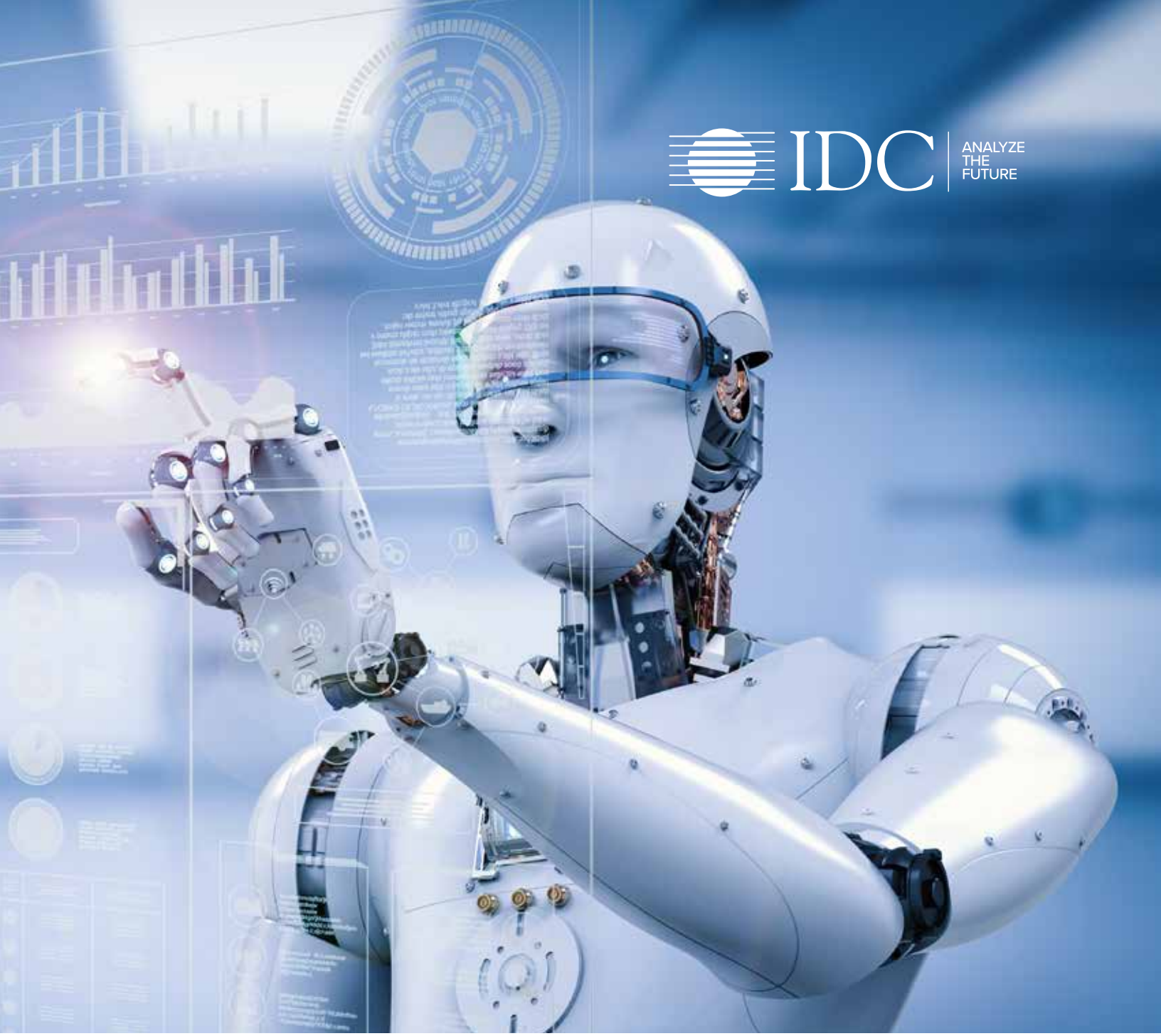


Embedded Artificial Intelligence: Reconfigurable Processing Accelerates AI in Endpoint Systems for the OT Market

Mario Morales | *November 2018*

An IDC White Paper, Sponsored by: Renesas Electronics Corporation

IN THIS WHITE PAPER

This IDC white paper provides a perspective on the transformation of embedded semiconductors as artificial intelligence (AI) scales and enables intelligent and adaptable endpoint systems that leverage real-time data to improve efficiency, lower fixed costs, operate autonomously, scale insights and, ultimately, enable companies to monetize services across a broad set of industry segments. The document also explores the key technologies necessary to enable the adoption of AI and emerging usage models that illustrate the potential that AI brings to embedded systems and the operational technology (OT) market. The OT market is poised to undergo significant change to become the next large and sustainable market that technology suppliers and incumbent system OEMs capitalize on.

SITUATION OVERVIEW

Much has been written about the explosive transformation we are witnessing in the cloud and datacenter environment over the past decade. Computing workloads continue to evolve at a breakneck pace, with each additional square foot added by cloud service providers and large enterprises. As the scale of cloud service providers and enterprises grows, so do the complexity and amount of data that is generated and used to drive transactions and run a large sustainable business. As sensors and processors have improved in terms of cost, capability, and performance, the same requirements of both technologies have begun to distribute further into embedded systems.

Industry Background

Over the past decade, the IT industry has been able to scale faster and larger by using more capacity to address the demand for performance and introduction of services. Hardware virtualization, new computing business models, and raw performance from semiconductor and system design have enabled the ability to deliver computing on demand and connections to cloud platforms anywhere on the planet. There are billions of transactions and hundreds of thousands of applications that have been able to leverage the computing power and the business model of the cloud. Today, the annual cadence of investment in cloud infrastructure has reached a run rate of close to $100 billion per annum, representing the fastest-growing part of overall IT spending.

The expansion of IP technologies continues forward, connecting billions of embedded systems in an expanding array of applications. Embedded systems are used to represent isolated devices that were intended to operate on their own, with a limited range of functions and limited understanding of their operating environment. Any ability to adjust operations to reflect local conditions had to be programmed, and programming post installation was limited or difficult.

However, the expansion of connectivity and internet technologies, the emergence of intelligent systems (embedded systems equipped with more powerful processors and robust operating systems), and the evolution of sensor technologies have created new innovation. Now, billions of endpoint devices and systems can be remotely managed, thereby capturing, generating, analyzing, and sending data back to the cloud.

IDC expects that, by 2025, the amount of data created globally will grow by a factor of 10 times ... More than half of all data created will come from endpoint devices.

The emergence of the cloud, the development of connected embedded systems, and the expanding reach of smartphones, tablets, and personal computers (PCs) have fueled a revolution in the creation and consumption of data. **IDC expects that, by 2025, the amount of data created globally will grow by a factor of 10 times** and reach 163ZB (1 trillion gigabytes). IDC expects that, in the same time period, **more than half of all data created will come from endpoint devices,** with the fastest-growing areas being embedded and Internet of Things (IoT).

Organizations will need to address the challenge of gaining useful insights from a growing ocean of devices and data.

There isn't enough computing power today to narrow the gap between the amount of data being created and the transformation required to make the data into useful and valuable insights for organizations. To address this growing divide, organizations have begun to harness artificial intelligence, especially deep learning methods. At the heart of this movement is the employment of training and using neural networks for inferencing operations. Neural networks leverage existing data to learn how to recognize some type of pattern in the data, and then the neural network applies this learned capability to make decisions based on new data sets with a high degree of accuracy. The learning can be applied to a broad set of problems, including image recognition, pattern identification in large sets of unstructured data, translation of complex languages in real time, recommendation of purchases for a customer, and even the prediction of a mobile user's behavior and location over time. The results can be applied to a broad set of tasks just like human beings would leverage their experiences, behavior, common logic, or sense across new and varied problems and situations.

Training and inferencing require a tremendous amount of compute today, but as frameworks and libraries become more established, training will be reused broadly, and inferencing will move closer to the endpoints where the data is being created to determine an outcome in microseconds (1,000,000 microseconds in a second) or in real time.

