Hype Cycle for Semiconductors and Electronics Technologies, 2020

Published: 6 July 2020 **ID:** G00441655

Analyst(s): Gaurav Gupta

Semiconductors are essential building blocks for all electronic equipment and, hence, critical enablers for all IT systems. This Hype Cycle profiles emerging semiconductor technologies to provide a comprehensive view for timely investment in disruptive innovation for competitive advantage.

Table of Contents

List of Tables

List of Figures

Analysis

What You Need to Know

Semiconductor chips are the foundation for all electronic products — from smartphones to medical equipment, to smartwatches, to factory automation robots, to self-driving vehicles. They are, therefore, the cornerstone of the entire IT ecosystem.

Semiconductor and electronics technologies continue to evolve with dynamic end-market demands. Although traditional markets, such as ultramobiles, smartphones, servers and PCs, still provide a significant percentage of semiconductor revenue, high-growth areas, such as electric vehicles (EVs), hybrid electric vehicles (HEVs), advanced driver assistance systems, wearables and security, offer new opportunities. Additionally, increased use of artificial intelligence (AI) in smartphones and servers and 5G implementation will contribute toward a major share of semiconductor revenue. Next-generation storage and data centers will continue to provide ample opportunities for growth. To exploit these end-market demands, chip vendors continue to invest in new and emerging semiconductor technologies, as well as make advances in novel materials and ecosystems.

This report offers an integrated view of semiconductor technologies, which IT end users can use to understand the potential impact and timing for emerging applications.

The Hype Cycle

The semiconductor technologies profiled in this document have been selected based on Gartner's expectation that they will address critical priorities for the electronics and IT industry.

The technologies on this Hype Cycle are the underlying enabling technologies for core benefits, such as:

- **Advanced analytics** provided by deep neural network (DNN) application-specific integrated circuits (ASICs), edge AI, embedded field-programmable gate arrays (eFPGAs), embedded AI, field-programmable gate array (FPGA) accelerators, neuromorphic hardware and sensor fusion.
- **Increased performance** delivered by carbon-based transistors, next-generation transistors, nano-RAM (NRAM), phase change memory, resistive random-access memory (ReRAM), chip-to-chip silicon photonics, and spin-transfer torque magnetic random-access memory (STT-MRAM).
- **Better power efficiency** enabled by energy harvesting from radio waves, gallium nitride (GaN)-on-silicon transistors, gallium oxide (GaO) transistors and silicon carbide (SiC) transistors.
- **Improved and extended communications** leverage cognitive radio, software-defined radio (SDR), mmWave and terahertz waves.

Depending on the maturity of technologies on this Hype Cycle, there are a range of engagement opportunities. In the case of high strategic value, it is crucial to engage early in the Hype Cycle to

improve visibility, establish an intellectual property (IP) position and influence the innovation trajectory.

These significant changes have occurred during the past year:

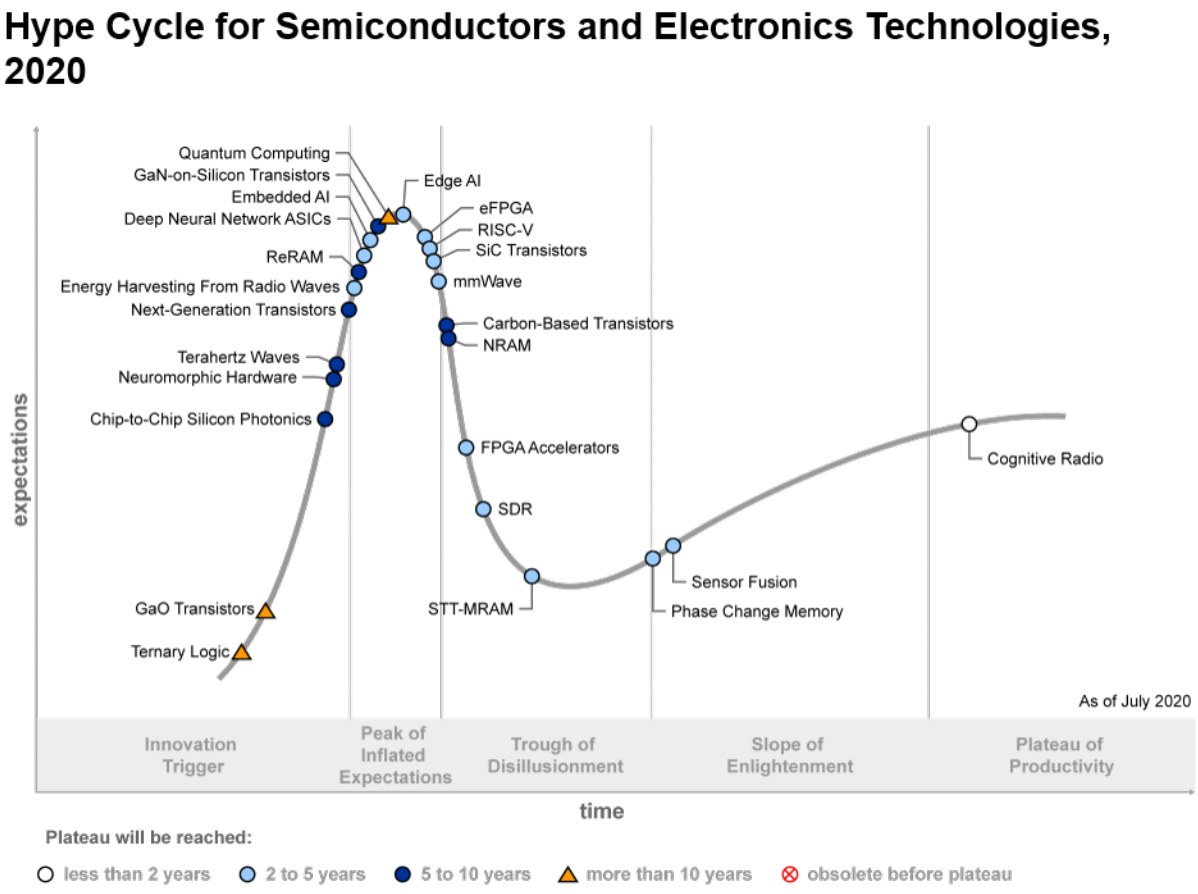
- Edge AI, eFPGA and RISC-V moved into the Peak of Inflated Expectations from the Innovation Trigger due to the strong excitement and buzz around these technologies across suppliers and investors, as evident from the number of announcements.
- Extreme ultraviolet (EUV) lithography and graphics processing unit (GPU) accelerators are past the Plateau of Productivity as they have been adopted in high-volume production and/or applications.
- mmWave is new to the Hype Cycle and has been included due to growing interest in enabling 5G communications.
- Ternary logic has also been added in this iteration of the Hype Cycle as it offers a new, low cost point for computing devices.

Over half the technologies on the Hype Cycle are clustered around the Innovation Trigger and the Peak of Inflated Expectations, and most are more than five years from reaching the Plateau of Productivity. These technologies will need continued capital funding investments, design wins and supplier consolidation before obtaining a sustainable market position. Companies that successfully navigate these challenges can leverage these technologies to cultivate significant benefits and competitive advantages.

The semiconductor industry continues to innovate across a broad range of new technologies for applications such as high-performance computing, AI, automated/advanced driving, and 5G. These applications provide opportunities for new technologies to demonstrate value. Technology adoption is strongly linked to use cases that have the potential to scale and show clear benefit.

IT end users, enlightened by the status and benefits of the technologies detailed in this research, must invest rapidly to accelerate innovations to market and claim competitive advantage. Comprehensive knowledge of the maturity and benefits of these technologies can be utilized for timely adoption to meet business challenges.

Figure 1. Hype Cycle for Semiconductors and Electronics Technologies, 2020



Source: Gartner
ID: 441655

The Priority Matrix

The Priority Matrix maps the time to mainstream adoption of a technology against its benefit rating.

Cognitive radio is expected to deliver a high benefit in less than two years as the Citizens Broadband Radio Service (CBRS) is expected to grow, leading to proliferation of this technology. Both edge AI and phase change memory (mainly Intel's 3D XPoint) have once more been rated as having transformational benefits. IT end users are advised to immediately review the business and technology impact of these innovations and act accordingly.

