Opportunities for the global semiconductor market

Growing market share by embracing AI

April 3, 2019

pwc

Positive outlook

Given today's rapid technological innovation, the semiconductor industry can look forward to growth. After what is expected to be a relatively weak 2019, we anticipate that the semiconductor market will recover in 2020 and continue to prosper. Semiconductor sales totaled US$481bn in 2018. Over the next four years, through the end of 2022, we expect sales to keep growing at a slower, but still healthy, compound annual rate of 4.6%, to US$575bn.

Of the seven component types that make up the semiconductor market—memory, logic, microcomponent, analog, optoelectronic, sensor, and discrete—sales of memory products will maintain the largest share of semiconductor revenues. However, Samsung's massive capital spending on its semiconductor division in 2017-18 will drive overcapacity in the memory market, leading to a decrease in sales of memory products—and of 3D NAND flash products in particular—in 2019, but the market should start to recover in 2020.

Demand for chips related to the rapidly growing use of artificial intelligence (AI) will contribute significantly to the industry's overall growth. Much of that demand is coming from the automotive and industrial markets, the two fastest-growing areas.

The automotive market will grow fastest, with a compound annual growth rate (CAGR) of 11.9% through 2022, due to stronger penetration rates of electric and hybrid autos and the large market potential for autonomous vehicles. Meanwhile, demand for chips in traditional cars remains strong. The industrial market continues to be driven by increasing demand for AI chips and strength in both the security and healthcare sectors. The industrial market overall is expected to grow at an annual CAGR of 10.8% through that period.

The communications market will grow by a CAGR of 2.2%, due to the replacement of smartphones, the introduction of 5G technology and growth in emerging markets. Meanwhile, the consumer electronics market will generate about 50% of its revenue from the sales of TVs, video-game consoles and handhelds, and digital set-top boxes until the end of 2022. The consumer electronics market will grow by a CAGR of 6.0% through 2022. Wearables are expected to grow heavily with a CAGR of 21.0%, but the share of wearables within the communications market is only about 10%.

Growth in the data processing market, at a CAGR of 2.1%, will stem largely from sales of servers and storage devices. While a year-over-year sales decline of 2.8% is expected for 2019, we anticipate sales will pick up starting in 2020. Although we expect the market for personal computers to decline by a compound annual rate of 5.2% through 2022, that decline will be offset by growth in the Internet of Things (IoT), machine learning and other forms of AI in servers and data centers.

The healthcare opportunity

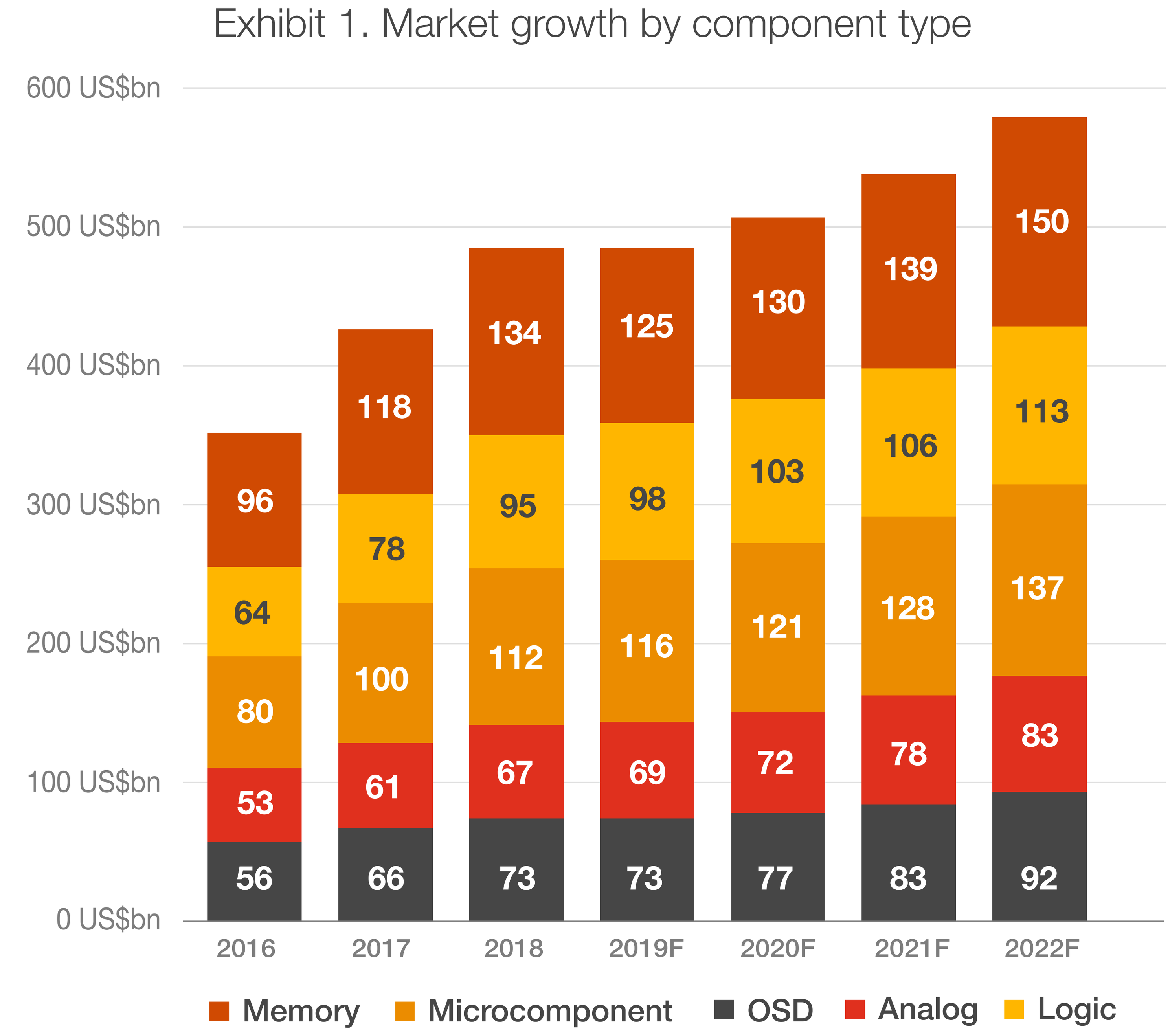
As companies in the healthcare space capture more and more of the benefits of big data, analytics and AI, these companies—including medical device makers, pharmaceutical firms and related technology companies—will have opportunities for unprecedented growth. For the US medical device market, much of that growth comes out of the Food and Drug Administration's 2016 Digital Health Innovation Plan, which encourages traditional healthcare companies to adopt new technologies and partner with digital health startups.

Tech giants such as Apple and Google have already demonstrated considerable interest in the growing healthcare market. Apple aims to consolidate a fragmented medical application market by providing all-encompassing platforms such as ResearchKit and CareKit, whereas Google is working to transform the industry through its DeepMind AI solutions. Other tech companies are also getting involved; for example, Nvidia is using GPU-powered deep learning to diagnose cancers sooner with ultrasound images.

Semiconductor industry by component type

Of the seven types of components the industry produces, the memory chip component will continue to claim the largest market share through 2022, although, as noted earlier, its growth will likely turn negative in 2019 before picking up again in 2020. Sales of logic and microcomponent chips will continue to account for the next-largest portion of overall semiconductor revenues through the forecast period (see Exhibit 1).

Exhibit 1. Market growth by component type



Source: PwC research

- **Memory.** A large portion of the growth in this segment will be driven by ongoing technological advancements such as cloud computing and virtual reality in end-devices such as smartphones. Sharply higher average selling prices (ASPs) for dynamic random access memory (DRAM) and NAND flash chips are also playing a significant role in generating revenues. In general, the expected price decreases would be offset by new capacity for flash memory and for DRAM, which would result in a better balance of supply and demand for these devices to support newer applications such as enterprise solid-state drives (SSDs), augmented and virtual reality, graphics, artificial intelligence and other complex, real-time workload functions. However, Samsung's massive capital spending on its semiconductor division in 2017-18 will create overcapacity in the memory market, the 3D NAND flash market in particular. The overcapacity will lead to excess supplies in the market, which will lower the market price for memory. As a result, revenues for these chips will drop in 2019, adversely affecting the entire semiconductor market.

- **Logic.** Demand from the communications, data processing and consumer electronics sectors will largely drive this market. Special-purpose application-specific integrated circuit (ASIC) and application-specific signal processor (ASSP) logic chips will make up the vast majority of the total market over the forecast period.

- **Microcomponent.** These chips are a crucial part of every electronic device, and the market's growth will be proportional to the number of these devices sold. Growth will stagnate in 2019 due to weak shipments of standard desktop and notebook PCs and tablet computers. Growth in microcomponents through 2022 will come from the automotive industry. Automakers are incorporating them in large numbers into intelligent cars, for powertrains and next-generation chassis and safety systems, and to process sophisticated, real-time sensor functions in safety and crash avoidance systems. The increased popularity of the IoT has also led to the need for powerful electronics, which has stoked demand for powerful processors. Local processing capability is most often provided by microcontrollers, hybrid microcontrollers/ microprocessors, and integrated microcontroller devices, which can provide real-time embedded processing, a top requirement for most IoT applications.

- **Analog.** We expect strong growth fuelled primarily by demand from the communications industry and, especially, the automotive industry. Use cases generating growth in demand include power management (to increase cellphone battery life), signal conversion (for data converters, mixed-signal devices and others) and automotive-specific analog applications (autonomous and electric vehicles and electronic systems).