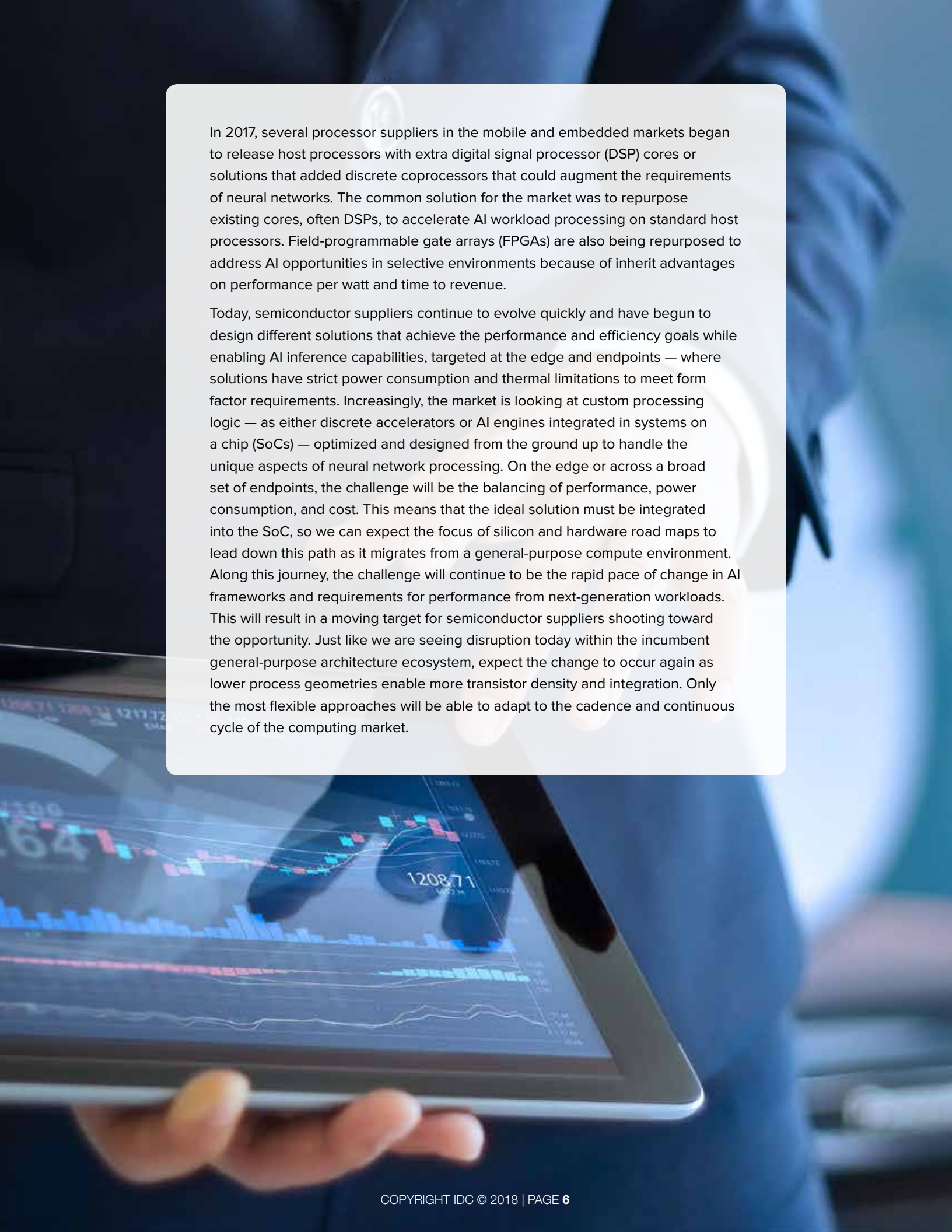
Artificial Intelligence Requires Re-Architecting of Computing Systems

The enormous amount of data being generated, processed, transmitted, stored, and analyzed demands that system companies and technology suppliers rethink how to build next-generation platforms beyond features and specs and toward addressing AI workloads and applications such as image recognition, natural machine translation, recognition of patterns and behavior, and predictive analytics. This reorientation is moving from using only general-purpose computing (microprocessor units [MPUs], which today account for the majority of processing) to a combination of large MPUs and accelerators and eventually to dedicated and custom architectures redesigned to address AI processing exclusively.

Large enterprises have been reluctant to change but are now being forced to reevaluate how to deploy their computing assets and augment operational requirements using AI. Cloud service providers, on the other hand, have been at the forefront of high-performance computing over the past decade, investing heavily in capacity, building large scalable and converging platforms, driving heterogeneous architectures, and using artificial intelligence to scale and transform their workloads and services.

On the edge and for endpoints such as wearables, industrial tools, and connected vehicles, we are witnessing a disruption as the use of various architectures becomes more commonplace to support AI. Only two years ago, AI workloads ran mostly on existing silicon in the system, such as the host processor, even though standard MPUs and graphics processing units (GPUs) are not very efficient at neural network processing. The development history for MPUs and GPUs has focused on greater raw compute performance per cycle and given power consumption levels. This path was primarily to adhere to growing demand for software applications. However, **AI and specifically neural network processing require less brute compute performance per cycle and instead benefit greatly when running multiple instances per cycle.** Today, the market is recognizing that traditional processing solutions are not going to address the requirements for AI moving forward.

AI and specifically neural network processing require less brute compute performance per cycle and instead benefit greatly when running multiple instances per cycle.



In 2017, several processor suppliers in the mobile and embedded markets began to release host processors with extra digital signal processor (DSP) cores or solutions that added discrete coprocessors that could augment the requirements of neural networks. The common solution for the market was to repurpose existing cores, often DSPs, to accelerate AI workload processing on standard host processors. Field-programmable gate arrays (FPGAs) are also being repurposed to address AI opportunities in selective environments because of inherit advantages on performance per watt and time to revenue.

Today, semiconductor suppliers continue to evolve quickly and have begun to design different solutions that achieve the performance and efficiency goals while enabling AI inference capabilities, targeted at the edge and endpoints — where solutions have strict power consumption and thermal limitations to meet form factor requirements. Increasingly, the market is looking at custom processing logic — as either discrete accelerators or AI engines integrated in systems on a chip (SoCs) — optimized and designed from the ground up to handle the unique aspects of neural network processing. On the edge or across a broad set of endpoints, the challenge will be the balancing of performance, power consumption, and cost. This means that the ideal solution must be integrated into the SoC, so we can expect the focus of silicon and hardware road maps to lead down this path as it migrates from a general-purpose compute environment. Along this journey, the challenge will continue to be the rapid pace of change in AI frameworks and requirements for performance from next-generation workloads. This will result in a moving target for semiconductor suppliers shooting toward the opportunity. Just like we are seeing disruption today within the incumbent general-purpose architecture ecosystem, expect the change to occur again as lower process geometries enable more transistor density and integration. Only the most flexible approaches will be able to adapt to the cadence and continuous cycle of the computing market.

Artificial Intelligence Is Compute Intensive and Will Be Adopted by Billions of Endpoints in the OT Market

Workloads in datacenters have been finely tuned to general-purpose MPU architectures supported by interconnect technologies and large amounts of storage and memory. This investment in semiconductors has been critical to the rapid innovation in semiconductor design and subsequent growth in the semiconductor market. Not only is demand from cloud infrastructure continuing to shatter growth expectations, but also billions of smartphones in our hands and pockets have made computing and data access ubiquitous.

As AI workload demands and applications are better understood, companies will look to extend AI to non-IT industry segments. Being more agile and responsive to business change and addressing evolving customer requirements have been important in driving the digital transformation in IT and will be critical to the OT market. IDC expects the more fragmented but larger OT market will become a key growth area as systems connect, sense, and become more intelligent, thus enabling a sea of intelligent endpoint consumers and OT systems.

This sea of **embedded and IoT devices**, which excludes PCs, servers, smartphones, tablets, and smart cards, is expected to reach over **8 billion unit shipments by 2022** at a CAGR of 8%. Each of these endpoints will be more intelligent, personal, and empowered by embedded AI and will surround us. We expect that, in five years, **20%+ of the cloud will be used to support the edge and the billions of endpoints shipping annually across large number of industry segments.** There are already early investments, announcements, and momentum at large cloud service provider incumbents such as Google, which has developed custom silicon for the edge; Amazon Web Services (AWS), which is rolling out Greengrass to support data created at the edge; and Microsoft's Azure IoT, which is targeted at industrial and other non-IT industries. IBM, Oracle, and other vendors are also introducing their own platforms to support the edge and endpoint devices.