Other technologies, such as neuromorphic hardware and next-generation transistors, are designated as having a transformational benefit with five to 10 years until mainstream adoption. These technologies involve a completely new architecture and processing elements that will disrupt current mainstream semiconductor capabilities. Organizations should assess these technologies to decide their engagement plans and investment priorities.

Most technologies in the Hype Cycle appear in the "middle ground" of high and moderate benefits, with two to five years and five to 10 years until mainstream adoption. This indicates that the semiconductor industry will see continual improvement over the next decade. IT end users must carefully track, assess and reassess these technologies, aligning their organizations and technology roadmaps to maintain their competitiveness.

Figure 2. Priority Matrix for Semiconductors and Electronics Technologies, 2020

Priority Matrix for Semiconductors and Electronics Technologies, 2020

benefit	years to mainstream adoption			
	less than two years	two to five years	five to 10 years	more than 10 years
transformational		Edge AI Phase Change Memory	Neuromorphic Hardware Next-Generation Transistors	
high	Cognitive Radio	Deep Neural Network ASICs eFPGA Embedded AI SDR Sensor Fusion SiC Transistors	Carbon-Based Transistors Chip-to-Chip Silicon Photonics GaN-on-Silicon Transistors NRAM	GaO Transistors Quantum Computing
moderate		Energy Harvesting From Radio Waves FPGA Accelerators mmWave RISC-V STT-MRAM	ReRAM Terahertz Waves	Ternary Logic
low				

As of July 2020

Source: Gartner
ID: 441655

Off the Hype Cycle

To retain a clear focus on those innovations that will have the greatest impact on business value, Gartner regularly assesses its innovation profiles and retires those that have either lost momentum or matured to being mainstream. The following are the changes to technologies that appeared in "Hype Cycle for Semiconductors and Electronics Technologies, 2019" report:

- Carbon nanotubes and graphene — Replaced by carbon-based transistors, a category that represents both these carbon-based technologies
- Solid-state DIMMs — Suspended from Hype Cycle coverage as phase change memory replaces this technology from the semiconductor perspective for application
- GPU accelerators — Suspended from Hype Cycle coverage as it moved onto the plateau into volume production
- EUV lithography — Suspended from Hype Cycle coverage as it moved onto the plateau into volume production
- Biochips — To be covered in the sensing technology Hype Cycle
- Printed electronics — To be covered in the printing and imaging services Hype Cycle
- Energy harvesting using thermal gradients, smart dust, energy harvesting using mechanical vibrations, and energy harvesting using specialized photocells — To be covered in the power and energy Hype Cycle
- Silicon photonics in chip interconnects — Replaced by chip-to-chip silicon photonics, a category that better represents the technology

On the Rise

Ternary Logic

Analysis By: Martin Reynolds

Definition: Ternary logic is an electronic logic system where the signals have three values, rather than the two values of binary. Ternary logic significantly increases logic density, but also programming complexity, over binary logic.

Position and Adoption Speed Justification: The extra logic level of a ternary system adds a surprising amount of complexity to computing, by increasing the number of logical operations from four to 27. This richness of operation makes ternary logic potentially denser and faster than binary logic, by using fewer bits to encode instructions and fewer clock cycles to complete a task.

There is increasing activity in the field. The USAF published a paper showing significant improvement in encryption speed for small devices.

A number of recent papers have explored ternary logic across multiple fields, including the benefits to quantum computing.

Although ternary logic has been known from the time of the first electronic computers, scaling binary logic has always been easier. Therefore, ternary logic makes only fleeting appearances in the history of computing. A rich description is available at The Ternary Manifesto.

Ternary logic is a long-term prospect, as it will take many years to explore and develop high-performance ternary microprocessors and associated software tools. However, the performance and code size benefits could lead to a generation of small microprocessors that deliver useful cost-performance trade-offs within five to 10 years.

User Advice: Published papers show that ternary logic is an emerging research topic in universities and research organizations. There are no practical implementations today, but we could see startups and prototype devices within the next five years. These devices will focus on low-power applications and will deliver superior performance and capabilities in the same power envelope. The USAF example cited above proposed substantial improvements in encryption. Ternary computing may also emerge from the quantum computing roadmap, although in a different form than for digital logic.

For now, IT leaders with energy-constrained small-device deployments in their portfolio should:

- Gain a basic understanding of ternary computing by following published research.
- Implement an internal development program as the first devices become available to build skills.
- Create a plan and roadmap if ternary computing devices intersect your business' computing needs.

Business Impact: Ternary computing, with a significant investment in software retooling, can deliver at least a double in performance from similar resources. Some functions, as evidenced by our cryptographic example, could increase performance well beyond double. Microprocessors based on ternary computing will initially enable new low-cost points for computing devices by increasing the capability of low-end, inexpensive microcontrollers. These devices will offer significant speedups for certain tasks, including security, over traditional devices. In certain applications, ternary computing will deliver products with superior cost/performance ratios.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

GaO Transistors

Analysis By: George Brocklehurst

Definition: A gallium oxide (GaO) transistor is a type of transistor built from a gallium oxide crystal, which offers a high breakdown voltage, thereby gaining the ability to operate at voltages up to 3kV (a review of the most recent progresses of state-of-art gallium oxide power devices). GaO transistors will significantly improve weight, size and efficiency of electric power trains, power conversion modules and high-power inverters used in renewable energy systems.

Position and Adoption Speed Justification: This technology is still in the laboratory research phase and is likely to be most suitable for lower-frequency high-power applications such as AC/DC

conversion — voltage tolerance up to 3kV reported. At this scale, applications will be in larger transport, e.g., trucks, rail, ships and industrial power. Significant time and investment will be required to pull this technology through to production and as a comparison it has taken several decades for silicon carbide to achieve the level of maturity it has today from this point.

A key step has been the move from lateral to vertically constructed transistors, required to achieve the density and voltage capabilities for a compelling value proposition. Germany's Federal Ministry or Education and Research (BMBF) in 2019 launched a £2M joint project bringing to the table industrial partners for a complete value chain perspective from manufacturing (AIXTRON) to deployment (ABB Power Grids). Meanwhile, FLOSFIA, a Japanese startup, announced results showing 50% improvement in channel resistance, with their "normally-off" transistors, when compared to silicon carbide. The company has a partnership agreement with DENSO and has signed a domestic distribution agreement with Hakuto and Kyoei Sangyo to initially bring a Schottky barrier diode (SBD) to market.

Supply chain progress includes a joint development between AGC and Novel Crystal Technology with aggressive plans for production launch of 650V diodes in 2021. Diode construction is far simpler than transistors but will help establish a supply chain for materials and help pave the way for commercialization of GaO transistors.

Aside from the power handling capabilities, an attraction of GaO is the availability of native substrates (from molten GaO similar to what has enabled silicon manufacture) at relatively low cost — which inhibited gallium nitride development. However, like gallium nitride, the thermal conductivity of GaO is very low, which will have thermal management considerations. Multiple technological, let alone manufacturing, challenges remain to bring GaO transistors to market.

With initial interest from industrial partners and the beginning of a supply chain for materials this innovation profile is moved a step further up the Peak of Inflated Expectations.

User Advice: Gallium oxide transistors are very embryonic, where progress is required through scientific breakthrough. The value proposition is an extension to those wide bandgap technologies (silicon carbide and gallium nitride) further along in the adoption cycle.

- Users with industrial and field operations using larger equipment and vehicles are likely to be earlier adopters. Users in this category should take the time to understand the implications and benefits of extending electrification to higher power use cases.
- Equipment and manufacturing vendors invested in markets where power density is core to long-term success should consider participation in university developments to improve visibility, learning and tracking of this technology.

It is unlikely that GaO will have any significant market presence in the next 10 years.