- **OSD (optoelectronic, sensor and discrete components).** Under this acronym, three component types are included, which exist next to integrated circuits. Emerging technologies that are now being produced in volume will stimulate demand growth for these chips. They include solid-state lighting, machine vision, image recognition, smart-grid energy, IoT and multi-sensor 'fusion' in intelligent portable systems.

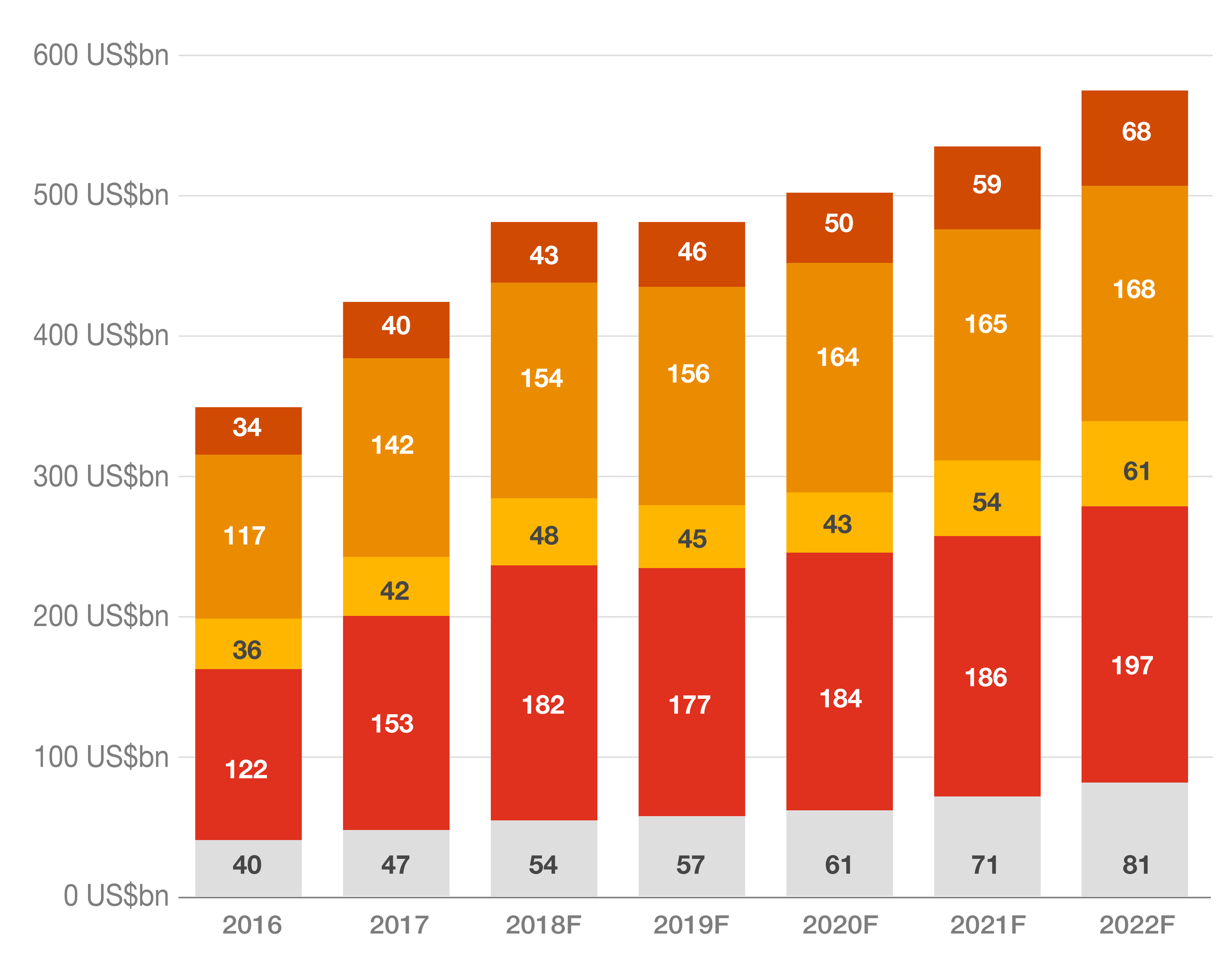
Growth in optoelectronic chips is expected to be strong, due to their increasing use in complementary metal oxide semiconductor (CMOS) image sensors for embedded cameras, automotive safety, vision-based automation and more powerful LEDs for solid-state lighting applications. In general, LED lighting solutions are rapidly transforming the market for a variety of residential, commercial and industrial applications. Among the factors propelling their growth: adoption of more energy-efficient lighting solutions, declining prices for LEDs, infrastructure modernisation and new development. LED lighting offers numerous advantages over many conventional lighting technologies in terms of energy efficiency, life span, versatility, colour quality and cost.

We also expect the sensor market to grow quickly, despite price erosion in recent years, due to an increase in unit shipments for new automated controls and IoT applications. The market for discrete components is expected to grow with stable demand increases in power transistors and other devices.

Semiconductor market growth by application type

Expectations of economic growth suggest that every application market will continue to expand, led by the automotive and data processing markets (see Exhibit 2).

Exhibit 2. Market growth by application type



Source: PwC research

• **Automotive.** We expect the automotive market to grow the fastest of all the markets, with a CAGR of 11.9%. This is due largely to strong penetration rates of electric and hybrid cars, which require about twice the semiconductor content of conventional cars, and the strong market potential for autonomous driving. Advanced driver-assistance systems (ADAS), light detection and ranging (LiDAR), infotainment, and safety and convenience functions are gaining more attention as cars become more automated and thereby require more semiconductors per vehicle. According to *IC Insights*, semiconductor content per vehicle is five times higher for full automation than for partially automated systems. Conventional cars, however, are still an important catalyst of semiconductor sales. In 2018, conventional car sales accounted for almost 95% of total revenues from the automotive market.

• **Communications.** Almost 80% of the demand for semiconductors from the communications market is driven by phones. Though the phone market is highly saturated, the introduction of 5G, the continuing high replacement rates of smartphones and the increasing demand for phones in emerging markets will maintain a CAGR of 2.2% for the market. And while demand for premium phones is expected to decline, this will be more than offset by strong growth in basic phones.

• **Consumer electronics.** Semiconductor revenue from consumer electronics applications will be generated by TV devices, driven by the increasing popularity of smart TVs, 4K ultra-HD TVs, 3D programming, video-on-demand content, a preference for large displays, and curved OLEDs. Gaming technology and set-top boxes will also be strong revenue boosters. As a result, the market will grow at a CAGR of 2.2%. Although the wearables market is still relatively small, it will grow the fastest of all the consumer electronics applications, at a CAGR of 6.0%. Revenues for chips for digital players, however, are declining, at a compound annual rate of 2.3%, as more appealing substitutes, such as Netflix and Amazon Prime, grow in popularity. The market for gaming consoles was also saturated through 2018, as consumers turned increasingly to mobile games.

• **Data processing.** Semiconductor sales in the data processing market, which includes devices such as PCs, ultra-mobiles, tablets, servers and storage devices, will grow at a moderate CAGR of 2.1% through 2022. A considerable portion of the market's growth will come from storage devices, with a CAGR of 12.3%, as smart functions in end-devices require more semiconductor content. Much of this growth will come from emerging solid-state drive technology, which overcomes the disadvantages of conventional data drives such as high turnaround time, a tendency to overheat and high power consumption. Strong sales of smartphones and other connected devices will accelerate demand for memory cards and storage devices. There is also pressure in the market to optimise server performance, which will increase the semiconductor content of each device.

• **Industrial.** After automotive, the industrial market will grow the fastest among all application types; we expect a CAGR of 10.8% through 2022. The largest share of that growth will come from demand for security, automation, solid-state lighting and transportation. We expect demand for semiconductors for security applications to grow the fastest, at a CAGR of 17.8%. This is led by the ongoing push for safer and smarter cities, especially in the Asia-Pacific region. Increasing numbers of terrorist attacks on airports and railway stations are spurring investment in advanced perimeter security and access control systems, while a growing emphasis on comfort and convenience is fuelling the popularity of fingerprint door entry systems and PIN and RFID access systems.