IDC expects that embedded AI will redefine markets and industries as it reshapes the way we derive value from data. The reach of embedded AI will only be bound by how quickly AI algorithms can be created and trained and how efficiently systems can prioritize the computing power necessary to address the growing tide of complex computational requirements at the endpoint. This will lead to next-generation semiconductors, hardware, software, and AI algorithms that will create disruptive change.

As large as the cloud infrastructure has become over the past 10 years, IDC believes we are still only at the beginning of a long journey as an even larger and broader OT market that has been deeply embedded in our lives pivots and changes traditional industries in radical ways with the adoption of AI. Industries like automotive, industrial, energy, and retail are all aligned to benefit from the application of embedded AI in their industries.

Operational technology industries are well established and span decades of development and structure. Collectively, these OT industries are larger than the IT market in terms of revenue, the installed base volume, and annual shipments. The majority of these traditional devices have been fixed in function, dedicated to monitoring and control, rooted in safety, and contained in large established vertical industries. **IDC estimated that the OT industry achieved $2.6 trillion in revenue in 2017, and the industry is expected to grow at a CAGR of 4% by 2022** (see Table 1).

Figure 1 provides revenue forecast for specific subsegments of the overall OT market. The segments include consumer, smart factory, and infrastructure, which collectively account for almost two-thirds of the overall OT market.

Table 1

Worldwide Operational Technology Systems Revenue, 2017 and 2022 ($T)

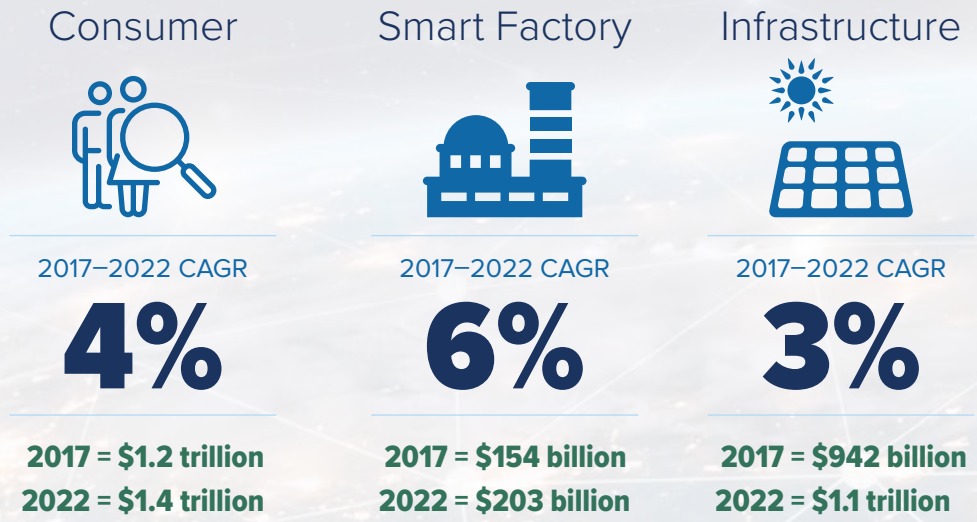
	2017	2022	2017–2022 CAGR (%)
Operational technology	2.6	3.1	4.0

Note: The OT market size is 2.5 times larger than the IT market and will grow at a CAGR of 4% by 2022. The consumer, infrastructure, and smart factory market segments account for two-thirds of the overall OT systems market in revenue.

Source: IDC, September 2018

Figure 1

Key Operational Technology Market Segment Revenue, 2017 and 2022



Note: The OT market size is 2.5 times larger than the IT market and will grow at a CAGR of 4% by 2022.

Source: IDC, 2018

Systems such as industrial tools (e.g., PLCs, wearables, home automation, video surveillance, smart meters, voice assistant speakers and earbuds, consumer and industrial robotics, digital signage, automotive ECUs, and ADAS subsystems) are all increasing in complexity and intelligence and the use of artificial intelligence to run inferencing and derive value from the real-time data being aggregated at each endpoint. That is the vision that makes the OT industry a prime candidate for disruption and potential growth for technologies suppliers.

Table 2 details the fastest-growing system submarket segments in the consumer, infrastructure, and smart factory industries.

Table 2

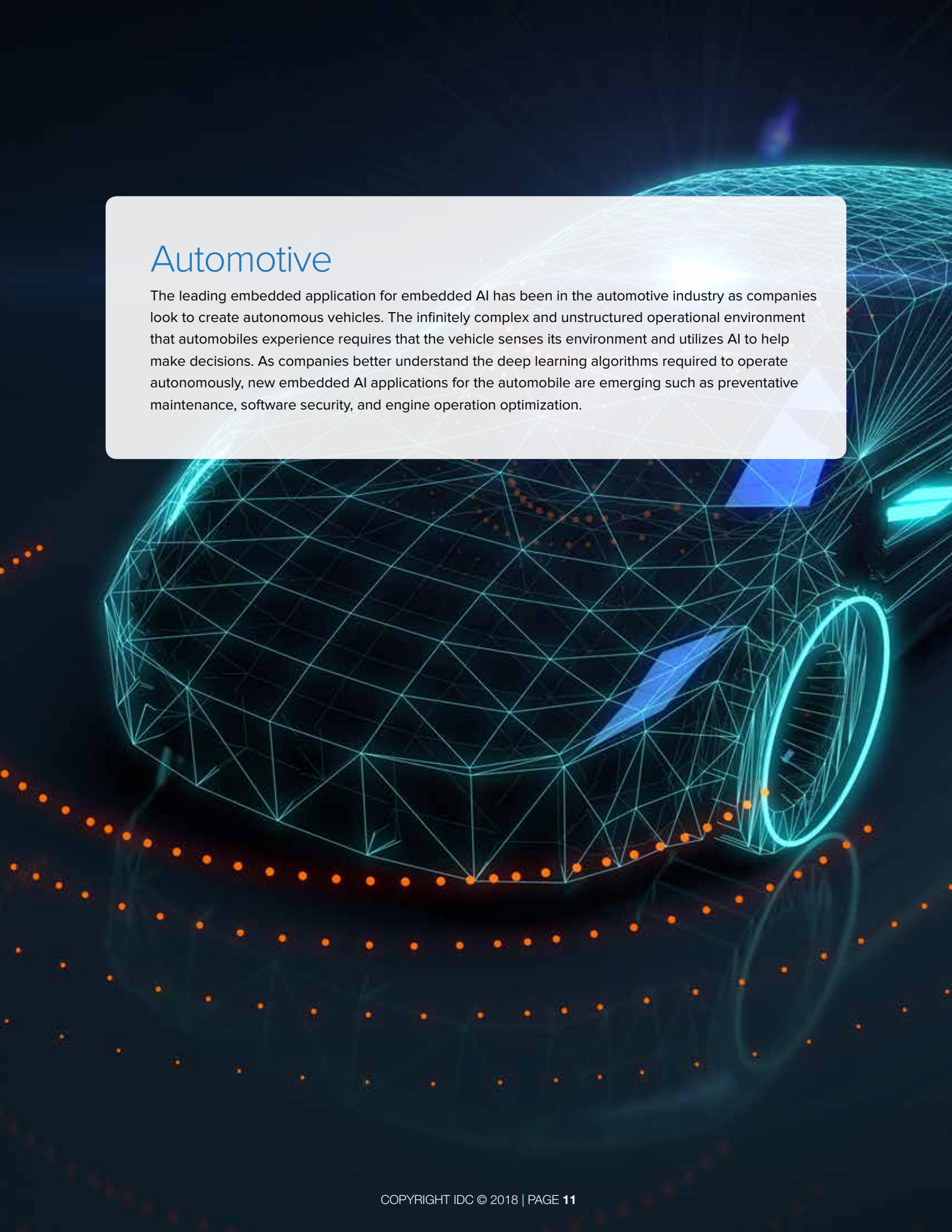
Top Market Growth Areas for Consumer, Smart Factory, and Infrastructure Industries

Consumer	2017–2022 CAGR (%)	Smart Factory	2017–2022 CAGR (%)	Infrastructure	2017–2022 CAGR (%)
Diagnostics and monitoring	25	PHV/EV industrial infrastructure	37	Time-sensitive networking	116
Drones	23	Digital signage	25	Industrial wearable	59
Healthcare gateway	17	Remote tracking	18	Industrial gateway	13
Wearable	14	Commercial motor vehicle telematics	13	Industrial automation	6
Smart home	5	Video surveillance	10	Functional safety	6

Note: The OT market size is 2.5 times larger than the IT market and will grow at a CAGR of 4% by 2022. The consumer, infrastructure, and smart factory market segments account for two-thirds of the overall OT systems market in revenue.