Business Impact: GaO has significantly greater power density and twice the voltage rating over silicon carbide simplifying high-power electronics, lowering cost and increasing efficiencies. Improvements using GaO for transistors will lead to electronics which provide large efficiency

savings for a wide range of use cases from electric vehicles and transport to industrial grid and renewable energy.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: FLOSFIA; NICT; Novel Crystal Technology; Tokyo University of Agriculture and Technology (TUAT); University at Buffalo

Chip-to-Chip Silicon Photonics

Analysis By: Anushree Verma; Bob Johnson

Definition: Chip-to-chip silicon photonics are a technology that enables high-speed laser-based data transfer between individual chips used in a complex system-in-package where optical components are integrated on the top silicon later at nanoscale dimensions.

Position and Adoption Speed Justification: The original silicon photonics (SiPh) developments were targeted for inter- and intra-die connectivity, by enabling high-speed data transfers without constraints incurred by using traditional copper (Cu) interconnects. However, optical components were 50 to 100 times the size of modern transistors, and were not viable solutions for improving data communications within a single chip.

Recent advances in packaging (heterogeneous integration) open up many new opportunities for expanded use of this technology. The most advanced package technologies today enable the integration of separate “chiplets,” each with its own dedicated function, into a single system-in-package. These chiplets have to communicate at high data rates with one another. While this communication is currently being addressed with conventional copper-based interconnects, there is an opportunity for in-package silicon photonics to address this issue in the future.

Note: Silicon photonics have been implemented as pluggables, and is being used for data interconnects to enable high-bandwidth optical signaling at data centers. Silicon photonics modules are already in use in optical network adapters, transceivers, and now co-packaged switches for applications in data centers and 5G network infrastructure. This is not the application of silicon photonics referred to in this Hype Cycle.

User Advice: Chip-to-chip silicon photonics have the potential for being an enabling technology for data interconnectivity between chips in advanced system-in-package implemented using the most advanced heterogeneous integration technologies. However, this technology is not yet ready for widespread adoption. Technology service providers should monitor advances in Silicon Photonics technology for in-package applications when planning their high-performance system designs and assess whether the technology will be a viable solution to high speed data transmission requirements within a single package.

Business Impact: Chip-to-chip silicon photonics have the potential for being a viable technology for high-speed chip-to-chip data transmission in advanced system-in-package in the future. It has the potential of eliminating existing constraints on system designs imposed by current copper- and silicon-based packaging technologies. However, advances must be made to reduce the size of silicon photonics components so that they can be successfully integrated within a system-in-package.

Silicon photonics have already shown the ability to enable high-performance data interconnect for data center operations. However, to be successful for in-chip and in-package applications, it must surpass the ever-increasing cost/performance threshold presented by alternate copper and silicon technologies.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: GLOBALFOUNDRIES; HPE; Intel; TSMC

Recommended Reading: "Predicts 2020: Semiconductor Technology in 2030"

Neuromorphic Hardware

Analysis By: Alan Priestley

Definition: Neuromorphic hardware comprises semiconductor devices inspired by neurobiological architectures. Neuromorphic processors feature non-von-Neumann architectures and implement spiking neural network execution models that are dramatically different from traditional processors. They are characterized by simple processing elements, but very high interconnectivity.

Position and Adoption Speed Justification: Neuromorphic systems continue to be at a prototype stage. IBM's TrueNorth and exploratory work on multilevel phase change memory technologies, the European Union's Human Brain Project (SpiNNaker and BrianScaleS), and BrainChip's Spiking Neuron Adaptive Processor technology are examples of neuromorphic hardware. Intel has developed a research chip, Loihi, and a range of servers and boards leveraging this chip to address a range of AI workloads: Loihi offers a higher degree of connectivity than competing implementations. Intel has also started training practitioners using its Loihi-based systems, as an early step to future adoption.

There are three major barriers to the deployment of neuromorphic hardware:

- Accessibility: Today GPUs are more accessible and easier to program than neuromorphic hardware; however, this could change when neuromorphic chips (NC) and the supporting ecosystems mature.
- Knowledge gaps: Programming neuromorphic hardware will require new programming models, tools and training methodologies.

- Scalability: The complexity of interconnection challenges the ability of semiconductor manufacturers to create viable neuromorphic devices.

At the moment, these projects are not on the mainstream path for deep neural networks (DNNs), but that could change with a surprise breakthrough in programming techniques, however, ongoing working in developing chips for neuromorphic computing continues, for this reason we have moved neuromorphic hardware closer to the peak.

User Advice: Neuromorphic computing architectures leverage spiking neural networks and have the potential to deliver extreme performance for use cases such as deep neural networks and signal analysis at very low power. Neuromorphic systems are also simpler to train than DNNs, with the potential of in-situ training. Furthermore, neuromorphic architectures can enable native support for graph analytics. Most of the neuromorphic architectures today are not ready for mainstream adoption. However, these architectures have the potential to become viable over the next five years. I&O leaders can prepare for neuromorphic computing architectures by:

- Creating a roadmap plan by identifying key applications that could benefit from neuromorphic computing.
- Partnering with key industry leaders in neuromorphic computing to develop proof of concept projects.
- Identifying new skill sets required to be nurtured for successful development of neuromorphic initiatives.

Business Impact: Rapid developments in DNN architectures may slow advances in neuromorphic hardware but NC holds the promise of enabling extremely lower power AI development. There are likely to be major leaps forward in hardware in the next decade, if not from neuromorphic hardware, then from other radically new hardware designs.

Neuromorphic systems promise of lower power operation makes them uniquely suitable for edge and endpoint devices, where their ability to support object and pattern recognition can support image and audio analytics.

We are in the midst of an extremely rapid evolution cycle, enabled by radically new hardware designs, suddenly practical DNN algorithms and huge amounts of big data used to train these systems. Neuromorphic devices have the potential to drive the reach of AI techniques further to the edge of the network, and potentially accelerate key tasks such as image and sound recognition inside the network. They will require significant advances in architecture and implementation to compete with other DNN-based architectures.

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: BrainChip; IBM; Intel

Recommended Reading: "Emerging Technology Analysis: Neuromorphic Computing"

"Forecast Database, AI Neural Network Processing Semiconductors, 1Q20"

"Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide"

"Forecast Analysis: Data Center Workload Accelerators, Worldwide"

"Product Managers Developing AI Chips Must Clearly Identify Target Markets"

"5 Questions a Product Manager Must Ask When Creating an AI-Enabled Edge Product Strategy"

Terahertz Waves

Analysis By: Bill Ray

Definition: Terahertz (THz) radiation lies in the range 300GHz to 3THz — above microwaves and below infrared. Most of this spectrum is currently unlicensed, allowing THz waves to be used for very wide bandwidth/high-data-rate communications. Important applications also exist in medical, security, industrial and many other uses. Major challenges persist in system design and THz emitter technology, but recent research is showing considerable promise.

Position and Adoption Speed Justification: THz frequencies will remain challenging for most communications applications over distances of more than a few meters. Progress has been made in reducing the size of high-power THz emitters, but it is experimental and largely targeted at the security and health applications of the band rather than telecommunications. THz frequencies do not travel well in air, so can only be used for short-range applications (a meter or so) within the Earth's atmosphere. However, the waves can penetrate common packaging materials such as plastics, fabrics and cardboard, so can be used to see into packaged objects and through clothing. Once through the outer layers, THz waves are effective in optical sensing and analysis of a wide variety of chemicals, which presents a more obvious market for the technology. There is interest in the use of THz waves for short-range, high-speed wireless, including within data centers and for satellite communications, where line-of-sight can be guaranteed and atmospheric interference isn't an issue.

Researchers have demonstrated semiconductor-based lasers capable of emitting THz waves, while graphene has the potential for quick control of both the intensity and the polarization of terahertz light. In 2019, the FCC started to issue 10-year experimental licenses for companies interested in using THz bands, spurring other regulators to start exploration.

Given the increased interest we have moved THz waves along the Hype Cycle, and reduced the time to maturity to less than 10 years.

User Advice: Applications of THz are by no means restricted to communications. Other applications are focused mainly on spectroscopy and on the detection of materials, such as explosives. Ongoing terrorism fears are accelerating development of the technology, but considerable technical obstacles remain.

It is too early for significant action by most users, so R&D functions should monitor the technology. Users in organizations that are more immediately affected (such as physical security screening and manufacturing inspection systems) should educate themselves in the emerging solutions and evaluate initial product offerings and research output. In communications, applications, industrial and military uses of various kinds are likely to emerge first, with mobile communications following.