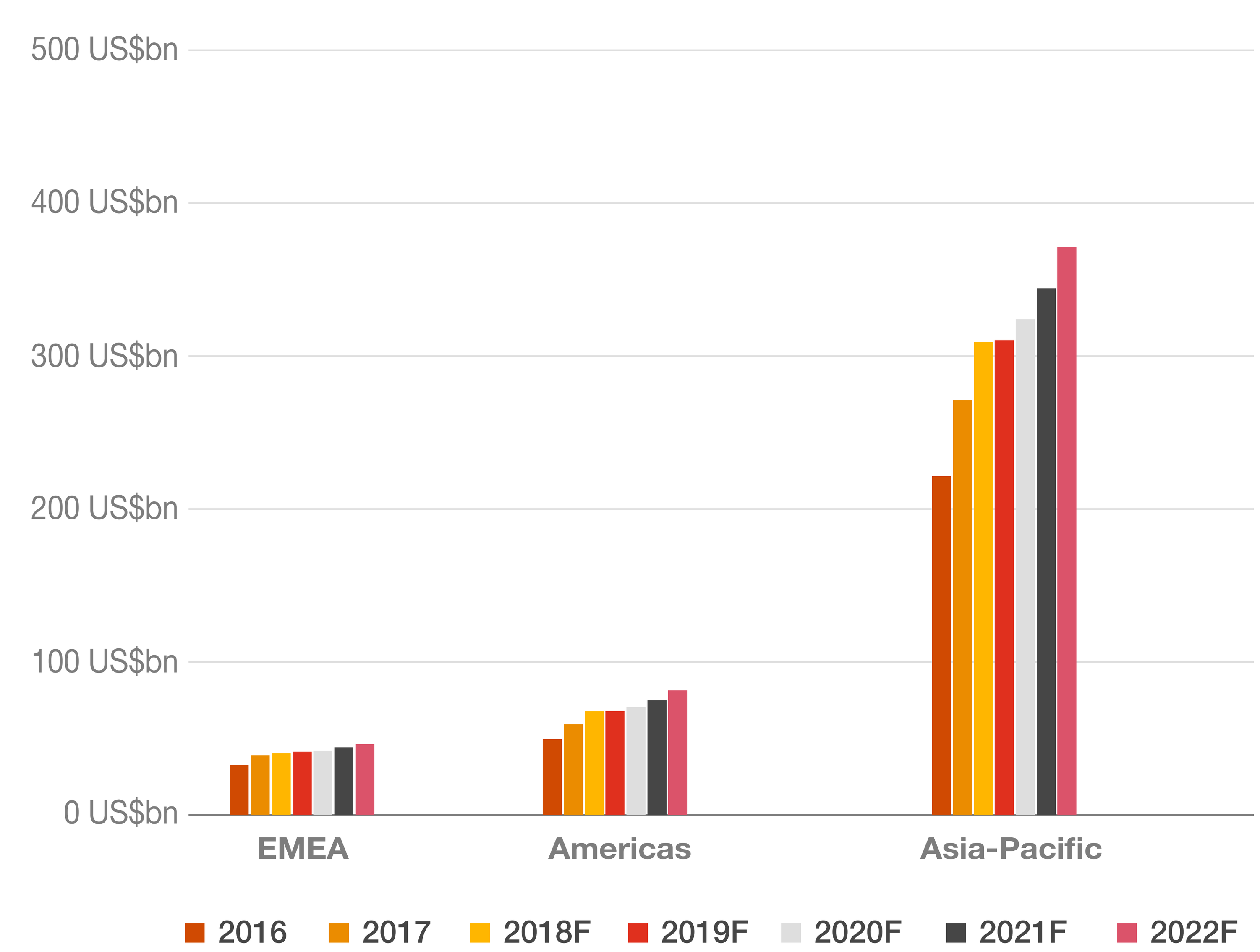
> “In the context of AI, we need to think about security, where, in general, European players are well positioned. This includes, on the one hand, secure elements and, on the other hand, dealing with environmental conditions and security layers, as in what happens around a vehicle. Both could be accompanied by a certification process, which needs to be offered at reasonable cost.”
> — Maurice Geraets, board member, NXP Netherlands

Semiconductor industry by region

Over the forecast period, we expect the semiconductor market to continue to grow quickly in every global market (see Exhibit 3).

• **Asia-Pacific.** This market will continue to be the main contributor of revenues for the semiconductor industry, with a CAGR of 4.8% through 2022. Electronic system production will continue to be centred in China. China has an ever-increasing demand for semiconductors and is the world's largest purchaser and importer of chips due to its manufacturing prowess, specifically in consumer electronics. Developing and strengthening the industry is a high priority of the Chinese government, with many budding companies in the early startup phase.

Exhibit 3. Semiconductor revenue by region



Source: PwC research

• **Europe, the Middle East and Africa.** This region will draw a CAGR over the forecast period of 3.5%. Data processing has been the largest end-use category in Europe, but we expect it to be surpassed by the automotive segment within the next two years. For many industries and applications, semiconductor content is essential. Europe has leading companies in a number of sectors like automotive, mobility (rail, air) and engineering. To secure leading positions in these industries and to facilitate new applications like AI, the EU should promote and protect its semiconductor industry. This includes R&D and manufacturing as well as the EU startup ecosystem.

• **The Americas.** This region will draw the second highest CAGR over the forecast period at 4.3%, propelled largely by projected gains in the market for NAND flash memory chips. In this region, the US is home to a number of leading semiconductor companies and has a strong startup ecosystem. Venture capital is a strong supporter of the industry. However, the US administration has recently vetoed a number of acquisition plans from non-US companies.

Rise of the learning machines

Demand in the semiconductor industry is typically fuelled by a disruptive new technology. Between 1997 and 2007, rapid increases in the popularity of PCs boosted demand for central processing units (CPUs) and memory chips, while the broad penetration of the Internet drove volume for Ethernet equipment, network processors and ASICs. The era of the smartphone began with the introduction of the iPhone in 2007, which increased demand for mobile processors, while the adoption of cloud computing has pushed growth for server CPUs and storage.

Now, artificial intelligence will likely be the catalyst for another decade-long growth cycle for the semiconductor sector. Though many of the compelling new use cases for AI will depend on algorithms implemented through software, not chips, the need for instant computing, connectivity and sensing will drive massive demand for AI-tailored semiconductors for the next decade.

> “AI accelerates learning, helps to improve customer interaction and can facilitate processes as well as procedures.”
> — Reinhard Ploss, chief executive, Infineon Technologies

AI and the semiconductor opportunity

AI is the capability of computers to simulate intelligent human behavior and make decisions or recommendations based on sophisticated analysis of data sets and predefined sets of rules. Semiconductors are instrumental to the development and acceleration of the AI opportunity and thus a key factor in boosting innovation in the field and AI’s potential for growth.

The use of AI typically depends on three kinds of algorithms:

- **Machine learning (ML):** the practice of using algorithms to parse data, learn from it and then make determinations or predictions about specific situations.
- **Deep learning (DL):** a type of ML based on analysing and learning from specific data sets, as opposed to task-specific algorithms.
- **Natural language processing (NLP):** an approach to analysing interactions between machines and humans, focussing on how to program computers to process and analyse large amounts of natural language data.

AI use cases can also be classified in terms of two primary implementation types:

- **Training systems:** These leverage massive data sets both to learn how to carry out specific activities and to continuously evolve the learning algorithm itself.
- **Inference systems:** These make real-time decisions using predefined models.

AI creates an unprecedented opportunity for semiconductor vendors due to its applicability across virtually every industry vertical, the strong forecast for the sheer number of chips needed both in the cloud and at the edge, and the growing need for specialized computing requirements to accelerate new algorithms.

> “For artificial intelligence, we need new forms of cooperation among different stages of the value chain and across industries. In a digitalised world with new ways of interaction between humans and machines emerging, both secure identities and a new ‘authority of trust’ are required. To this end, an international institution acting as an ethics commission could play an important role. Existing platforms fostering dialogue about the ethical use of AI, on the European and United Nations levels, could serve as building blocks for this institution, creating benefits from AI but also protecting the interest of the individual.”
> — Reinhard Ploss, chief executive, Infineon Technologies

AI-driven growth forecast by industry

We expect the market for AI-related semiconductors to grow from a current US$6bn in revenues to more than US$30bn by 2022, a CAGR of almost 50.0%. Although AI-driven use cases are expected to find their way across every industry segment over time, their adoption will be determined by the size of investment in the technology, the pace of its development and the speed at which its benefits are realised.

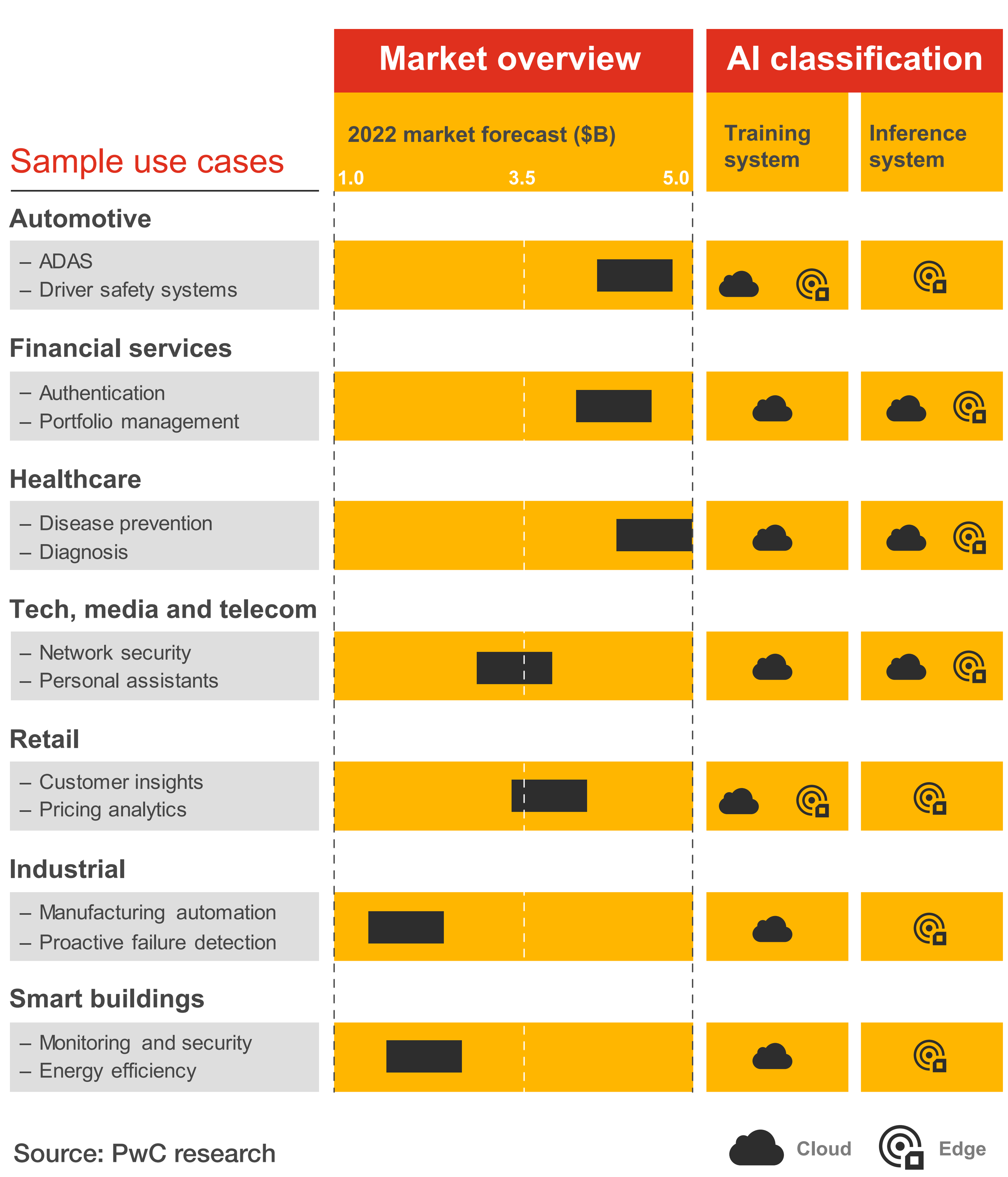
The market for semiconductors powering inference systems is likely to remain fragmented, because each of the broadly varied potential use cases — facial recognition, robotics, factory automation, autonomous driving and surveillance — will require a tailored solution. In comparison, training systems will be primarily based on traditional CPU, graphics processing unit (GPU) and field-programmable gate array (FPGA) infrastructures and ASIC.