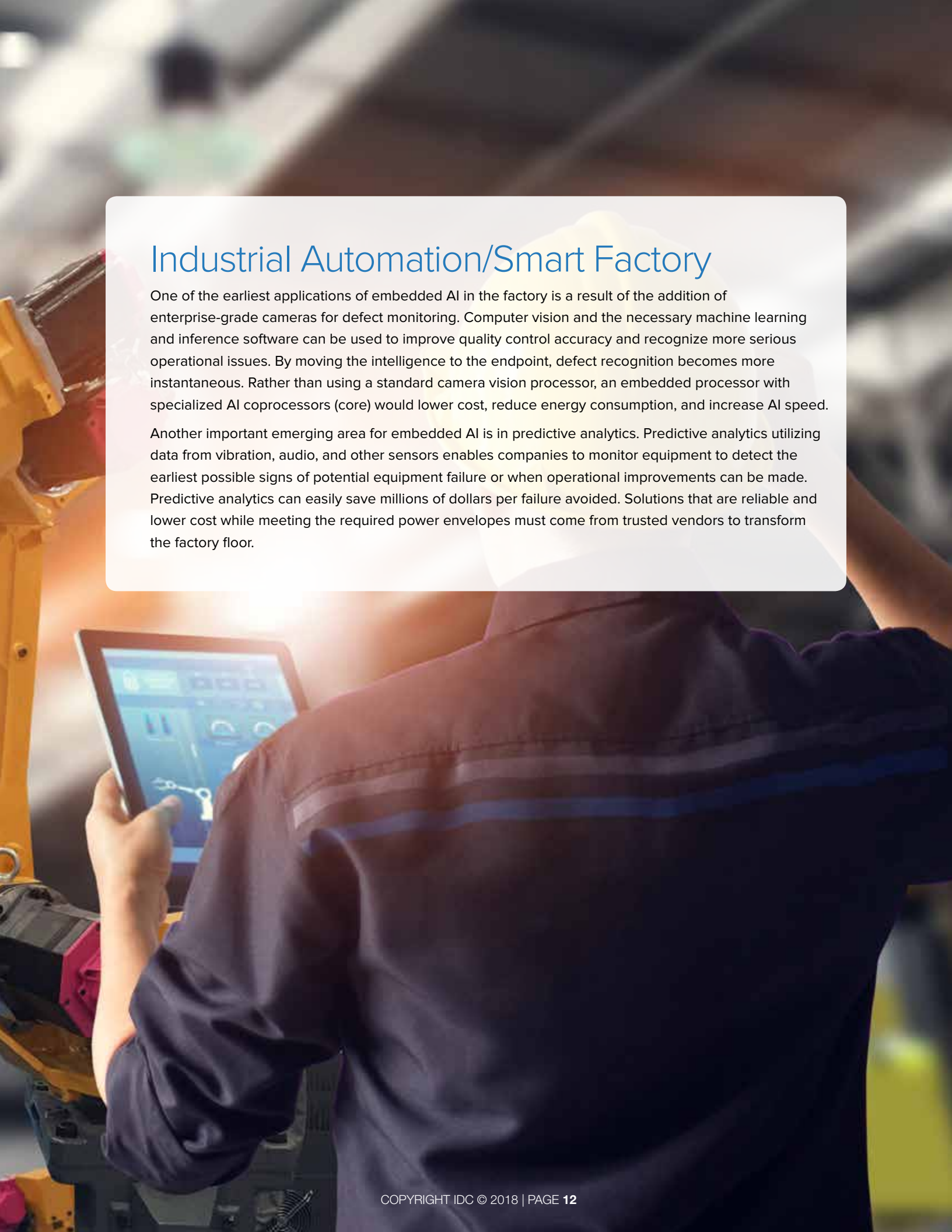
Source: IDC, September 2018

EXAMPLES OF ENDPOINT AI APPLICATIONS IN OT-SPECIFIC INDUSTRIES



Automotive

The leading embedded application for embedded AI has been in the automotive industry as companies look to create autonomous vehicles. The infinitely complex and unstructured operational environment that automobiles experience requires that the vehicle senses its environment and utilizes AI to help make decisions. As companies better understand the deep learning algorithms required to operate autonomously, new embedded AI applications for the automobile are emerging such as preventative maintenance, software security, and engine operation optimization.



Industrial Automation/Smart Factory

One of the earliest applications of embedded AI in the factory is a result of the addition of enterprise-grade cameras for defect monitoring. Computer vision and the necessary machine learning and inference software can be used to improve quality control accuracy and recognize more serious operational issues. By moving the intelligence to the endpoint, defect recognition becomes more instantaneous. Rather than using a standard camera vision processor, an embedded processor with specialized AI coprocessors (core) would lower cost, reduce energy consumption, and increase AI speed.

Another important emerging area for embedded AI is in predictive analytics. Predictive analytics utilizing data from vibration, audio, and other sensors enables companies to monitor equipment to detect the earliest possible signs of potential equipment failure or when operational improvements can be made. Predictive analytics can easily save millions of dollars per failure avoided. Solutions that are reliable and lower cost while meeting the required power envelopes must come from trusted vendors to transform the factory floor.