Business Impact: The impact in some areas of business will be high. Security systems will be affected by this technology. Terahertz manufacturing inspection systems can be expected to reduce costs and improve quality, particularly in products based on composite materials. If the range can be extended then telecommunications infrastructure will achieve significantly higher data rates, which will be important for establishing business success and competitive positioning.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: Advantest; TeraView; Thruvision

Recommended Reading: "Top 10 AI and Sensing Technology Capabilities for Personal Assistant Robots in 2020"

Next-Generation Transistors

Analysis By: Samuel Wang

Definition: Next-generation transistors are new types of transistors extending Moore's Law scalability beyond the 3 nm class of FinFET transistors. It will break the limitation of current generation transistor technology to enable the manufacturing of future chips. Next-generation transistors will come in a variety of structures but all will consist of an electrode that switches current on and off within a channel using an electric field.

Position and Adoption Speed Justification: With a better understanding of the characteristics of 3D transistors, technologists are able to scale FinFET down to 3 nm node class, which should be ready for volume production in 2022; further transistor scaling will require new device structures and innovation of materials. Progress has been made in the following areas:

Nanosheet FET and nanowire FET: The next practical option will emerge by nanosheet and nanowire FET, both called Gate All Around (GAA) transistors. Such transistor structure allows current flow through the nanowire or is pinched off under the control of voltage on the gate electrode that surrounds the nanowire. However, because of their small size, single nanowires cannot carry enough current to make transistors efficient. The solution is a structure that consists of stacked multilayer nanosheets with increased width that are under the control of the same gate and so act as a single transistor. Progresses have been made by large foundries on the development of 3 nm GAA technologies. Samsung Electronics has intended to introduce a 3 nm manufacturing process by 2021 based on GAA transistors. But as all companies are reassessing the financial impact of the COVID-19 pandemic, Samsung's 3 nm production schedule could also be affected.

2D monolayer: Monolayer materials are considered for high-performance transistor innovation due to their atomic thickness, which offers better scalability in comparison to Si and III-V channel FETs. Within the 2D materials family, monolayer black phosphorus-based FET has gained popularity due to its superior transport properties, while research efforts to stabilize BP under ambient conditions are ongoing. Researchers at New York University have disclosed a new way to fabricate atom-thin processors involving patterning metal contacts on 2D monolayer semiconductors in vanishing Schottky barriers. Such use of thermal nanolithography is considered as a breakthrough but is still at initial experimental stage. In a separate research work, scientists at TU Wien, Vienna, have reported the prototype of 2D monolayer transistors that surpassed all expectation by using calcium fluoride as the insulating material. The commercialization of the 2D monolayer transistor to volume production remains to be years away.

Negative capacitance transistors: The concept of negative capacity will have a broad implication for device physics. The existence of negative capacitance has been demonstrated in association with micro-electromechanical system (MEMS)-based switches that can be stabilized within unstable region. Capacitance for the storage of electrical charge normally has a positive value. However, using the ferroelectric material hafnium zirconium oxide in a transistor's gate allows for negative capacitance, which could result in far lower power consumption to operate transistors. The experience of HfO_2-based ferroelectric films are relatively new to the material/device communities and the dynamics of such a thin film have drawn strong interest by various institutes and semiconductor companies.

Nanogap transistor: Another promising proof-of-concept design for nanochips is a combination of metal and air gaps. As gaps become smaller than the mean-free path of electrons in air, there is ballistic electron transport. Therefore, instead of sending electrical currents through silicon, these transistors send electrons through narrow air gaps, where they can travel unimpeded as if in space. Although some studies claim the devices don't use any semiconductor material, practical device may use gates made of silicon run beneath the gap. This approach is still limited at discovery stage.

User Advice: Semiconductor manufacturers should maintain close collaboration with academic institutions and conduct research and development work on next-generation transistors.

Semiconductor equipment manufacturers should evaluate the equipment requirements from various options of next-generation transistors on their particular segments and develop plans to support volume production of 3 nm processes.

Semiconductor design companies need to work closely with foundries on choosing the right technology for chip designs. Specifically, the approach to 3 nm technology by TSMC and Samsung seems to be very different, the true benefits of the cost, density, performance and power dissipation must be carefully evaluated.

Business Impact: Next-generation transistors will have a high business impact on the semiconductor industry, as they are key enablers of continued technology advancement. Next-generation transistors have the potential to provide more than 10 times the improvement in speed, power dissipation and area reduction for future integrated circuit devices, allowing future systems to perform better, become more portable and be manufactured at a lower cost. Without these next-

generation transistors, the economy of semiconductor manufacturing will hit a wall in five to 10 years as Moore's Law stops and products cannot be further improved in cost, performance and power dissipation.

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Sample Vendors: imec; IBM; Intel; Samsung Electronics; TSMC

Recommended Reading: "Invest Implications: 'Market Insight: The Era of FinFET Semiconductor Manufacturing Technology for Mobile Applications Has Arrived'"

At the Peak

Energy Harvesting From Radio Waves

Analysis By: Bill Ray

Definition: Energy harvesting from radio waves covers the generation of electrical energy from the ambient radio waves already common in the environment, making use of Wi-Fi, cellular and broadcast signals to trickle-feed batteries or capacitors.

Position and Adoption Speed Justification: Improvements in this technology have been slow despite the technique being common in the early days of radio. Recent developments owe more to developments in low-power computing than improvements in harvesting efficiency, though such improvements do increase demand. The amount of power being gathered has increased only marginally, while the amount of work to which that power can be put has increased exponentially. RF harvesting also faces a challenge of the ongoing reduction in broadcast power, as cell density increases and transmission efficiency increases. Beamforming techniques now threaten to further reduce the ambient power available.

Beamforming enables a transmitter to direct the radio signal, reducing the scattered signal on which RF harvesting relies. Such a directed signal reduces the ambient energy available for passive harvesting, making the technology easier to use in systems which do not use beamforming (such as Bluetooth).

RFID systems use a form of energy harvesting, in that the reflected radio signal is created by the transmission from the reader. Some companies, notably Farsens, have built harvesting systems that use the signal from an unmodified RFID reader to power sensors, LEDs, and even a screen.

The last year has seen significant developments from market leaders, with Wiliot gaining investment from various supply-chain companies including Amazon. This investment has enabled the company to take its energy-harvesting SoC from prototype to production and discuss customers in retail and shipping logistics. Therefore, we have moved this profile along the Hype Cycle, and reduced the time to plateau.

User Advice: There are few situations where alternative sources of energy, such as solar, vibrational or thermal, are not able to provide comparative supply. The technology is most useful where a known radio source can be harvested, such as the sensors developed by Farsens and Wiliot (RFID and Bluetooth, respectively). These companies use that known source for interaction and power harvesting, ensuring that a minimum amount of power is available.

Energy harvesting of RF signals is a proven technology but lacks a killer product to demonstrate its utility. Users looking at harvesting ambient RF should certainly consider the long-term impact of changes in the radio landscape to mitigate the risk of the harvested signal disappearing, but where a known radio signal is available the technology is quite practical.

Business Impact: Further improvements in energy efficiency, as well as increased efficiency in harvesting RF, will create the potential for everlasting devices dependent only on the continuing presence of radio transmissions. Such devices will prove useful in hard-to-service areas, where solar panels could become obscured and battery replacement is expensive.

Where sensors need to be wirelessly interrogated then the use of the interrogation signal for both power and communications makes sense. Extremely low-cost sensors could (for example) be placed around a building, or a workspace, to report environmental data when interrogated, consuming no power at all at other times.

In more-general deployments, however, the presence of any particular radio transmission (such as broadcast television or 2G GSM) cannot be guaranteed indefinitely, and improved beamforming will reduce the available power, so energy harvesting of RF has a limited window of opportunity.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Farsens; Nikola Labs; Wiliot

Recommended Reading: “Emerging Technology Analysis: Energy Harvesting Enables Autonomous IoT Endpoints”

“Unlock the Innovation Potential of Future Location-Sensing Technologies”

ReRAM

Analysis By: Joseph Unsworth

Definition: Resistive random-access memory (ReRAM) is a type of nonvolatile memory in which the presence or absence of a conducting path within a dielectric between two electrodes is used to determine the stored state. The conducting path is formed either by generating defects in an oxide layer or by the formation of a metallic filament. A broader definition of ReRAM could include phase

change memory (PCM) and magnetoresistive random-access memory (MRAM) but, we have chosen to treat these as separate technologies on the Hype Cycle.