• **Automotive.** Again, this is the segment with the largest market potential. We expect that it will bring in revenues of US$4.0bn to US$4.7bn in 2022 in ADAS and self-driving assistance use cases (see Exhibit 4). These will include both inference-based systems, for self-driving and safety assistance in the car and at the edge, and training-based systems, for traffic avoidance mapping. The relative sizes of the two will determine the types of semiconductors that will witness the most growth in demand—GPUs and ASICs for edge computing and CPUs and FPGAs for cloud computing.

• **Financial services.** We believe this segment will bring in US$4bn to $4.5bn, largely from use cases involving identity authentication for transactions and smart portfolio management. As with the automotive industry, financial services will likely implement both inference and training systems, depending on the use case. Authentication-based use cases will depend largely on inference-based AI on the edge, primarily for facial recognition on smartphones and fingerprint sensing through mobile CPU or dedicated AI semiconductors. Training-based AI will be used primarily to analyse massive data sets to recognize trends for smart investing and portfolio management; these activities will typically reside in the cloud, given their need for heavy computation based on CPU or GPU infrastructures.

• **Industrials.** Among all sectors, this one will likely present the smallest opportunity—between US$1.5bn to US$2bn—primarily from manufacturing optimisations and proactive fault detection. This is because these applications are heavily weighted toward training systems that can leverage existing infrastructure, and thus are unlikely to require best-in-class computation power and lower latency. Moreover, realizing the benefits of AI will likely take this industry longer than other sectors, due to longer deployment and refresh cycles for industrial customers.

Exhibit 4. Adoption of AI-driven use cases across industries



Source: PwC research

AI in the automotive industry

The automotive industry spends heavily on electronic components and is advancing quickly in terms of how it uses AI to further innovation. Of the US$67.9bn global automotive electronic-component market, AI's impact will be most felt in infotainment at US$8.5bn, advanced driver-assistance systems (ADAS) at US$12.9bn and safety US$5.7bn by 2022.

Infotainment systems will be used for personal assistance, navigation and entertainment. Already standing out in the market are Apple's Car-Play and Google's Android Auto platform.

ADAS and safety applications will center on driver assistance and autonomous driving, either through proprietary solutions from car manufacturers (such as GM's Cruise Automation solution) or broadly available platforms (like Intel's Mobileye and Nvidia Drive). The components for these AI applications will focus on sensing (optoelectronics and non-optical sensors), computing (ASICs, ASSPs, general-purpose logic and microcomponents), and storage (memory) with a combined available market of US$20.8bn by 2022. Analog and discrete components will be part of the overall solution but not a dominant factor for AI use.

Of these components, AI-infused logic will be designed for ASSP, ASIC and microcomponents, whereas memory, optoelectronics and non-optical sensors will function as complementary components to support the overall subsystems design. The market for AI-related chips for ADAS, safety and infotainment is forecast to reach between US$4.0bn and US$4.7bn by 2022, making up about 19.2% to 22.6% of the overall market for these application segments.

The use of AI in autonomous vehicles will depend on the extent of self-driving capabilities incorporated into the car, typically classified in five levels, depending on the amount of human intervention required for normal operation.

Level 0 involves no automation at all. At levels 1 and 2, ADAS provides automatic braking, stability control and cruise control. levels 3 involves autonomous driving under some circumstances, while at levels 4 and 5, driving is fully autonomous.

At levels 4 and 5, the autonomous driving subsystem must leverage all of its components to provide assistance in both typical and special circumstances, fully eliminating the need for a driver and even a steering wheel. Camera, radar and LiDAR sensors must detect objects to avoid. The infotainment module acts as the primary source of data delivery for navigation, sensor control and voice commands. Finally, the core autonomous platform performs the role of AI inference system for real-time computation and for making critical safety and navigation decisions.

For tasks that rely on learning systems, including real-time route navigation, personalised infotainment recommendations and digital voice assistance, on-board connectivity capabilities will route the requests to the cloud. AI algorithms for optimising these use cases will be executed through AI-tailored cloud infrastructure, typically hosted by public cloud vendors or data centres managed by the automakers or service providers.

The AI solution stack

> "The one who manages the convergence between heuristic and algorithmic programming best will benefit most from AI. This convergence can be a great opportunity for European companies, which have solid experience in dealing with heuristics as well as with complex systems."
> — Reinhard Ploss, chief executive, Infineon Technologies

Hardly a day goes by without mention of a new and innovative AI-enabled product or service that will change the world as we know it. We are already captivated by how virtual assistants can initiate a human-like phone call and reserve a table at a restaurant. But to understand better where the AI opportunities lie, it is necessary to take a deeper look into the underlying components of the AI technology stack, the scaffolding on which the applications are built.

In our view, the AI technology stack consists of five elements or layers: hardware, libraries, frameworks and tools, the platform, and applications and services (see Exhibit 5). Since most of the attention devoted to AI has focussed on AI-enabled customer experiences, it is logical to begin with applications and services. This is the topmost layer of the solution stack, where the most tangible AI capabilities — such as Amazon's Alexa virtual assistant and Apple's Face ID — come together at an application level. Some of these capabilities are also being offered as services, such as recommendation engines embedded into software.

Exhibit 5. Elements of the AI technology stack

Stack elements	Description
Applications and services	Software applications leveraging AI for 'intelligence,' including vision processing, customer-support chat bots, smart assistants, algorithmic trading.
AI platforms	Ready-to-use building blocks and services that provide capabilities including machine learning, data analytics, NLP, agents, data solutions that can be leveraged to build AI applications.
AI frameworks, tools and interfaces	Technology that leverages underlying ML algorithms to design, build and train deep-learning models for specific applications. Many are open-source and broadly supported.
AI libraries	Low-level software functions that help optimise the deployment of an AI framework on a specific target hardware.
AI hardware	Processor units and semiconductor logic circuits designed and optimised to accelerate execution of AI workloads and computations.

Source: PwC research

But what are applications without the underlying reusable components that provide their core functionality? This is the task of the platform layer. Several companies are producing AI platforms, promising the ability to build AI-enabled apps without having to deal with the complexity of the underlying dynamic neural network (DNN) and algorithms. These companies have built platforms that aim to provide 'ready-to-use' building blocks and software services — foundational AI capabilities such as NLP, agents and decision engines, which help speed up development of AI applications and services. Examples include:

- **Rainbird Technologies.** A software-as-a-service-based AI platform from Rainbird Technologies is aimed at making business operations smarter. It provides a rules-based automated decision engine that can enable complex tasks such as making predictions, recommendations and business decisions. It also captures the rationale for why certain decisions were made by the platform, which is valuable for audits, especially in regulated industries.
- **Semantic Machines.** This startup, based in Berkeley, Calif., and recently acquired by Microsoft, has developed a fundamental machine learning–based technology platform that enables users to interact effortlessly with information systems. This approach, dubbed 'conversational AI,' is expected to have far-reaching implications for how we transact business on e-commerce sites, interact with social media and even use productivity software and our devices on a day-to-day basis.

The middle of the stack — AI frameworks, tools and interfaces — enables developers to design, build and deploy the actual models and algorithms. Independent software vendors (ISVs) are offering AI frameworks, tools and interfaces for developers to use the underlying AI algorithms to build deep-learning models for specific use cases. Several of these frameworks are also open source, which will allow for their broad adoption and considerable support from most players in the AI ecosystem.

The bottom two layers of the stack consist of hardware — the processors, logic circuits and other components that run AI software — and libraries, which are essentially low-level software functions that help optimise the AI models and algorithms for the underlying silicon chipset. We expect the traditional semiconductor vendors such as Intel, Nvidia, Qualcomm and Xilinx to supply the silicon chips optimised to accelerate AI use cases for this portion of the stack. These companies will also likely provide the libraries needed to facilitate the development and scaling adoption of their proprietary architectures and in turn help deploy AI frameworks on their silicon products. Examples of AI libraries include Intel DL SDK/Vision SDK, Nvidia cuDNN TensorRT and ARM NN.