Service Robotics/Commercial Robotics

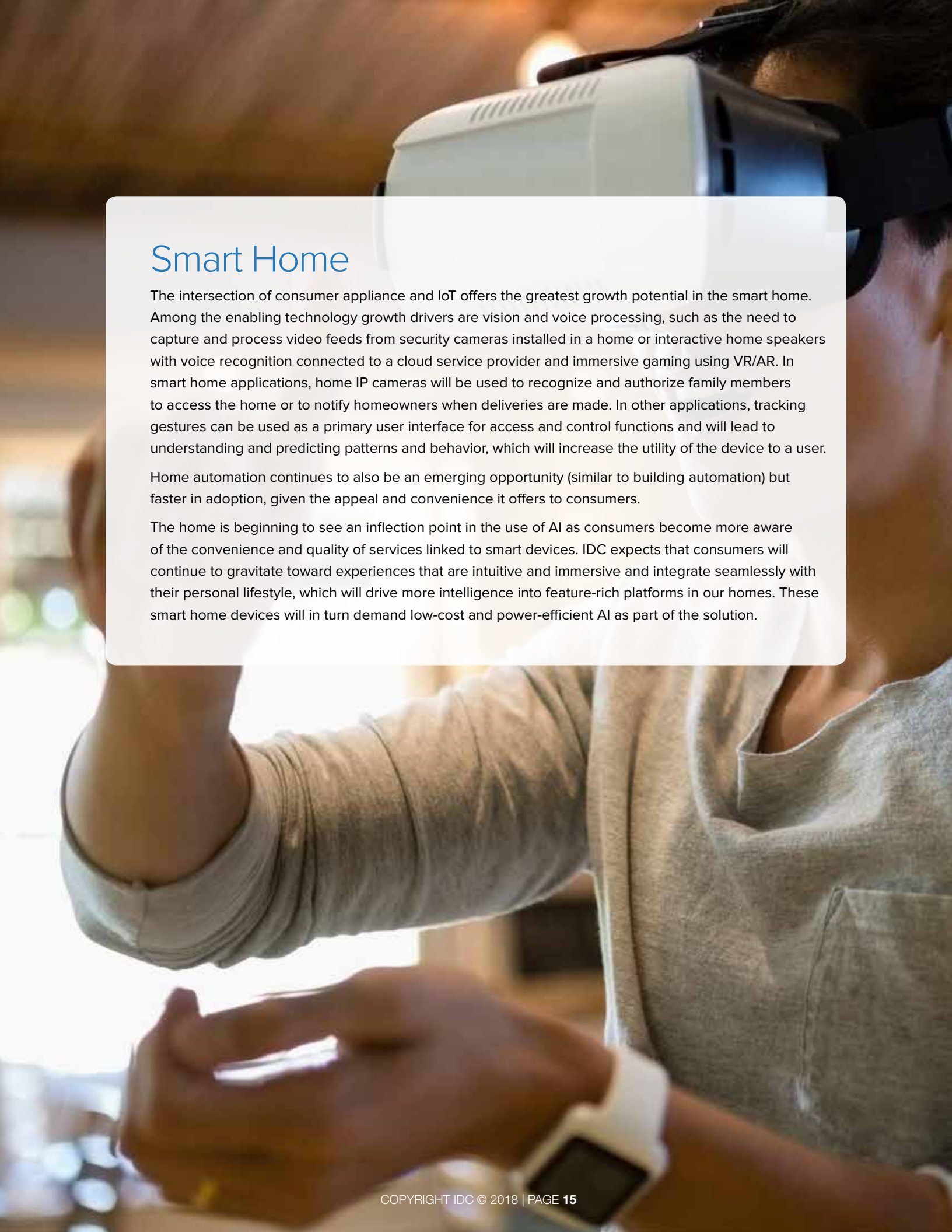
Commercial robots are used across a wide array of industries, but the commonality among them is that they will increasingly leverage embedded AI. This is especially the case as the commercial robotics market is seeing a shift from robotic product sales to a service model, sometimes referred to "robotics as a service." In the case of warehouse robots that move inventory around for storage or delivery, embedded AI can help the robots navigate faster, coordinate better with other robots moving around them, and find more efficient paths to take to get the work done. The better that these robots can learn on their own, the more rapidly all that learning can be aggregated and improved inference can be distributed to all the robots. As robots increasingly operate autonomously, reducing power consumption everywhere, from servos and actuators to all the embedded processing power using AI to make decisions, requires processors that can learn and make decisions within a specific operating power range.

Building Automation Systems

The need to reduce cost and lower the carbon footprint of buildings by optimizing the use of energy and addressing occupant requirements is driving investment in the automation and redesign of commercial and retail buildings. Today, building automation systems (BASs) are increasingly utilizing the cloud as facility managers rely on more sensor data to manage and control the building environment to provide more value to customers.

As we look forward, BASs will continue to introduce more intelligence and features across lighting systems, ventilation, and electrical systems to improve occupant comfort, increase automation, unify resources, optimize services, and reduce costs. Systems will have more sensors, require more monitoring and control, and generate additional real-time data that provides owners and managers with actionable insights on utilization. Building management systems, which already utilize high-end processors, will leverage AI to self-monitor and control building parameters across a site or multiple sites without any human interaction.

As building automation continues to evolve, expect to see more demand for flexible architectures that can be optimized for real-time workloads in order to improve efficiency. IDC expects this requirement will increase the use of silicon technology as it enables electrification across all major systems in a typical site including heating, ventilating, and air-conditioning (HVAC) systems; lighting; security surveillance; and central control units. This electrification will drive more durability and utility across smart buildings, which will result in an increase in the value of the commercial property.



Smart Home

The intersection of consumer appliance and IoT offers the greatest growth potential in the smart home. Among the enabling technology growth drivers are vision and voice processing, such as the need to capture and process video feeds from security cameras installed in a home or interactive home speakers with voice recognition connected to a cloud service provider and immersive gaming using VR/AR. In smart home applications, home IP cameras will be used to recognize and authorize family members to access the home or to notify homeowners when deliveries are made. In other applications, tracking gestures can be used as a primary user interface for access and control functions and will lead to understanding and predicting patterns and behavior, which will increase the utility of the device to a user.

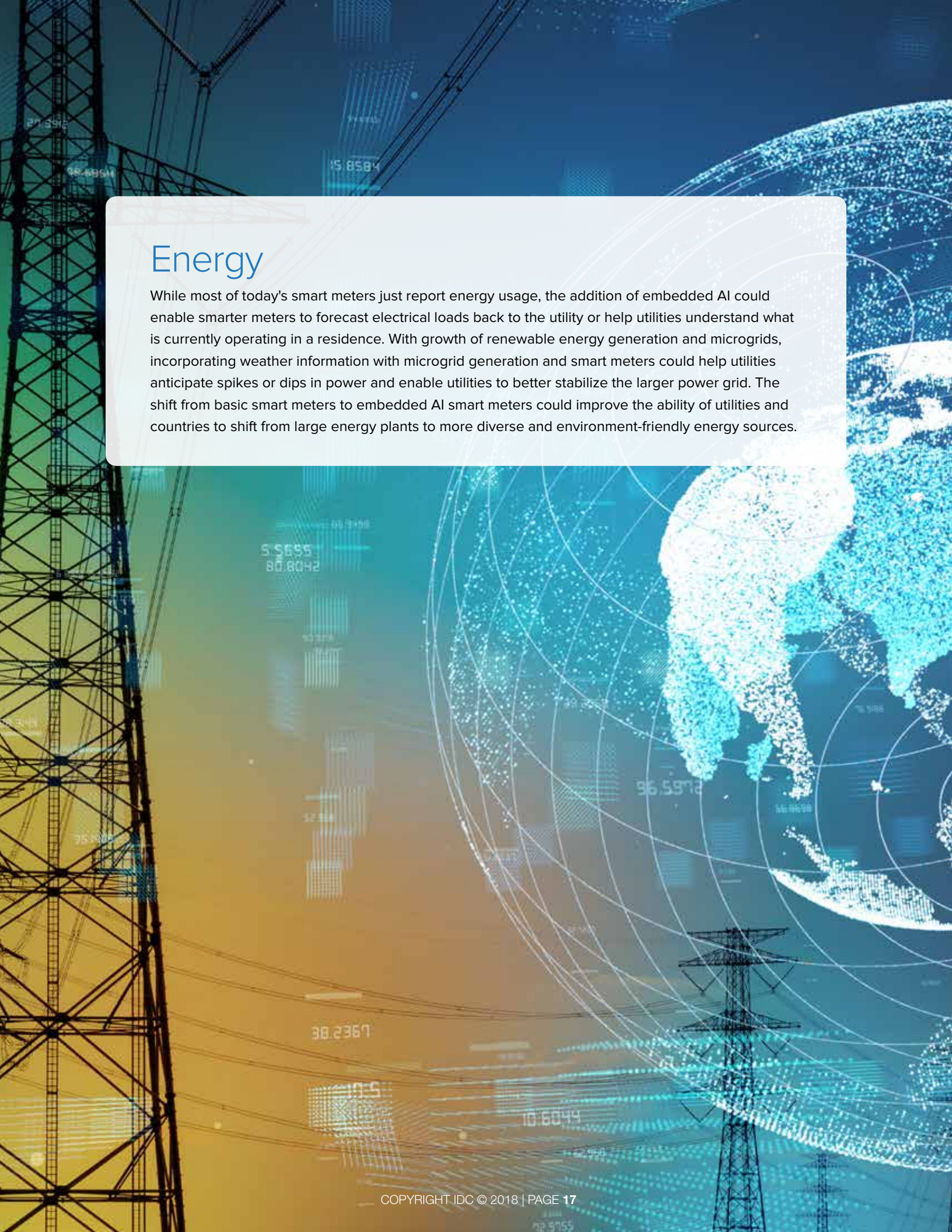
Home automation continues to also be an emerging opportunity (similar to building automation) but faster in adoption, given the appeal and convenience it offers to consumers.