Position and Adoption Speed Justification: ReRAM promises significant performance benefits over other memory technologies including faster read/write speed and lower power consumption. However, the use of specialized materials, such as chalcogenide and silver, to create ReRAMs means that new manufacturing processes need to be developed and optimized, which increases the cost of production, at least in the short term. In long term, through 2D and 3D scaling efforts, it is possible to achieve high density to compete with existing mainstream nonvolatile memory.

There are three main types of ReRAM:

- Oxide Vacancy or Displacement Memory (OxRAM)

In this type of ReRAM, the dielectric material between the electrodes is a chalcogenide glass, the resistance of which is altered by the removal or replacement of oxygen ions by passing a current through it in one direction or the other.

An early variant of OxRAM was Hewlett Packard Labs' memristor, which has never been commercialized as a product despite more recent joint development programs with the likes of Western Digital and SK hynix. However, OxRAM has been available as embedded ReRAM in Panasonic Semiconductor's MCU products for several years. Most recently Panasonic has partnered with Fujitsu Semiconductor, which brought to market a stand-alone 4Mbit serial interface ReRAM as an extension to its electrically erasable programmable ROM (EEPROM) and serial flash portfolio.

- Nanofilament RAM

In this type of ReRAM, the resistance of a silicon-based switching medium between the two electrodes is altered by the formation of a conductive filament, which is formed by applying a voltage between the two electrodes.

Nanofilament ReRAM has yet to come to market although engineering samples in embedded memory form have been produced by Crossbar and its foundry partner Semiconductor Manufacturing International Corporation (SMIC). Most recently, in 2018, Crossbar licensed its intellectual property (IP) to Microsemi (subsequently acquired by Microchip Technology) for "next generation" embedded memory products but these are still in development.

- Conductive Bridging RAM (CBRAM)

In this type of ReRAM, when a current is applied between the two electrodes, a conductive path is created from anode to cathode by the migration of silver ions from the anode into the dielectric material. A reverse current moves the ions back into the anode, removing the conductive path.

CBRAM started life as programmable metallization cell (PMC) memory and has been in development for more than 20 years. In recent years, it has been commercialized and brought to production by a single vendor — Adesto Technologies. Adesto Technologies actively markets and sells 32Kb to 512Kb EEPROM products based on CBRAM technology, which are targeted at specialist markets such as battery-backed IoT devices.

ReRAM will need more investment from major memory vendors and will require a greater technology breakthrough to achieve economies of scale, lower cost and mainstream adoption. In May 2020, GigaDevice announced a patent licensing agreement with Rambus as well as a licensing agreement with Reliance Memory, which is a joint venture between Rambus, GigaDevice and several investment partners. While this is an interesting development, it remains to be seen when ReRAM technology would be brought to market. The applications thus far are limited, with some early use in specialized IoT applications, and there is potential for future penetration in this market. For these reasons, the position of ReRAM on the Hype Cycle has only seen minor progression.

User Advice: Product leaders at mainstream semiconductor vendors should evaluate ReRAM as an embedded memory technology, especially if they are developing solutions targeted at applications that require fast read/write and low power consumption such as battery-backed IoT devices. Including ReRAM in any product roadmap will involve partnering with IP and manufacturing service providers with ReRAM expertise.

Business Impact: ReRAM is a niche technology but has seen some market traction, either as embedded nonvolatile memory in microcontrollers or as an alternative to serial EEPROM and serial flash memory, targeted at specific applications in the IoT, industrial and automotive markets.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Adesto Technologies; Crossbar; Fujitsu Semiconductor; GigaDevice; Microchip Technology; Panasonic; Rambus

Recommended Reading: “Market Definitions and Methodology: Semiconductor Devices and Applications”

“Emerging Technology Analysis: The Future and Opportunities for Next-Generation Memory”

“Market Trends: New Memory Technologies Enhance Energy Savings, Security and Functionality of IoT Devices”

Deep Neural Network ASICs

Analysis By: Alan Priestley

Definition: A deep neural network (DNN) application-specific integrated circuit (ASIC) is a purpose-specific processor designed to execute the DNN computations utilized in a wide range of artificial intelligence applications.

Position and Adoption Speed Justification: DNN ASICS are being used in a divers set of data center, edge and endpoint applications, some examples include object detection and classification

in images and video streams, natural language processing, social media recommendation engines, autonomous vehicles and pharmaceutical analytics.

There are two major phases of DNN-based AI application development:

- Training: Large volumes of known data are processed by the DNN ASIC. These operations are data throughput intensive and typically require the use of floating point math.
- Inferencing: New or unknown data is analyzed by the DNN. These tasks are latency dependent and can utilize integer math.

A majority of training and inferencing tasks currently use GPUs, but the use of DNN ASICs can deliver significantly higher performance and lower power consumption than CPUs or GPUs when executing DNNs.

While it is possible for the same DNN ASIC to be used for both training and inference tasks, devices are being developed that are optimized for a specific task and often for a specific class of DNN. The training phase typically takes place in a data center and leverage large scale designs, typically high power chips, optimized for data throughput. Having developed and trained a DNN-based AI application it is typically deployed on a DNN ASIC optimized for inference operations. These chips may be used in data center deployments but often will be utilized in edge or endpoint systems where there may be constrained formfactors and power resources, requiring highly optimized chip designs.

Many companies have announced plans to DNN ASICs for both training and inference ranging from traditional semiconductor vendors to startups. The large hyperscale cloud service providers are also developing ASICs optimized for their specific DNN-based workloads, examples include Google's tensor processing units (TPUs) optimized for its TensorFlow-based applications.

User Advice: The benefits of DNN ASICs in processing the highly parallel operations required for today's AI-based applications are significant. However, widespread use of DNN ASICs will require the standardization of neural network architectures and support across diverse DNN software development frameworks. Plan an effective long-term DNN strategy comprising DNN ASICs by choosing ASICs and vendors that offer or support the broadest set of DNN frameworks to deliver business value faster. General purpose CPU vendors are also adding new instructions to their CPUs to support DNN-based workloads and these should also be evaluated when assessing the use of ASICs to accelerate these DNN-based applications.

Business Impact: Hardware acceleration will enable DNN-based workloads to address more opportunities in a business through improved cost and performance. Use cases that can benefit from DNNs include video analytics and object detection, image recognition, natural language processing and recommendation systems.

IT leaders deploying deep neural network applications should include DNN ASICs in their planning portfolio. We expect this market to mature quickly, possibly within the three-year depreciation horizon of new systems.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Amazon; Baidu; Google; Graphcore; Hailo; Huawei; Intel; SambaNova Systems; Syntiant

Recommended Reading: "Forecast Analysis: Data Center Workload Accelerators, Worldwide"

"Forecast Database, AI Neural Network Processing Semiconductors, 1Q20"

"Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide"

Embedded AI

Analysis By: Amy Teng; Alan Priestley

Definition: Embedded AI refers to the use of AI/ML techniques within embedded systems to enable analysis of locally captured data. This requirement is particularly critical for electronic equipment where decision latency must be minimized for operational efficiency and safety. It can also enable always-on use cases targeting battery-operated devices requiring low-power operation.

Position and Adoption Speed Justification: There is an increasing demand for embedded systems to analyze and interpret the data they capture by leveraging AI/ML locally.

Virtually all major MCU vendors have expanded their toolchain to include compilers, model conversion tools, libraries and application samples (such as object and gesture recognition) to enable embedded AI. Additionally, the emergence of Tiny machine learning (tinyML) has encouraged many new lightweight ML algorithms. In February 2020, Apple acquired an AI star-up, Xnor.ai, focusing on BNN (Binarized Neural Network), which is a type of tinyML.