It is becoming increasingly evident that the hardware layer is arguably the most interesting part of this AI solution stack, for two key reasons: First, there is a growing recognition that AI demands unique processing requirements for its underlying hardware. This has led to a renewed race to choose the optimal processing architecture. Which architecture will win — GPU, digital signal processor (DSP), FPGA or custom ASIC — remains to be seen. Second, the number of players developing AI hardware has increased beyond the usual list of the traditional chipmakers, which could threaten entrenched vendors and redraw their market positions dramatically.

Exhibit 6. High expectations of AI opportunity reflected in array of products

		Semiconductor companies						IP licensors	
		Intel	**Nvidia**	**Qualcomm**	**Xilinx**	**NXP**	**ST**	**ARM**	**Cadence**
Target applications		• *Data centre* • *Self-driving cars* • *AR/VR* • *Drones* • *Surveillance*	• *Data centre* • *Self-driving cars* • *Retail analytics* • *Smart cities* • *Surveillance*	• *Computer vision*	• *ADAS* • *Drones* • *Data centre* • *Medical*	• *Vision processing for ADAS*	• *Computer vision* • *Smart driving*	• *Voice assistants* • *Consumer robots* • *Smartphones*	• *Automotive* • *Surveillance* • *Drone* • *Mobile/wearable*
AI libraries		**MKL DL SDK Vision SDK Movidius Myriad Dev Kit**	**CUDA DNN TensorRT**	**Snapdragon NPE SDK**	**reVISION SDAccel Toolkit**	**S32 Design Studio IDE Vision SDK**	**STM32 STM32 Cube**	**ARM NN**	**Tensilica NN mapper toolkit DSP SDK**
AI processing, hardware/silicon	**CPU**	Xeon PHI						Cortex-A75 Cortex-A55	
	DSP			Hexagon DSP (Snapdragon NPE)					Tensilica Vision C5 DSP
	GPU		Pascal (Volta) Maxwell Tesla						
	FPGA	Arria 10			Zynq MPSoC				
	Custom	Nervana NNP Myriad X Loihi NMP	ASIC - TPU			S32V Vision Processor	AL SoC for DCNN	ARM ML	

Source: PwC research

■ Training ■ Inference ■ Both

The emerging battleground for AI silicon

In a clear sign of the high expectations surrounding the market opportunity for semiconductors in AI, every major vendor is offering AI silicon. The most commonly targeted applications are ADAS, drones, surveillance and computer vision.

There is a wide variation in their choice of architectures, ranging from general-purpose CPU, DSP, GPU and FPGA to custom ASICs (see Exhibit 6). As expected, most vendors' choice of architectures for AI silicon is closely aligned with their core capability or area of strength. Xilinx's Zynq MPSoC is a customisable variant of its FPGA product, for example, and most of Nvidia's products are based on its core GP-GPU architecture.

On the other hand, vendors such as NXP and STMicroelectronics mostly provide AI-specific accelerations and extensions that augment the capabilities of their existing product portfolios, rather than AI-specific chips.

In yet another variation, IP licensing vendors ARM and Cadence offer soft CPU and DSP IP cores, on the assumption that in the future, AI processing will be embedded in ASICs, rather than handled by stand-alone chips specialised for AI workloads. Their model allows silicon vendors to license the AI soft cores to develop their own chips targeting AI applications. Meanwhile, vendors with deep pockets, such as Intel, are making broad bets on a variety of different architectures — CPU, FPGA and custom ASICs — designed to address different processing requirements.

Another distinction we've observed is in whether companies are producing chips designed specifically for either training or inference systems. Intel and Nvidia offer the most diverse set of chips, intended for either the training or inference markets. Intel's Arria 10 FPGA and Myriad X ASIC are designed for inference workloads, while its Nervana NNP is positioned as best for training. Similarly, Nvidia makes Pascal and Volta chips for training workloads and Maxwell for inference. Both companies also make chips — Intel's Loihi NMP and Nvidia's Tesla — designed to work well in either application. No single approach is best for every scenario, we believe; the best approach will likely be different for each use case, depending on the types of data sources to be analysed, data gravity considerations and requirements for real-time processing.

Custom solutions

Amid this flurry of innovation, some companies are likely to try to develop custom chips that promise to deliver the AI 'holy grail' — superior performance at lower power and cost than any of the standard architectures introduced in the first generation. This battle of architectures is likely to continue into the foreseeable future, and we believe that it is too early to call winners. Custom chips will likely perform best overall, but their economics might not work out if they address only a very narrow set of applications and use cases, as the resulting lower volumes may not justify their upfront development costs.

Several nontraditional chipmakers have joined the arms race for AI semiconductor superiority and are testing the waters with custom chips designed for their specific AI needs (see Exhibit 7). We're witnessing an unmistakable trend among the top public cloud vendors — notably Amazon, Google and Microsoft — all of which are exploring custom AI chips as an alternative to GPUs and FPGAs in order to gain a competitive advantage in the performance and cost of their cloud offerings. Amazon recently announced an AI chip for Alexa, its edge computing home device; Microsoft is developing an AI chip for its HoloLens smart eyeglasses; and in 2017, Google launched its tensor processing units (TPUs) for neural networks, claiming that they will deliver 15 to 30 times higher performance and 30 to 80 times higher performance per watt than CPU/GPU chips for similar workloads.

Several companies, including Apple, Samsung and Tesla, are developing their own AI silicon, customised to the requirements of their products. Apple introduced its A12 Bionic for iPhone XR and iPhone XS smartphones; the product includes a neural engine for Face ID and Animoji (animated emoji) applications and contains an image processor for computational photography and pixel processing functions.

These major shifts in the race for AI dominance, with public cloud providers and product firms developing their own custom silicon to optimise their applications and use cases, are bound to threaten the market position of the traditional chipmakers such as Intel, Nvidia and Xilinx, and potentially upend their traditional business models. Despite all the custom activity, we expect that GPUs and FPGAs will continue to coexist in the cloud to accelerate AI workloads. In the near term, most applications will be served through commercially available silicon products, given the substantial investments and resources required to develop a new silicon design and the high sales volumes needed to do so profitably.

The AI startup landscape

Yet another trend with which semiconductor makers must contend is the huge contingent of startups, which are developing and commercialising revolutionary new chip architectures optimised for AI. The key question: Will any of these young companies pose a threat to the incumbents, or even perhaps find an opportunity to leapfrog the competition and win in AI?

The rush to capitalise on the promise of AI is triggering plenty of innovation across the solution stack. Venture capital funding for AI startups has been increasing dramatically in recent years, culminating in a record US$11bn invested in AI and machine learning companies in 2017. Unsurprisingly, much of the action is taking place in the software and algorithm space at the top of the stack, where startups are building scalable platforms focused on specific AI use cases, and are looking to develop AI software that can be integrated into existing applications to make them more intelligent.

Still, there is also considerable interest and excitement at the bottom of the stack, where a rising number of startups are targeting new silicon architectures that are optimised to meet the unique processing requirements posed by AI workloads. Our analysis of VC funding shows a resurgence of interest in semiconductor startups, which captured close to US$750m in VC funding in 2017, three times more than the funding of the prior two years combined, and 12 times more than all the funding directed to AI chip startups in all the years prior to 2015.

As Exhibit 8 shows, 11 of the top 19 AI semiconductor startups are based in the US, and most are exploring specialised processor architectures tailor-made for various AI and deep-learning workloads. Nine of them are building deep-learning processors, and three are developing so-called neuromorphic processors, which are based on radical new architectures that attempt to mimic the way the human brain functions.

Exhibit 7. Other players join the race for AI

		Cloud providers				Other product vendors		
		Amazon Web Services	Microsoft	Google Cloud Platform	Baidu	Samsung	Tesla	Apple
Target applications		• *Facial recognition* • *Text-to-speech* • *Smart assistant*	• *Digital transformation* • *Intelligent assistant*	• *Image search* • *Voice search* • *Translate* • *Smart reply*	• *Search* • *Voice assistant* • *Computer vision*	• *Facial recognition* • *Animated emoji*	• *Self-driving cars*	• *Face ID* • *Animoji*
AI libraries		AWS DL AMI Xilinx SDAccel	Bing API Face API Analytics API	Video API Vision API Speech API NL API				Core ML
AI processing, hardware/silicon	GPU	Nvidia GPUs	Nvidia GPUs				Nvidia GPUs	
	FPGA	Xilinx AWS EC2 F1	Intel Project Brainwave		Xilinx Cloud Server			
	Custom	AI chip for Edge (Alexa)	AI chip for Hololens	Cloud TPU TPU		Natural Engine (Exynos 9)	AMD In-car chip	Neural Engine (A12 Bionic)