The home is beginning to see an inflection point in the use of AI as consumers become more aware of the convenience and quality of services linked to smart devices. IDC expects that consumers will continue to gravitate toward experiences that are intuitive and immersive and integrate seamlessly with their personal lifestyle, which will drive more intelligence into feature-rich platforms in our homes. These smart home devices will in turn demand low-cost and power-efficient AI as part of the solution.

Healthcare

In addition to smart devices in our home, low-power wearable applications will also augment the growing demand for remote healthcare and assistant living. Wearables already monitor our heart, but they will be used to extend services and reach of devices (e.g., body-worn fabric). Low-power computing intelligence, sensors, and AI will create new markets for infusing intelligence and augmenting healthcare. Health services will also be able to provide a higher level of quality and flexibility to each patient and will allow doctors to increase their level of personal care.

The increasing number of older adults will continue to grow exponentially and drive the opportunity for healthcare services directly into our home. The United Nations and the World Health Organization estimate that, by 2030, over 12% of the world's population will be 65 years or older, which equals over 1 billion people on the planet, with the majority primed and in need of healthcare. The majority of these older adults will come from developed countries who will be educated on services and will have access to more data and AI-enabled hardware. AI will be the only way to address all these people, especially with unsustainable growth in cost. AI will help find patterns in patient data, automate record process, digitize patient records, and enable higher-resolution images for early detection and tracking of chronic states of patients. The potential for AI in this industry will come down to how quickly it can be infused at every level of the industry and how patients, doctors, and insurance companies measure results.



Energy

While most of today's smart meters just report energy usage, the addition of embedded AI could enable smarter meters to forecast electrical loads back to the utility or help utilities understand what is currently operating in a residence. With growth of renewable energy generation and microgrids, incorporating weather information with microgrid generation and smart meters could help utilities anticipate spikes or dips in power and enable utilities to better stabilize the larger power grid. The shift from basic smart meters to embedded AI smart meters could improve the ability of utilities and countries to shift from large energy plants to more diverse and environment-friendly energy sources.

MARKET EXPERIENCE AND EXPERTISE RELEVANT IN THE OT MARKET

The value chain for the OT market is broad and fragmented across a large set of vertical industries that span decades of development, investment, and commercialization.

Vendors in the OT market typically command a large percentage of a vertical segment, given the decades of investment and experience. GE, Siemens, ABB, Rockwell, Honeywell, Emerson, GM, Ford, Bosch, BMW, Hitachi, Toyota, Honda, and Denso all have a strong and established base of engagements with the public sector and B2B customers, giving them a global presence and footprint of business. Each vendor differentiates itself by leveraging deep system and market expertise established over decades, which results in an entrenched ecosystem that is very difficult for new market entrants to penetrate. The business channels established for technology suppliers that serve these vendors and markets have been established in parallel with the growing levels of expertise, support, and experience with years of service in each industry and customer. IDC expects that as OT vendors march forward in their transformation and use of AI, they will evaluate technology suppliers, especially semiconductor companies, by the following business attributes:

- Technology suppliers must be organized to sustain relationships at multiple levels with OT vendors whose markets generally dictate lower volumes in fragmented industries compared with the high-volume and concentrated IT markets.
- Partners must offer extensive development and technical support to vendors whose resources are limited in specific technologies and emerging market segments.
- Experience delivering solutions that enable customers to create unique value propositions for end users that take account of system and operating environments as well as security and industry-specific protocols will differentiate incumbent suppliers from new market entrants.
- Support long product life cycles (often at least 7, 10, or 15 years) not typically matched in the IT industry.
- **Experience and expertise in specific industry segments and applications are relevant to the needs of OT vendors,** including curating the right mix of hardware and software know-how and working in partnership and trust with the entire ecosystem of each customer.

Ultimately, more than ever before, OEMs and technology suppliers must work alongside each other to drive innovation and technology development for each market segment. AI will require that the partnership leverage their respective system expertise, hardware, tools, and software development ecosystem. The goal for systems companies will not just be to roll out the new AI and hardware technologies, but also, more importantly, the resources must meet the cost and return-on-investment goals and business outcomes of the organization.

FUTURE OUTLOOK

Embedded AI Will Drive the Next Wave of Growth for Semiconductor Suppliers in the OT Market

The quantity of data we generate overall as an industry will continue to outpace the computing power necessary to transform the data into value for any organization. Data has become the currency for doing business, and the only way to truly mine the potential is to embrace artificial intelligence and a new computing paradigm across every layer of the OT value stack.

There will be a range of hardware and software solutions to address the new computing requirements of our industry. OEMs and technology suppliers should not be confined to traditional measures and features when embracing AI into their enterprise and industry segments. Just like computing, AI will be ubiquitous.

Some key factors necessary for this new world of embedded AI to become a reality are:

- AI training will remain in the cloud and be reused across the edge and endpoint systems over the next five years.
- AI inferencing will migrate to the edge and endpoint systems as an emerging class of hardware architecture solutions come to market that bring a real balance to the requirements of the industry segment including addressing power efficiency, market flexibility, functional safety requirements, and real-time response.
- The inevitable introduction of autonomous systems in factories and infrastructure and in our homes will drive inferencing and, eventually, machine learning directly to endpoint systems; this will shift computing architectures from general-purpose microcontroller units (MCUs) and MPUs to SoCs with highly tuned and dynamically reconfigurable processing engines.
- A dynamically reconfigurable processor (DRP) is a programmable architecture that dynamically switches data paths that enable highly complex accelerators to run parallel instructions without increasing power requirements. FPGAs and DRPs use similar wiring structures and parallel techniques in the logic, but a DRP is made up of large subcomponents instead of fine-grained configurations like an FPGA. This means that it takes less time for a DRP to reconfigure than other alternative architectures including FPGAs. The configurability of DRPs allows the architecture to continue to adapt and support the changes in deep learning neural networks. A DRP can be integrated into an application-specific standard product (ASSP) and an application-specific integrated circuit (ASIC) and also as an IP block in an MPU and MCU architecture. Power efficiency and system market requirements determine the path of integration.
- The OT market will drive embedded intelligence to billions of devices demanding low-power dissipation, low latency, and real-time response, which DRPs can viably address. DRPs could enable AI inference to run in real time on many of these endpoint systems, especially in combination with other general-purpose architectures already designed in the system.
- DSPs and MCUs serve the majority of existing OT system requirements today as fixed-function computing architectures; however, as we look into the future, DRP implementations will provide a solid alternative to address embedded AI inference across a fragmented set of industries. The success of one architecture over another architecture will come from the ability to understand the market requirements and tie those needs with cost-efficient local computing to address real-time applications and workloads. As previously mentioned, brute compute will not win in this new AI-rich paradigm, especially with OT endpoint systems starting to implement inference and machine learning.

Figure 2 provides IDC's assessment of varying architectures in the market today that are vying to be the solution of choice across the OT and IT markets. IDC expects that multiple architectures will coexist over the next five years across the spectrum of computing systems in the IT and OT market segments. However, as we look at the next wave of embedded and IoT systems, the list of architectures becomes more concentrated because most architectures might address one or two technology attributes based on their strength, but not be comprehensive enough for the unique demands of the OT market. Ultimately, the OT market will center on a solution that balances the key technology attributes such as power efficiency, market flexibility, functional safety reliability, real-time operational response, market flexibility, and ease of design.

Figure 2

Technology Attributes of Architecture Solutions

	Performance efficiency	Low power	Market flexibility	Functional safety reliability	Neural network framework support	Ease of design	Real-time operation	Raw computing performance
CPU			✓	✓	✓	✓		✓
GPU					✓	✓		✓
FPGA	✓		✓	✓	✓		✓	✓
SoC					✓			✓
DRP	✓	✓	✓	✓			✓	
DSP	✓	✓	✓			✓	✓	
MCU		✓	✓	✓		✓	✓	
ASIC					✓			✓
VPU	✓	✓					✓	

VPU = visual processing unit — primarily used for imaging and computer vision. There are a few variants of these type of solutions in the market addressing markets like drones, robotics, and consumer electronics.

DRP = dynamically reconfigurable processor — a new category being used to illustrate the growing requirements for a reconfigurable solution to optimize and address the requirements of the operational technology industry segments as the market begins the adoption of AI inferencing and moves beyond using traditional MCUs, MPUs, and DSPs.