Vendors are also enhancing the AI capabilities of their embedded processors by integrating hardware logic blocks into chips to optimize and advance inference performance. Renesas Electronics has introduced an MPU with an embedded Dynamically Reconfigurable Processor (DRP), a programmable on-chip logic block that can be reconfigured via firmware updates. This enables the processor to be easily updated with the latest AI algorithms. NXP has a general-purpose MCU with heterogeneous cores (ARM Cortex M33 and Cadence Tensilica HiFi 4 DSP) targeting audio/video analytics applications.

ARM's Cortex-M55 is the first Armv8.1-M based MCU core with Helium vector extensions focusing at DSP/ML compute capabilities, and Ethos-U55 — the first micro-NPU that will co-work with Cortex-M by providing configurable MCAs and weight compressions. These two technologies together with ARM's software development frameworks enable partners and developers to quickly expand to embedded AI/ML applications by reusing current assets and experiences. Semiconductor vendors are integrating these hardware IP block into their product lineups, and more products are expected to be available for market adoption from 2021.

In addition to the aforementioned vendor activities, we expect the market continue to vibrant throughout the year, as a result we updated its position toward to the peak of HC.

User Advice: Adoption of embedded AI requires a clear workflow and vendor support on tools, especially where the embedded system is used for real-time response and control. As the market is at an early stage of adoption, IT leaders must:

- Determine where (endpoint, edge or cloud) is best to execute AI based data analytics.
- Identify the subset of applications in your OT system or product portfolios that can be meaningfully impacted using embedded AI.
- Evaluate the availability of reference designs that are close to your target application, chip vendors, their solutions and design partners. Focus on their ability of translating and optimizing your trained model into local systems.
- Evaluate the process of updating algorithms — ensure no security vulnerability is created due to changing designs.

Business Impact: Embedded AI enables devices to analyze captured data using AI/ML techniques locally, reducing the need to transfer data to a remote data center for analysis. This can reduce latency and enhance operational efficiency. Companies who own, sell or serve IoT and industrial electronics, ranging from OT machines, factory equipment, IoT sensors to consumer electronics, will be positively impacted depending on inclusion of and the value created by AI.

Initial justification will come from business cases focusing on first-order operational savings, e.g., predictive maintenance — these are the easiest and clearest to define. As adoption picks up, Gartner expects to see additional value created through dynamic and real-time optimization of manufacturing lines to incoming orders and workloads, intelligent buildings that optimize employee productivity.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Arm; Cartesiam; NXP Semiconductors; One Tech; Renesas Electronics; STMicroelectronics

Recommended Reading: “Market Share Analysis: Microcontrollers, Worldwide, 2019”

GaN-on-Silicon Transistors

Analysis By: George Brocklehurst

Definition: Gallium nitride (GaN) is a compound semiconductor whose properties yield high bandwidth and high-voltage transistors. GaN alone has significant manufacturing challenges therefore it needs a substrate. Previously, silicon carbide has been used which is cost prohibitive for

many use cases, therefore moving to a low cost silicon substrate has reduced many barriers to adoption.

Position and Adoption Speed Justification: GaN-on-silicon transistor revenues are still relatively small, $45M compared to a power transistor TAM of $12B, but the supply chain and ecosystem is maturing. A growing number of leading semiconductor vendors have GaN products or supporting chips for its implementation. Most recently, ST the leading vendor of silicon carbide transistors, acquired a majority stake in Exagan, a GaN startup, and announced TSMC as its GaN-on-silicon manufacturing partner.

Highest growth adoption is coming from USB-C charge adapters with multiple third-party vendors announcing new solutions 50% smaller and lighter to those using incumbent silicon technology — OEM solutions will follow. Adoption over the next one to two years will be rapid in this segment taking advantage of 100W capable USB-C power delivery adoption.

Adoption in industrial power supplies has been slower to realize in part due to slower new technology adoption rates in these segments. We are seeing broader field implementations and a growing number of case examples across end markets. These examples are proving opex and capex savings as well as new use cases — and business opportunity — enabled by the technology. Growth rate will build over the next three years in these higher value industrial and automotive use cases to become the dominant segment for this technology.

Within RF infrastructure applications, the key partnership (announced Feb 2019) remains MACOM and ST. The driver here is supporting an attractive cost down versus GaN-on-silicon carbide, ST and MACOM expect 10 to 20 times decrease in cost per power amplifier. This comes at a point when 5G is driving significant content increases with broadening adoption of massive MIMO.

We are close but still not quite at the Peak of Inflated Expectations.

User Advice: This technology is applicable for any roadmap where power density is an inhibitor from a few tens of watts to many kilowatts. High value use cases include edge data centers, on-site renewable energy generation and energy storage, electric transport/vehicles, charging of drone fleets, mobile/warehouse robots and industrial/mechanical automation.

- **Existing use cases (see business impact section):** Prioritize a business case analysis with some urgency. Leverage existing case examples available online or direct from vendors to clarify returns and benefits from implementing the technology.
- **New and emerging use cases:** Reach back in the value chain to the technology vendors to leverage their learning and expertise on what solutions could look like and the likely benefits realized.

Business Impact: GaN on silicon case examples are demonstrating the end-to-end value of this technology:

- **Electric vehicles:** With multiples savings in space, weight and efficiencies as well as cost savings, GaN will have a significant impact on EV adoption. Onboard chargers demonstrating

3x power density improvement support the fast charge roadmap. As high-power versions are introduced GaN will challenge IGBT and SiC for the power train.

- **Data center:** The ability to double the power density of server supplies and halve losses has a significant impact across the data center spectrum. Enablement of edge data centers in new space restricted locations through to 50% utilization improvement and millions of dollars in opex savings in larger data center operations.
- **Renewable energy and microgrids:** The efficiency and space savings afforded by GaN provide material improvement in the ROI of on-premises renewable energy generation and storage. Case examples show 8% round-trip efficiency increases.
- **RF infrastructure:** Cost savings of moving GaN to silicon often an attractive proposition for RF infrastructure at a time when 5G is driving up to 10x content increases in the Remote Radio Head (RRH).
- **Other use cases of note:** Improvements to wireless power transfer for faster and more frictionless charging of drone fleets. Compact yet higher power motor controllers enabling better integration for factory automation.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Dialog Semiconductor; EPC; Exagan; GaN Systems; Infineon Technologies; MACOM (AppliedMicro); Qromis; Texas Instruments; Transphorm

Recommended Reading: "Cool Vendors in Technology Innovation Through Power and Energy Electronics"

Quantum Computing

Analysis By: Martin Reynolds; Matthew Brisse; Chirag Dekate

Definition: Quantum computing is a type of nonclassical computing that operates on the quantum state of subatomic particles. The particles represent information as elements denoted as quantum bits (qubits). A qubit can represent all possible values of its two dimensions (superposition) until read. Qubits can be linked with other qubits, a property known as entanglement. Quantum algorithms manipulate linked qubits in their entangled state, a process that can address problems with vast combinatorial complexity.

Position and Adoption Speed Justification: Quantum computers are not general-purpose computers. Rather, they are accelerators capable of running a limited number of algorithms with orders of magnitude of speedup over conventional computers. These problems fall into a broad category of optimization, where a traditional algorithm would take impossibly long to find a solution. Quantum computers are superior for problems with small input and output, but enormous

combinatorial complexity. Quantum computers will scale using a mix of existing technology, and a combination of new algorithms and conventional computing.

Hardware based on quantum technology is unconventional, complex and leading-edge. Current qubits are good enough to demonstrate the potential of quantum computing, but quantum systems face challenges in scale, noise and connectivity that require yet unknown breakthroughs to offer business value.

The technology continues to attract significant funding, and a great deal of research is underway at many university and corporate labs. Most practitioners are working on gate-model quantum computers, which sequence the qubits through operations that prepare input data and create a solution. In such systems, input data is embodied in the program steps.

Quantum annealing is a style of quantum computing that uses entanglement and superposition in a way more akin to analog computing. D-Wave's quantum annealers, with thousands of qubits, are demonstrating practical solutions, but also face scaling and noise challenges, and do not yet deliver a business advantage.