■ Training ■ Inference ■ Both

Source: PwC research

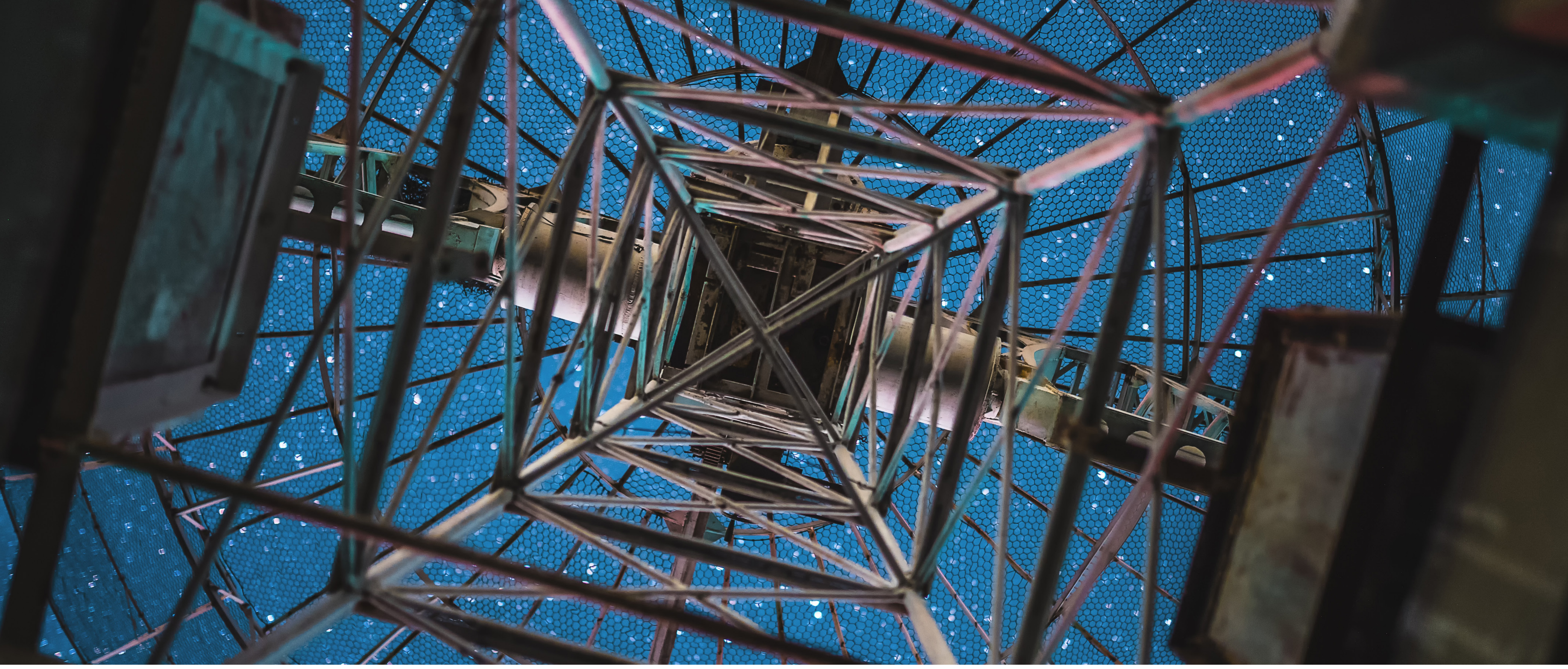


Exhibit 8. Venture capital funding for AI startups is increasing

Startup	Founded	HQ (GEO)	Stage	Funding to date (US$m)	Strategic investors	Technology
Horizon Robotics	2015	Beijing, China	Series B	700	SK Hynix, SK China	Vision DSP
Graphcore	2016	Bristol, UK	Series D	310	Microsoft, BMW	Deep-learning processor
Cambricon Technologies	2016	Beijing, China	Series B	210	SDIC	Deep-learning processor
Wave Computing	2010	Campbell, CA	Series E	203.3	Samsung	Deep-learning processor
Vicarious	2010	San Francisco, CA	Series C	120	Samsung	Neuromorphic processor
Rigetti Computing	2013	Berkeley, CA	Series B	119		Optical/quantum AI computing
Cerebras	2016	Los Altos, CA	Series C	112		Deep-learning processor
Vayyar	2011	Yehud, Israel	Series C	80		Vision DSP
ThinkForce	2017	Shanghai, China	Series A	68		AI acceleration engine
Movidius	2006	San Mateo, CA	Series E	56	Intel (Acquired)	Neural compute engine accelerator (Appl: Vision DSP)
Mythic	2012	Redwood City, CA Austin, TX	Series B	55.2		Neuromorphic processor
KnuEdge	2005	San Diego, CA	Seed	47		Neuromorphic processor
Nervana	2014	San Diego, CA	Series A	40	Intel (Acquired)	Deep-learning processor
Xanadu	2016	Toronto, Canada	Seed	25		Optical/quantum AI computing
Reduced Energy Microsystems	2014	San Francisco, CA	Seed	6.9		Deep-learning processor
LightOn	2016	Paris, France	Seed	3.3		Optical/quantum AI computing
CyberSwarm	2017	San Mateo, CA	Seed	2		AI-assisted cybersecurity CPU
Tenstorrent	2016	Toronto, Canada	Seed	1		Deep-learning processor
Vathys	2015	Portland, OR	Seed			Deep-learning processor

Series A Series B Series C Series D Series E Seed

Source: PwC research

It is worth noting that only a minority of these companies have strategic investors, the most active of which are Intel and Samsung. In fact, Intel has already acquired both Movidius and Nervana and has begun to integrate their offerings into its AI road map. One notable absentee from the strategic investor list is Nvidia, which has instead used its venture arm to invest further up the stack, in companies that are building platforms and applications.

Total VC investment in AI startups to date is evenly split between early-stage Series A and B investments and late-stage Series C and D deals. Although most of the late-stage startups are based in the US and Europe, the bulk of the early-stage outfits are in the Asia-Pacific region. Among the biggest Chinese startups are Cambricon Technologies, Horizon Robotics, ThinkForce and DeePhi Technologies, which have raised a combined US$300m in VC funding to date, with the first two accounting for roughly two-thirds of that amount.

Based on our analysis thus far, we believe that the race to build the best silicon for AI is just getting started. The competition is expected to be intense, and threats to incumbents are likely to emerge from as-yet-unforeseen directions. Though only time will tell who will be the winners and losers, every semiconductor company with hopes of playing on this highly competitive field must begin to prepare now.

Capturing the AI opportunity

History has shown that although semiconductor companies have profited handsomely from disruptive growth cycles, they have repeatedly failed to capture their fair share of each cycle's full value, either through monetisation of new technologies beyond the actual chips themselves or through expansion into new business models enabled by those technologies.

The rise of AI will almost certainly be the most powerful driving force in the semiconductor industry for the coming decade. As our analysis shows, incumbents and startups alike are working hard to develop the hardware that will power AI. But can they also capture the full value of this opportunity? Can they go beyond the profits to be made from the development and sale of AI chips to participate in the AI revolution as a whole?

We believe they can, but to do so they must take a deliberate approach to reevaluating their AI strategy and business models, crafting their technology and product strategies, and figuring out how to leverage their role in the entire AI ecosystem. Following are a few key considerations and suggestions for companies as they prepare to navigate the changes wrought by AI and make the most of this opportunity.

Strategy and business models. Develop the company's vision for AI, and then use that vision to better understand the core areas on which to focus, at least in the early stages of the game. Part of this exercise should involve exploring new ways to monetise company assets and areas of expertise. Can data be leveraged? Are there relevant services to offer? This approach will also inform how best to evolve current portfolios to become more AI relevant and AI friendly, while staying closely aligned with overall strategy. Here are three focus areas on which companies may choose to concentrate:

- *Growth segment.* Identify and target new AI-enabled use cases in market segments that take advantage of a company's current footprint and offer significant growth potential, such as ADAS and IoT.
- *Monetising beyond silicon.* Explore opportunities to either license AI-specific intellectual property (IP) or offer AI-related services that could be monetised. These could include hosted AI-as-a-service and anonymised AI use case data to improve training systems and algorithms.
- *Product portfolio.* Evaluate carefully where to make big bets to build new AI capabilities as opposed to making incremental investments to increase your current portfolio's AI friendliness.

Technology and product offerings. Choosing the right technology and architecture for chips, given their specific use cases, is crucial, but companies must also ensure that the chosen technology path is supported elsewhere in the AI stack. One proven approach is to define product offerings to broadly encompass the libraries, tool kits and other software elements within the stack, but a firm must determine whether to build all the elements itself or to integrate its offerings with elements from technology partners. Here are some ways companies may focus their offerings:

- *Evolving hardware architectures.* Explore designs with custom architectures, such as neuromorphic processing, and subsystem designs that are tailored to accelerate deep-learning algorithms through their unique self-learning capabilities — by integrating logic and memory functions, for example.
- *AI libraries and tool kits.* Develop software development kits (SDKs) and compliers to optimise and accelerate AI algorithms for existing products within the portfolio, wherever possible.
- *Full-stack offerings.* Work with partners to offer a full-stack solution across silicon, platforms, tools and libraries to enable ease of application development and differentiation.

Partnerships and the AI ecosystem. The deep dive into the AI solution stack above clearly demonstrates that success in AI is heavily dependent on building a complete ecosystem of partnerships, both through technology alliances and as companies take their offerings to market. Smart players will identify and effectively leverage the ecosystem to reduce time-to-market, build strong sales efforts to make inroads into specific industry verticals or applications, and evaluate strategic partnerships and investments, such as joint ventures and acquisitions, that can help drive long-term growth. Here are some options for companies to explore:

- *Licensing.* Explore options to both reduce risk and speed up AI-specific product development by licensing IP cores from third-party vendors such as ARM, CEVA and Cadence.
- *Partnerships.* Formulate new strategic partnerships with players across the stack to drive adoption of your AI chips and solutions.
- *Strategic investments.* Consider investing in AI startups to enhance capabilities for use cases aligned to your core business.