ASIC = application-specific integrated circuit — a custom semiconductor solution sold to a single OEM. Most, if not all, of the intellectual property comes from the system vendor in this design. ASIC includes Google's TPU.

SoC = system on a chip — primarily being illustrated to describe mobile baseband processors with integrated AI engines and neural network coprocessors.

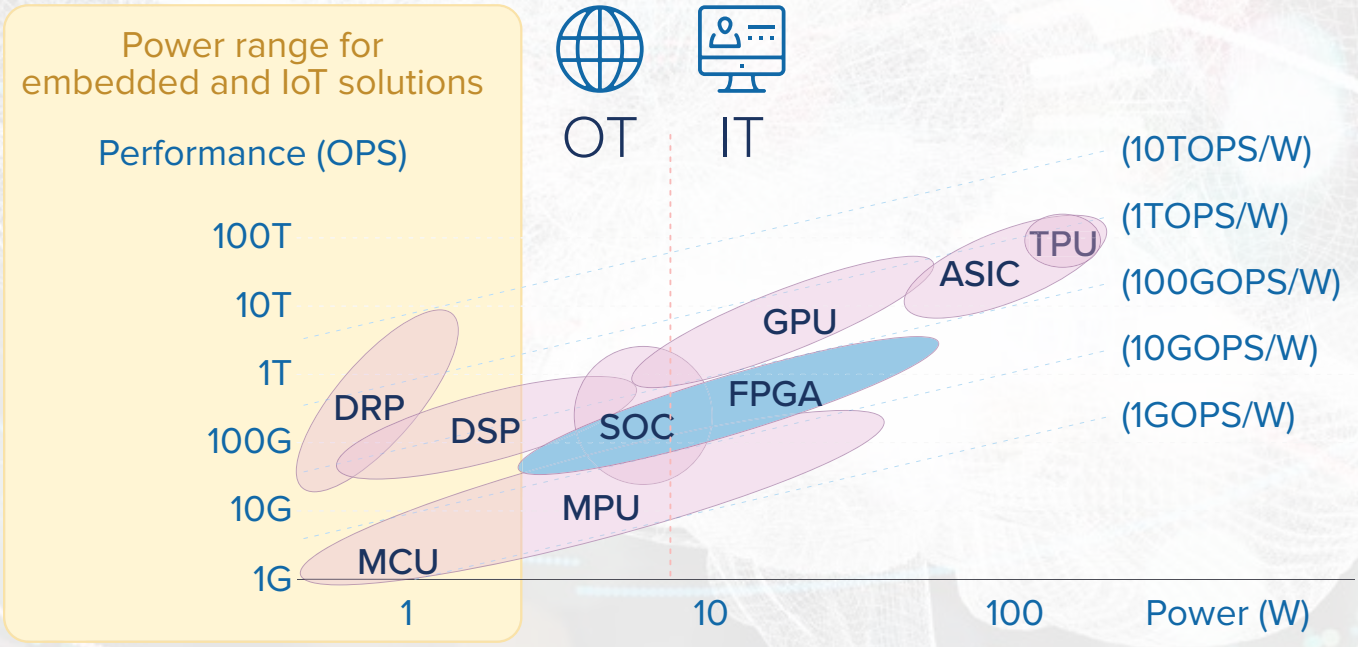
Source: IDC, 2018

As shown in Figure 3, no one type of processing option performs well against all the necessary technology attributes for AI training or inference. Established MCU, DSP, and MPU architectures serve the OT market today. As we continue to see the momentum for inference increases at endpoints, an emerging and new class of solution we are calling DRP will be able to address most of the key attributes required to run AI inference in the embedded, mobile, and industrial IoT markets over the coming years. IDC expects DRP has the opportunity to thrive as embedded and IoT systems adopt machine learning and AI.

As we look at the broad and fragmented embedded and IoT markets, performance must always find a balance with power. Today, raw computing architectures measured in operations per second (OPS) are being designed and used for training and inferencing in the cloud. In the datacenter and cloud infrastructure markets, performance remains the key focal point as illustrated by the large position of MPUs, GPUs, FPGAs, and the growing design activity of ASICs in the market by cloud service providers. However, for the OT industry, power efficiency and low power (refer back to Figure 2) are most critical to address the real-time response, actuation, and AI inference needed for the broad set of applications and workloads emerging in areas like industrial tools, building automation, and other markets previously discussed. Not only will being able to process billions of operations per second matter, but also critical operations must be done in real time and consume less than 10W of power for any key application in the OT market. These parameters prohibit cloud-centric MPUs, GPUs, and ASIC architectures from easily penetrating the OT market and open the door for alternative power-efficient solutions designed, tuned, and optimized for the key attributes of these industries.

Figure 3

Power Efficiency of Embedded and IoT Solutions



SoC = system on a chip — primarily being illustrated to describe the mobile baseband and coprocessors that integrate a neural network processor or an AI engine.

DRP = dynamically reconfigurable processor — a new category being used to illustrate the growing requirements for a reconfigurable solution to optimize and address the requirements of the operational technology industry segments as the market begins the adoption of AI inferencing and moves beyond using traditional MCUs, MPUs, and DSPs.

ASIC = application-specific integrated circuit — a custom semiconductor device sold to a single OEM. Most, if not all, of the intellectual property comes from the OEM in this design. Varieties of custom ASICs include cell-based ASICs, gate array ASICs, new structured ASICs, and FPGAs. ASIC includes Google's TPU.

VPUs = visual processing unit — falls between DSPs and MPUs and is primarily used for imaging and computer vision.

Performance = operations per second (y-axis)

Power = watts and milliwatts (x-axis)

TOPS = trillions of operations per second

GOPS = billions of operations per second

Note: The dotted red line is an estimated market threshold for power requirement for OT versus IT industry segments. Power efficiency is critical in OT systems.

Source: IDC, 2018

ESSENTIAL GUIDANCE

AI technology will continue to play a critical role in redefining how computing must be implemented in order to meet the growing diversity of devices and applications.

Technology and hardware suppliers are marching with conviction toward AI with existing architectures and early market solutions, but the journey will be met with challenges and opportunities. OT vendors are at the start of their business transformation, and what they need from their partners is no longer just products and technology. OT vendors are buying into road maps, not chips — buying a journey not a point solution. This is a fundamental change. It means technology suppliers must continue to invest not just in hardware but also in the balance of hardware and software in order to provide a solution that addresses the appetite for data and the potential of AI across the entire value stack.

Semiconductor companies will focus on consolidating workloads at the endpoint. Having experience in specific industries will be essential in enabling the transformation of large OEMs in the OT market. In addition, having hardware solutions that offer reconfigurable processing to address efficiency will be vital in order to address AI inferencing.

Figure 4 details key technology attributes for embedded AI and illustrates IDC's long-term view and essential guidance.

Over the decade, there will be key technology attributes that will be necessary in the next era of computing. Technology suppliers must be able to offer a road map for each of these attributes as the industry moves to market-ready solutions and not just point products or platforms. Speeds and feeds and brute compute are no longer the winning formula. Design cycles will give way to product life cycles and sustainable and proven technology road maps.

The competitive landscape and innovation are thriving, with no incumbent leader established in AI inferencing at the endpoint. What is clear is that over 25 billion embedded and connected computing systems will ship by 2022. Computing will be more distributed. AI frameworks will define the application and hardware as the battleground moves from cloud to endpoint systems. The competitive advantage will center on efficient performance, low power, and adaptability and flexibility of road map.

