjavik as well as New York and Palo Alto. The firm is a member of the London Stock Exchange.

Bryan, Garnier & Co Technology Equity Research Coverage

adyen	altice	altran	ALTEN	ASM INTERNATIONAL
ASML	Atos	Besi	bouygues TELECOM	Capgemini
DASSAULT SYSTEMES	dialog SEMICONDUCTOR	easyVISTA™	ekinops	indra
iliad	infineon	ingenico GROUP		orange™
SafeCharge	sage	SAP®	Schneider Electric	software AG
soitec	sopra steria	SWORD ACTIVE RISK	ST life.augmented	TEMENOS The software specialist for banking and finance
TIM	wirecard	worldline e-payment services		

7 Analysts | 50+ Stocks Covered

With more than 150 professionals based in London, Paris, Munich, Stockholm, Oslo and Reykjavik as well as New York and Palo Alto, Bryan, Garnier & Co combines the services and expertise of a top-tier investment bank with a long-term client focus.

Edge Computing empowers a new set of applications requiring lower latency and more privacy

BRYAN, GARNIER & CO

LONDON
Beaufort House
15 St. Botolph Street
London, EC3A 7BB
UK

T: +44 (0) 20 7332 2500
F: +44 (0) 20 7332 2559

Authorised and regulated by the Financial Conduct Authority (FCA)

PARIS
26 Avenue des Champs Elysées
75008 Paris
France

T: +33 (0) 1 56 68 75 00
F: +33 (0) 1 56 68 75 01

Regulated by the Financial Conduct Authority (FCA) and the Autorité de Contrôle prudential et de resolution (ACPR)

MUNICH
Widenmayerstrasse 29
80538 Munich
Germany

T: +49 89 242 262 11
F: +49 89 242 262 51

NEW YORK
750 Lexington Avenue
New York, NY 10022
USA

T: +1 (0) 212 337 7000
F: +1 (0) 212 337 7002

FINRA and SIPC member

STOCKHOLM
Malmskillandsgatan 32, 6th Floor
114 55 Stockholm
Sweden

T: +46 706 337 503

OSLO
Grundingen 2
0250 Oslo
Norway

T: +47 22 01 64 00

Regulated by the Norwegian Financial Supervisory Authority (Norwegian FSA)

REYKJAVIK
Höfðatorg, Katrínartún 2
105 Reykjavik
Iceland

T: +354 554 78 00

PALO ALTO
394 University Avenue
Palo Alto
California (CA) 94 301
USA

T: +1 650 283 18 34

FINRA member

DISCLAIMER

This document is based on information available to the public and other sources deemed reliable. No representation or warranty, express or implied, is or will be made in relation to, and no responsibility or reliability is or will be accepted by Bryan Garnier & Company or any of its officers, employees or advisers as to the accuracy or completeness of this document or any other written or verbal information available to the recipient or its advisers. While all reasonable care has been taken to ensure that the facts stated are accurate and the opinions given are fair and reasonable, neither we nor any of our affiliated companies nor any of our, or their directors, representatives or employees, accepts responsibility or liability for any loss or expense arising directly or indirectly from the use of this document or its or its contents. This document is not and should not be construed as an offer, or a solicitation of any offer, to buy or sell securities. Bryan, Garnier & Co is authorised and regulated by the Financial Conduct Authority (FCA) in the United Kingdom.