Digitisation in the industry

The semiconductor industry has been a pioneer in digitisation since its inception, offering digital services and pursuing new digital business models. There were times during the 1970s, for example, when Intel made more money from selling testing equipment for its chips than from selling the chips themselves. In the 1980s, with the proliferation of digital design and simulation tools and communications technologies, the fabless and foundry models emerged and disrupted the then-prevalent IDM model. Later, pure IP plays emerged, pursued by Qualcomm, ARM and others, further eroding the semiconductor value chain.

Today, other industries, notably automotive, have clearly outpaced the semiconductor industry in terms of digitisation. That's ironic, given that automakers' own success in digitisation has been largely supported by the products made by the semiconductor industry. It is now more critical than ever for semiconductor companies to consider how best to leverage digitisation, and which of the possibilities make the most sense for their organisations.

Chipmakers can weigh three general strategies — digitised product and service offerings, digital business models and the digital semiconductor value chain — as they consider how best to pursue digitisation. Note that all of these strategies must be supported by robust data and analytics.

Digitised product and service offerings. We see three options for semiconductor companies to digitise their product and service offerings: data monetisation, augmentation and customisation (see Exhibit 9). Data monetisation allows semiconductor firms to capitalise on the massive amounts of data being produced by their own or others' devices. Augmented products and services enrich a company's already existing technology with AI or by integration with other products or services. Customisation brings additional value to the customer through a higher level of precision and efficiency.

Exhibit 9. Digital strategies for semiconductor companies

	Data monetisation		Augmentation			Customisation
Play	• Provide raw data • 'Cleanse' and organise data • Generate insights from data	• AI-assisted integrated circuit (IC) design service	• AI-assisted IC integration service	• IoT platform integration	• Complete devices	• Chip on demand
Description	• IC design data provided for customer • Algorithm provided to customer for data cleansing, insight generation	• AI-based design and bug-fixing services provided to customer	• AI-based circuit design services working primarily around the IC product base of a vendor	• Chip vendors work with hyperscale cloud platform providers to ensure tools to program IC products are available in the cloud platform's integrated development environments and are enabled via application programming interfaces	• Better enables capturing of IoT opportunities by owning the data • Moving closer to the customer enables more successful application-specific designs	• Looking ahead, proliferation of AI applications for chip design will allow the customer to set performance and functionality parameters for IC, with design options created by an AI engine
Example/ potential use case	• Samsung's Artik Cloud IoT Platform	• Intel's EDA Design Flow (chip)*	• Intel's EDA Design Flow (board-level design)*	• Qualcomm/Linaro* • Cloud service	• Intel's Mobileye	• Intel (experimental phases)

* Potential use cases
Source: PwC research

Digital business models. We see five types of innovative digital business models for semiconductor firms to choose from (see Exhibit 10):

1. ***Razor and blades:*** Firms can capitalise on this model by offering a core product — such as a computing architecture — for a lower profit margin and then developing an additional product, such as cloud platform services, that depends on this architecture but carries a higher margin. In the future, semiconductor vendors will be able to provide devices for rent to cloud service providers, which will monetise hardware utilisation.
2. ***Platform:*** Companies can create value from this opportunity by facilitating data- or hardware-based exchanges between chipmakers and customers. This benefits the platform participants and allows the hosts to influence standards and increase switching costs to competitors.
3. ***Open source:*** With this opportunity, semiconductor firms create a platform that allows customers to build customised, open source–enabled chips. This facilitates the public sharing of software source code and integrated circuit design with third parties, thus dividing R&D costs and reducing time-to-market.
4. ***XaaS (everything as a service):*** Semiconductor makers can use this approach to develop innovative services around infrastructure, hardware and software, such as computing as a service. Then hardware and functionality updates can be billed as a service.
5. ***Marketplace:*** Develop a two-sided marketplace that increases in value through the positive network effects created as more and more people use it. The marketplace can include cloud-based algorithms-as-a-service, for example, for shared AI training data. Customers can access the algorithms with a general interface and developers can upload new algorithms and models to enrich the marketplace.

Exhibit 10. Innovative business models for semiconductor companies

Type	Play	Description	Example
Razor and blades	• Cloud platform and novel computing	• Provide cloud platform services supported by leading computing architectures • Support application-specific computing architectures (e.g. deep learning) • In the future, chip vendors will be able to rent devices to cloud service providers, monetising hardware use	Google's cloud platform Nvdia's GPU cloud
Platform	• Data platform	• Create data interchange and monetisation platforms for collaboration partners, which benefit the platform participants and allow hosts to influence standards and increase switching costs to competitors	TSMC's open innovation platform
	• Platform SW infrastructure	• Develop standard APIs as part of the AI silicon IP that allow for third-party developers to create and deploy AI models on the chip • Create standard tool kits, compilers and/or libraries that build 'developer stickiness' for chips within specific applications	CEVA and Nvidia's deep learning SDK ARM's NN SDK TensorFlow library
	• IoT platform	• Enable the interoperability of IC products with the value-added services of external vendors, making possible business models such as hardware-as-a-service	ARM and Linaro's open source ARM platform Intel's IoT platforms
Open source	• Open source IC designs	• Create a platform that allows customers to build customised, open source–enabled silicon chips • Customisation can be quickly incorporated and delivered to the customer faster and at lower cost than traditional designs	SiFive's open source designs
XaaS	• Computing-as-a-service	• Hardware and functionality updates billed as a service • Service dynamically reprograms field-programmable gate arrays as new and improved inference models are downloaded onto chips • Chip-features-as-a-service (unlocked for a fee)	IBM's Q (quantum computing)
Marketplace	• AI algorithms/models	• Create marketplace for cloud-based algorithms-as-a-service • Customers can access algorithms via a general interface • Developers can upload new algorithms/models to enrich the marketplace	Algorithmia's AI algorithm marketplace
	• Data for end-user applications	• Provide services that help consolidate AI-related data across customers and use cases, anonymise it and make it available for sale in an AI data marketplace • Potential buyers of this data can obtain feedback to improve the training/algorithms and learn how end-customers are using AI	Bottos' distributed AI data marketplace OpenMind and Ocean's AI data sharing marketplace

Source: PwC research

Digital semiconductor value chain. Semiconductor players could profit significantly by digitising both their vertical and horizontal value chains, from end-to-end, not only to leverage new AI-driven capabilities, but to fully exploit the other digital opportunities they offer. Exhibit 11 shows a number of initiatives companies should consider pursuing.

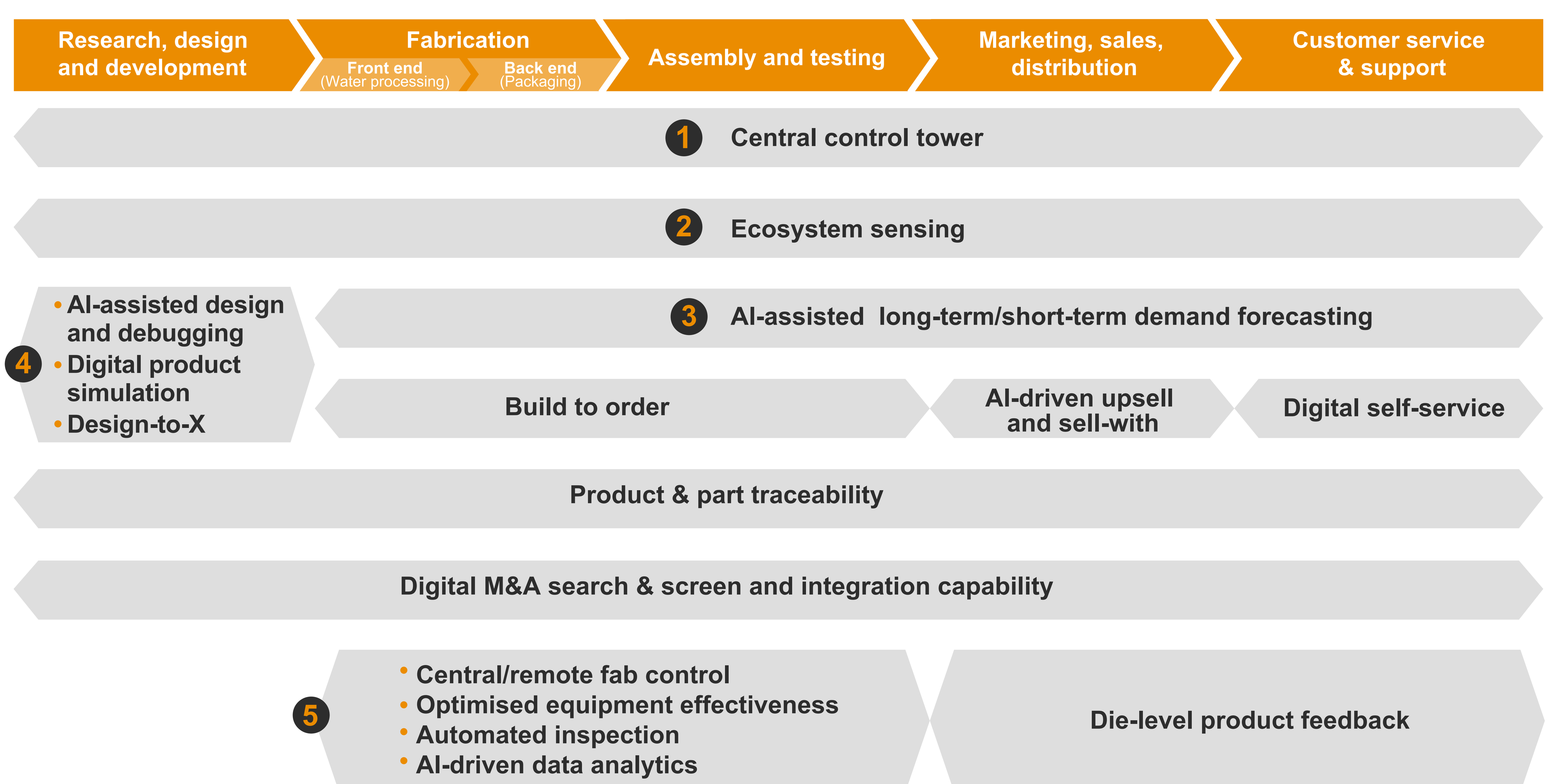
Taken together, semiconductor companies should consider the full range of opportunities available to them through all three digital strategies (see Exhibit 12).