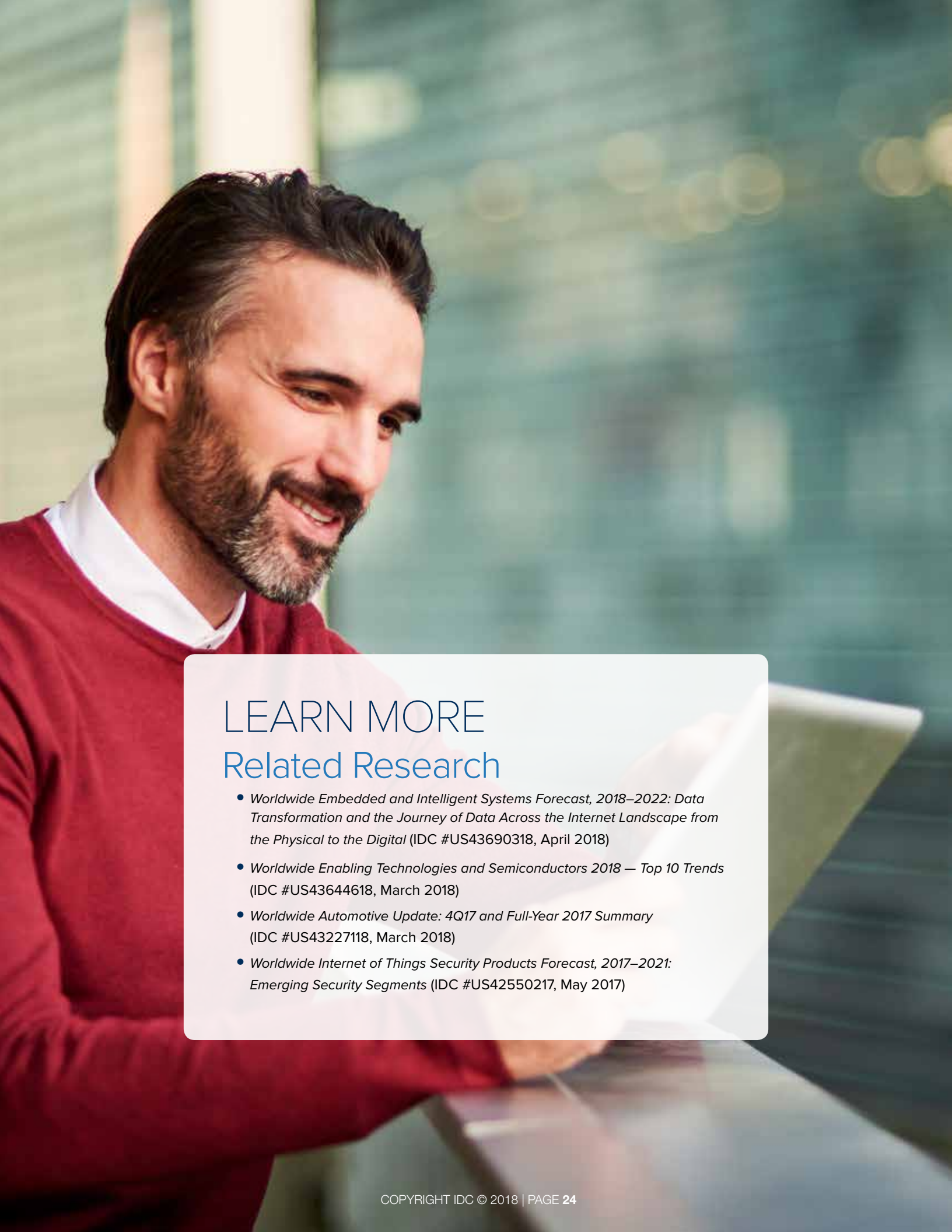
Low-power computing, sensor fusion, integration, and AI inferencing will redefine legacy markets and transform them into new opportunities for hardware suppliers and software and service companies. We are on our way; a long but inevitable journey where systems will become autonomous, requiring real-time data and decisions. These autonomous systems will be capable to create their own algorithms, self-heal, predict behavior of users, and ultimately become invisible in our everyday lives.

Are we ready for this new world? Who are the essential partners you will trust and partner with on this journey?

Figure 4

Key Technology Attributes for Embedded Artificial Intelligence — Journey into the Future

Source: IDC, 2018



LEARN MORE

Related Research

- *Worldwide Embedded and Intelligent Systems Forecast, 2018–2022: Data Transformation and the Journey of Data Across the Internet Landscape from the Physical to the Digital* (IDC #US43690318, April 2018)
- *Worldwide Enabling Technologies and Semiconductors 2018 — Top 10 Trends* (IDC #US43644618, March 2018)
- *Worldwide Automotive Update: 4Q17 and Full-Year 2017 Summary* (IDC #US43227118, March 2018)
- *Worldwide Internet of Things Security Products Forecast, 2017–2021: Emerging Security Segments* (IDC #US42550217, May 2017)

APPENDIX

Definitions

- **Artificial intelligence silicon:** It is a term that describes a new demand for raw processing capability to drive machine learning and intelligence across a multitude of different computing systems in datacenters, edge, and endpoint. The most disruptive facet of AI silicon is that when you think about how computers are programmed today or how information is processed and operating systems evolve, what's happening now is a very big shift toward doing more user training and inferencing using neural networks. Silicon is being designed with the same structure of logic as a human brain.
- **Internet of Things (IoT):** IoT is a network connecting either wired or wireless devices, or "things" that are autonomously provisioned, managed, and monitored. Embedded and intelligent systems are a subset of IoT. There is confusion in the market, however, as a large base of systems are classified by the market as IoT and intelligent systems, while a larger established set of systems are part of the deeply embedded world. The difference is in the fact that intelligent systems can process and manage data and activities, while the deeply embedded "things" tend to have a much more limited functionality, dedicated to sensors, motor controls, or actuators. Systems that are connected, without necessarily being driven by a user, are defined as part of the IoT.
- **Traditional embedded systems:** Embedded systems are computer-based products with a limited range of functions and dedicated applications, as opposed to mainstream computing segments, such as personal computers, general servers, and cellular phones. Embedded systems range from set-top boxes, digital TVs, routers, industrial automation equipment, automotive systems, and medical devices to smart cards.

Traditional Technology Solutions

- **Accelerator:** A data processing element that offloads AI-specific functions from a host processor. The accelerator may exist on the same die with the host processor or separately (discrete) from it. In detail:
  - **Integrated accelerator:** The accelerator is on the same silicon die as the host microprocessor or microcontroller.
  - **Discrete accelerator:** The AI accelerator exists separately from the host microprocessor or microcontroller and connects to the host processor through a high-speed interconnect.
- **Application-specific integrated circuit (ASIC):** An ASIC is a custom semiconductor device sold to a single OEM. Most, if not all, of the intellectual property comes from the OEM in this design. Varieties of custom ASICs include cell-based ASICs, gate array ASICs, new structured ASICs, and FPGAs.
- **Application-specific standard product (ASSP):** An ASSP is an off-the-shelf semiconductor device sold to multiple OEMs. Most, if not all, of the intellectual property comes from the semiconductor supplier in this implementation.
- **Dynamically reconfigurable processor (DRP):** This new category is being used to illustrate the growing requirements for a reconfigurable solution to optimize and address the requirements of the operational technology (OT) industry segments as the market begins the adoption of AI inferencing and moves beyond using traditional MCUs, MPUs, and DSPs.
- **Digital signal processor (DSP):** A DSP is a specialized processor that performs mathematical operations to manipulate analog information that has been converted into a digital form (e.g., in VoIP applications, DSPs are used to perform voice channel processing, echo cancellation, and compression/decompression functions). DSPs are primarily used in analog systems to process real-time data. ADCs and DACs needed to provide such functionality may sometimes be integrated on the DSP chip.
- **Discrete graphics processing unit (GPU):** Discrete GPUs are data processing semiconductors with the primary purpose of offloading processing functions from the system's main processors for accelerating certain functions, such as graphics display, artificial intelligence, and sophisticated mathematics.
- **Field-programmable gate array (FPGA):** An FPGA is a data processing semiconductor that can be programmed for specific functions after manufacturing.
- **Microprocessor unit (MPU):** A microprocessor is the central logic processing semiconductor that enables intelligence in a system.
- **Microcontroller unit (MCU):** An MCU is a semiconductor device that provides processing support for applications such as servo and motor control. The major difference between MPUs and MCUs is that it has some form of ROM, EPROM, or EEPROM, which stores programmed customer-supplied instructions. These instructions allow MCUs to carry out control functions in various applications.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street

Framingham, MA 01701

USA

508.872.8200

Twitter: @IDC

idc-community.com

www.idc.com

Copyright Notice