An alternative solution, Fujitsu's Digital Annealer, offers the equivalent of 8,000 slow, but high-quality, qubits in a rack mount package. Digital Annealers may solve problems otherwise assigned to quantum computers, potentially delaying the point at which quantum computers offer business advantage.

User Advice: In the few known applications, quantum computers could operate exponentially faster than conventional computers. Quantum computers will act as accelerators to problems preprocessed and postprocessed by digital computers.

Early applications will likely be in the class of optimization, such as routing or portfolio optimization. Later applications will address organic chemistry, drug discovery, materials science and code breaking (as prime number factoring).

Quantum computers will eventually compromise today's cryptographic key exchange protocols. Quantum-safe cryptographic algorithms are in the final stages of the standardization process, and should be a midterm strategic initiative for organizations where data must be protected over decades.

If a practical quantum computer offering appears, check its usefulness across the range of applications that you require. It will probably be dedicated to a specific application, and this is likely to be too narrow to justify a purchase. For those customers interested in quantum computing, Gartner recommends the use of quantum as a service (QaaS). QaaS providers such as IBM, Rigetti Computing, Xanadu, D-Wave and others are offering application development and test tools that work with remote hardware.

Business Impact: Quantum computing could have a huge effect, especially in areas such as optimization, machine learning, cryptography, drug discovery and organic chemistry. Although outside the planning horizon of most enterprises, quantum computing could have strategic impacts in key businesses or operations.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: 1QBit; Alibaba Cloud; D-Wave; Google; IBM; Microsoft; QC Ware; QinetiQ; Rigetti Computing; Zapata Computing

Recommended Reading: "Strategy Guide to Navigating the Quantum Computing Hype"

"Emerging Technology Analysis: Act Now on Quantum-Safe Encryption or Risk Losing Deals"

"Quantum Computing Planning for Technology General Managers"

Edge AI

Analysis By: Alan Priestley; Erick Brethenoux; Eric Goodness

Definition: Edge AI refers to the use of AI techniques embedded in IoT endpoints, gateways and edge servers, in applications ranging from autonomous vehicles to streaming analytics. While predominantly focused on AI Inference, more sophisticated system may include a local training capability to provide in-situ optimization of the AI models.

Position and Adoption Speed Justification: Edge AI will be implemented in a range of different ways, depending on the application and design constraints of the equipment being deployed — form factor, power budget (i.e., battery powered vs. mains powered), data volume, decision latency, location, security requirements etc.:

- Data captured at an IoT endpoint and transferred to an AI system hosted within an edge computer, gateway or aggregation point: In this architecture, the IoT endpoint is a peripheral to the AI system. The endpoint acts as a data gatherer that feeds this data to the AI system. An example of this is environmental sensors deployed for a smart agriculture application.
- AI embedded in the IoT endpoint: In this architecture, the IoT endpoint is capable of running AI models to interpret data captured by the endpoint and drives some of the endpoints' functions. In this case, the AI model (e.g., a machine learning model) is trained (and updated) on a central system and deployed to the IoT endpoint. An example is a medical wearable that leverages sensor data and AI to help visually impaired people navigate the world in their daily lives.

The applications that are starting to see increasing adoption of edge AI include those that are latency sensitive (e.g., autonomous navigation), data intensive (e.g., video analytics), and require an increasing amount of autonomy for local decision making. While many of these applications are still in R&D or trial phases, and widespread adoption is at least a few years away, other such as video analytics (leveraging deep learning methods and deployed as deep learning models at the endpoints or in edge servers) are starting to see adoption — driven by the rapid growth in deployment of surveillance cameras and the need for real-time interpretation of captured video streams.

User Advice: Enterprise architecture and technology innovation leaders should:

- Determine whether the new AI developments in machine learning (ML) are applicable to their IoT deployments, or whether traditional centralized data analytics and AI methodologies are adequate.
- Evaluate when to consider AI at the edge vs. a centralized solution. Applications that have high-communications costs are sensitive to latency, or ingest high volumes of data at the edge are good candidates for AI.
- Assess the different technologies available to support edge AI and the viability of the vendors offering them — many potential vendors are startups, which may have interesting products but limited support capabilities.
- Estimate carefully and pragmatically the appropriate level of autonomy and trustworthiness for AI systems deployed on edge systems.
- Assess the risk associated with the nondeterministic nature of many AI techniques where it may not be possible to control or replicate the analysis results.
- Use edge gateways and servers as the aggregation and filtering point to perform most of the edge analytics functions. Make an exception for compute-intensive endpoints, where AI-based analytics can be performed on the devices themselves.

Business Impact: By incorporating AI techniques at the edge, enterprises may be positively impacted as follows:

- Improved operational efficiency, such as enhanced visual inspection systems in a manufacturing setting.
- Enhanced customer experience, with faster execution time, performed at the edge.
- Reduced latency in decision-making, with the use of streaming analytics and migration to an event-based architecture.,
- Communication cost reduction, with less data traffic between the edge and the cloud.
- Increased availability even when the edge is disconnected from the network.
- Enhanced local decision autonomy for edge systems.
- Reduced storage demand through a more reactive exploitation of the data and a better estimate of its potential obsolescence.

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Amazon; Baidu; Google; Huawei; Intel; Matroid; Microsoft; Neurala; NVIDIA; Qualcomm

Recommended Reading: "Exploring the Edge: 12 Frontiers of Edge Computing"

"5 Questions a Product Manager Must Ask When Creating an AI-Enabled Edge Product Strategy"

"Use Edge AI to Drive Revenue Growth, Forecasting, Customer Engagement and Workforce Planning"

"Cool Vendors in AI Semiconductors"

eFPGA

Analysis By: Alan Priestley

Definition: Embedded field-programmable gate arrays (eFPGAs) are third-party IP blocks available for inclusion by semiconductor device developers into ASIC and ASSP designs to provide flexible logic blocks to customize the device for use in a specific application. eFPGAs can also be FPGA die that are mounted on a multidie package with other die, such as ASSPs, MPUs or MCU to provide programmable logic blocks to enable additional functionality to be added to the device.

Position and Adoption Speed Justification: eFPGA can provide equipment designers with the flexibility to increment the functionality to their designs, potentially extending the instruction set of a processing element to add additional processing resources. Or, integrate custom elements of their design that would otherwise have required the use of multiple discrete devices. While FPGAs are stand-alone chips that also provide similar functionality, the ability of an eFPGA logic block to be integrated on-die, on-package, with a multitude of other functional block provides chip designers with a higher degree of flexibility.

With the growth in demand of highly integrated ASSPs and ASICs for use in IoT endpoints, designers are faced with the challenge of developing ASSPs and ASICs that can be used in a range of devices and are also able to support new functionality as the devices evolve and mature. The adoption of eFPGA technology, either as an IP block or as an on-package die provides device developers with the flexibility required to design ASSP and ASICs to current and future IoT devices' design requirements. However, while eFPGAs can bring hardware flexibility to chip designs programming them remains a task for logic designers rather than software programmers.

A number of companies are now specializing in the development of eFPGA IP. In addition, advances in packaging technology enabling the integration of FPGA die onto multidie designs are now also becoming viable for use in many IoT endpoint designs. This is resulting in an increased use of this technology by IoT designers. For this reason, we have positioned eFPGAs in the post-peak region of the Hype Cycle, with the expectation of reaching broad scale deployment within the next two to five years.

User Advice: Flexibility is rapidly becoming a design criterion for many semiconductor devices, especially when developing chips to support emerging AI workloads, technology strategic planners must evaluate the integration of eFPGA technology into their devices to enable design flexibility and

win new business. The market for semiconductor devices that include eFPGA functionality covers a wide range of electronic equipment types, including:

- IoT endpoints
- Smartphones
- Data processing equipment such as servers, storage and networking subsystems
- Wireless communications equipment
- Automotive applications such as infotainment, driver-assist systems and autonomous control systems
- Robotics applications
- Industrial, medical, and test and measurement applications

Technology product managers should leverage eFPGA logic blocks to enable design flexibility and reconfigurability of semiconductor devices to facilitate device reuse in multiple designs. As part of their evaluation of the various eFPGA vendor offerings, product managers must also take into consideration the software tools available to enable developers to configure the eFPGA logic blocks and the integration of these tools into their design workflow.