Operations. In addition to the various digital strategies chipmakers can employ to boost the top line, they should also consider the opportunities to improve earnings by applying AI and machine learning to improve their operations. Among the options:

- *Central control tower:* This allows companies to maintain a detailed real-time view into all supply chain segments and operations, including the operations of suppliers and customers. Dashboards can be designed for virtual control room operation and to provide visibility for key stakeholders.
- *Ecosystem sensing:* Use AI to help gather data and insights from your supplier and customer ecosystem, to identify the most pertinent signals indicating further opportunities and to make suggestions for taking action in response.
- *AI-assisted long-term/short-term demand forecasting:* Gather demand signals from the ecosystem such as moves by large players, signals from the supply chain, relevant news and the like, and analyse their implications to improve real-time production mix and supply scheduling.
- *AI-assisted design and debugging:* Use machine-learning systems to suggest IC design solutions, identify potentially buggy design elements and implement more efficient product design branching.
- *Fabrication process optimisation:* Locate fab control centres in remote facilities or alongside office space to improve learning and responsiveness, thus improving equipment efficiency via analysis of sensor logs of production equipment and related events. Employ computer vision tools to identify faulty assemblies. AI will support the release of wafer lots in the front and back ends.

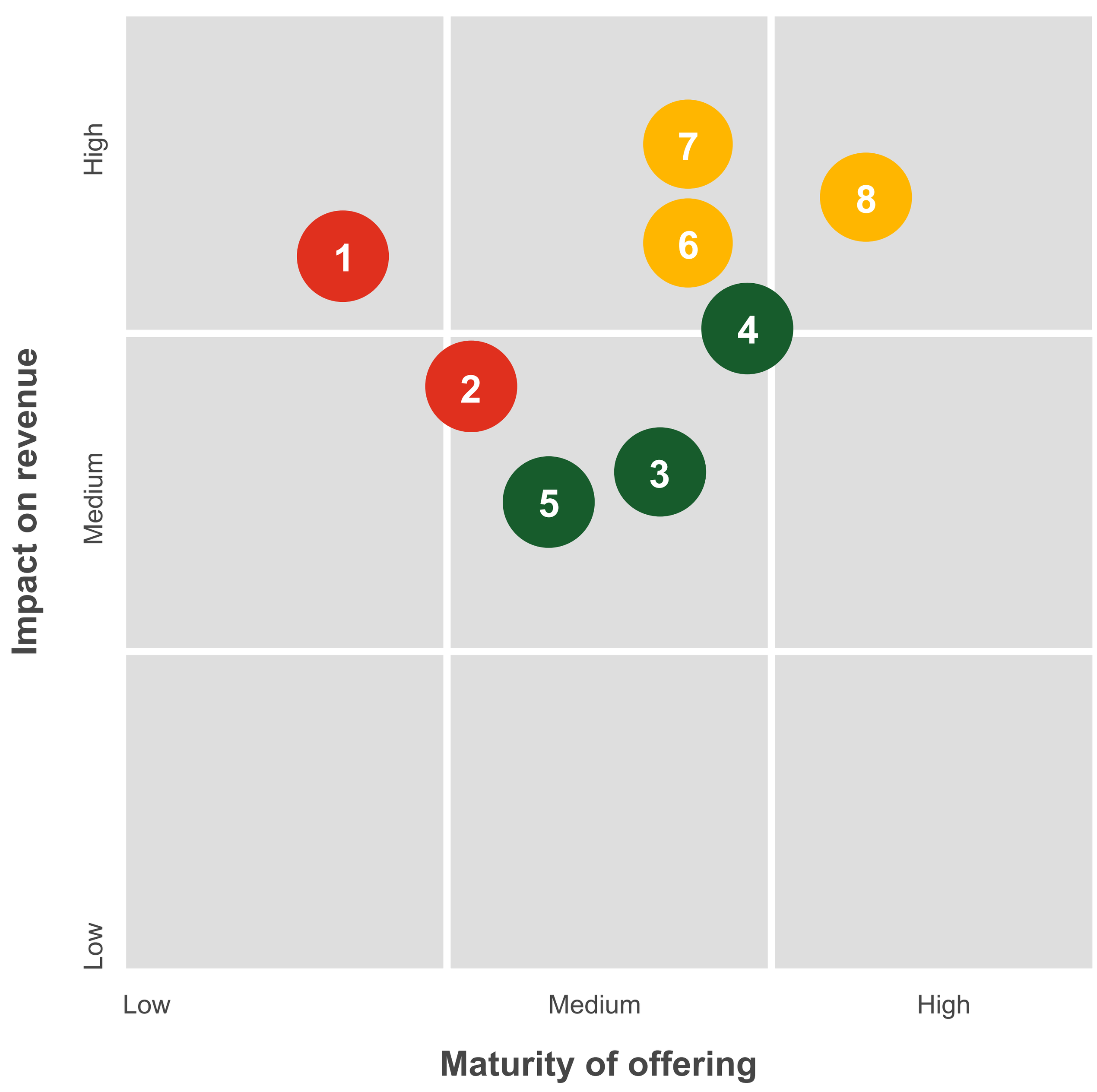
Semiconductor companies looking to benefit from further digitisation and the advent of AI should proceed systematically, through the following carefully thought-out steps: The first step is exploration and learning. Companies must select suitable opportunities for the further understanding and development of new use cases and products through pilot programmes. Second, companies will need to build up capabilities. This requires identifying existing and required capabilities and then creating any missing ones. Designing and rolling out a digital transformation programme to implement new capabilities would follow. Third, companies will need to scale up and identify further digital opportunities and related use cases. Then they will have to define the road map for incorporating those opportunities in the new digital organisation. And firms must leverage the new capability set and organisation to ensure ease of new product development and introduction.

Exhibit 11. Range of opportunities in the value chain



Source: PwC research

Exhibit 12. Opportunities by maturity and impact



Digitisation of product and service offerings

1. Data monetisation – AI-assisted IC design service
2. Product customisation – Chip on demand

Data monetisation – Handling services

Augmented product – AI-assisted IC integration service

Augmented product – IoT platform integration

Augmented product – Device play

Innovative digital business models

3. Open source – IC design
4. X-as-a-service – Compute as a service
5. Marketplace – AI algorithms/models

Razor and blades – Cloud platform and novel computing

Platform – Data

Platform – IoT

Marketplace – AI end-user/application data

Digitisation of value chain

6. Ecosystem sensing
7. AI-assisted design and debugging
8. Fabrication process optimisation

Central control tower

AI-assisted long-term/short-term demand forecasting

Source: PwC research

The road ahead

It is clear that the coming years, and even decade, hold many possibilities for semiconductor firms to profit. Over the forecast period through 2022, we expect that the semiconductor market will continue to grow quickly in every global market to US$575bn.

Of the seven component types, memory chips will continue to maintain the largest market share through 2022, with a large portion of this uptick driven by cloud computing and virtual reality in end-devices such as smartphones.

Further, amid the positive outlook for the worldwide economy, every application market is likely to grow through 2022, led by the automotive and data-processing markets. Lifting these segments will be the demand for chips related to AI.

The semiconductor firms that are able to take the most advantage of this growth and fully realise their market potential will likely be those that harness the possibilities that AI brings. With competition from new startups and entrants from other corners of the tech world, the race to capture the market is only intensifying. Semiconductor firms, in addition to providing chips, must find ways to either monetise new technologies beyond the actual chips themselves or expand into new business models enabled by those technologies. Companies that do so will thrive. Those that don't will be overtaken by more agile competitors.

Contacts

Düsseldorf

Werner Ballhaus
Partner, PwC Germany
+49-211-981-5848
werner.ballhaus@pwc.com

Seung-Min Shin
Manager, PwC Germany
+49-211-981-1228
seung.min.shin@pwc.com

Munich

Marcus Gloger
Partner, PwC Strategy& Germany
+49-89-5452-5572
marcus.gloger@pwc.com

Tanjeff Schadt
Director, PwC Strategy& Germany
+49-89-5452-5521
t.schadt@pwc.com

Sören Tolzmann
Director, PwC Germany
+49-89-5790-5202
soeren.tolzmann@pwc.com

Andris Ogrins
Senior Manager, PwC Strategy& Germany
+49-89-5452-5632
andris.ogrins@pwc.com

San Jose

Conall Dempsey
Partner, PwC US
+1-215-518-0149
conall.dempsey@pwc.com

Mark McCaffrey
Partner, PwC US
+1-408-817-4199
mark.mccaffrey@pwc.com

Anand Sundaram
Director, PwC US
+1-480-319-1931
anand.sundaram@pwc.com

Amit Adalti
Director, PwC Strategy& US
+1-408-624-6714
amit.adalti@pwc.com

Shanghai

Jianbin Gao
Partner, PwC China
+86-21-2323-3362
gao.jianbin@cn.pwc.com

Thank you