Business Impact: Small form factors and high-volume production are driving demand for more highly integrated electronics solutions. Increasingly ASSPs and ASICs are being designed to meet the challenges of high volume IoT endpoint production. Many product managers at equipment developers face the challenge that design requirements may change or there is a need to reuse an existing semiconductor device in a new design. These circumstances often require the use of additional semiconductor devices in order to meet the equipment design goals. For many developers, FPGAs have been an answer to this problem; however, in many designs, the use of an additional device may not be possible or desirable — real estate, power, incremental BOM cost, etc. The use of eFPGAs within an ASSP or ASIC design provides a level of flexibility to the core functionality of the device.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Achronix Semiconductor; Flex Logix; Intel; Menta; QuickLogic

Recommended Reading: “Forecast Analysis: Semiconductor Devices With Integrated eFPGA Technology, Worldwide”

“Market Trends: Edge Computing Creates New Equipment Design Challenges”

“Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide”

RISC-V

Analysis By: Bill Ray; Martin Reynolds

Definition: RISC-V is an open, free, instruction-set architecture for embedded-chip designs which may be used by vendors in their chip designs without patent or licensing fees. The architecture has consistent 32- and 64-bit instruction sets. Developers may extend the functionality, while retaining a common core architecture to ensure compatibility across implementations.

Position and Adoption Speed Justification: The advantage of using RISC-V over competitors, such as Arm, is not only the reduction in license fees but also the flexibility that permits designers to extend the chip architecture (something which generally requires a rare and technically challenging architectural license). The core RISC-V IP is freely available but use of the RISC-V logo on commercial products requires membership of the RISC-V International.

RISC-V is the first major new processor core architecture in decades. From the same genealogy as the MIPS architecture, its design reflects modern practices in computer architecture. More important, however, is the relatively complete ecosystem of IP, toolchains and software. Moving applications to RISC-V is a relatively straightforward process and brings significant savings to volume users.

Several major semiconductor companies have invested in products based on RISC-V, including Microchip Technology, which uses a RISC-V-based microprocessor soft core in its PolarFire FPGA SoC, and Samsung Electronics, which has incorporated RISC-V into its Exynos system. Membership of the RISC-V International includes Espressif Systems, Google, Huawei, Marvell, NVIDIA, NXP Semiconductors, Qualcomm, Samsung Electronics and Western Digital. RISC-V is displacing embedded controllers at Western Digital and NVIDIA. SiFive offers RISC-V cores instantiated with low-cost custom logic, reducing barriers to entry for custom chips.

Concerns over restrictions on trade have created a lot of interest in RISC-V from China, where open platforms are seen as an important tool in avoiding reliance on foreign suppliers. With such interest, and millions of RISC-V incarnations shipping, the profile has been nudged over the peak.

User Advice: Arm and competing cores are inexpensive and capable, as long as there is sufficient volume to cover the license fees. As the RISC-V license makes no restriction on commercialization of the core, nor imposes any open-source requirements on the added enhancements (in contrast to projects using the GPL, such as the Linux OS), RISC-V is an attractive alternative to license-bearing cores such as Arm or MIPS, and offers more capability than low-cost cores available as design libraries.

Therefore, RISC-V opens opportunities where either low-volume customization, or high-volume cost reduction, represent a product or business opportunity. For example. NVIDIA and Western Digital are switching their product lines to use RISC-V cores in place of licensed cores from other vendors. SiFive offers RISC-V cores in conjunction with a low-cost, small-volume custom silicon product that can bring a custom chip to lower price points than otherwise possible.

Companies in the MCU or SoC business should evaluate how RISC-V could be integrated into their product offerings, potentially reducing license fees and increasing flexibility in embedded

environments currently dominated by Arm cores. Software companies should be ready to offer RISC-V-compatible versions of their embedded offerings.

Business Impact: RISC-V is unlikely to compete with flagship A-series designs from Arm, at least in the medium term. However, in embedded applications aligned with Arm R or M cores, RISC-V can usefully be combined with custom architectural features that can enable new feature price points in a final product. This capability maps well to IoT endpoints where existing solutions are not feature-competitive when built with competing solutions.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Andes Technology; Bluespec; Codasip; SiFive

Recommended Reading: "Emerging Technology Analysis: Opportunities of Open Instruction Set Architecture RISC-V"

"How Your Business Should Realize the Benefits of the OEM-Foundry-Direct Revolution"

"OEM-Foundry Direct: Evaluate Your ASIC Business Case"

SiC Transistors

Analysis By: George Brocklehurst

Definition: Silicon carbide (SiC) is a compound (silicon and carbon) semiconductor and compared with silicon has 10 times higher breakdown voltage, higher power handling capability and faster high-voltage switching capability. SiC transistors that benefit from these properties are relatively new, brought to production in the last decade. These properties lend SiC transistors to high-voltage power conversion use cases.

Position and Adoption Speed Justification: Manufacturability has been the main adoption inhibitor, as expensive raw materials, a limited supply chain and low yield have kept cost high. Volume is the key to yield improvements and cost reductions to make this technology more accessible to broader markets.

Significant volume ramp up occurred through 2018 and 2019 as Tesla adopted STMicroelectronics' Silicon Carbide transistors in its Model 3 electric power train. This is the fastest growth business for Silicon Carbide transistors and represents one-third of 2019 revenue in this segment. We are seeing growing design wins with global automotive OEMs for the power train and onboard chargers (which will challenge GaN). Volkswagen and Cree announced their partnership in 2019 and Delphi announced a $2.7B order for power train inverters using Cree's transistors. This market is set to grow 45% compound annual growth rate (CAGR) through 2024.

Industrial power modules form the largest end market for SiC with a growing number of cases ranging from 1kW to over 100kW. Growth and adoption rates range from 10 to 20% CAGR through 2024. Faster growth will come from electric vehicle charging infrastructure and electric transport.

Meanwhile, capacity is a growing concern to meet electric vehicle demand. In 2019, Cree announced its intention to invest $1B to establish a new 8-inch SiC foundry. STMicroelectronics, market share leader in SiC transistors, announced its acquisition of SiC wafer manufacturer Norstel, extended its Cree supply agreement to $500M and established a $120M supply agreement with SiCrystal, a ROHM Group company.

In 2019, SiC transistor revenue edged past 5% share of the DC power transistor market. With ramping volumes, a supply chain set for scale and multiple announced design wins this innovation moves closer to plateau.

User Advice: Multiple established vendors are promoting SiC transistor products; while, early low volumes equate to this being a high-cost technology, volume is ramping. Adoption is more compelling where there is a high value placed on compactness and efficiency of power electronics rates in kW.

The costs of implementing SiC transistors should not be compared stand-alone to incumbent technologies; they should be evaluated at the system level, where aspects such as a simplified thermal/switching design are taken into account. Particularly in industrial applications implications for capex vs. opex must be carefully considered and evaluated against risk appetite for change.

SiC transistors are not a drop in replacement to insulated-gate bipolar transistors (IGBTs), as the electrical interface requires different drive conditions and the benefit is realized only when the surrounding circuitry is optimized. As identified in the adoption section, EVs, industrial power supplies, transportation, EV charging infrastructure and renewables are markets where SiC has a strong value proposition. Competing technologies, such as GaN on silicon, which is increasing in power capability, should also be considered and evaluated.

Business Impact: SiC transistors has significant impacts on key use cases summarized below:

- **Electric vehicles (EV):** Efficiency improvements result in lower battery costs and more miles per charge. ST has cited a 20% improvement in miles per charge as well as reduced battery costs.
- **EV charging infrastructure:** SiC transistors enable greater power delivery and, therefore, opportunity to halve or better the charging time achieved by today's silicon solutions.
- **Renewable energy:** SiC transistors halve power loss directly lowering cost of energy generation. In addition, Infineon Technologies quotes a 20% reduction in system costs due to thermal design simplifications.
- **Industrial Power Supplies:** Up to 10% efficiency improvements in kW power supplies provide for compelling opex improvements — lower runtime and maintenance.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Sample Vendors: Cree; Fuji Electric; Infineon Technologies; Microchip Technology; Mitsubishi Electric; ON Semiconductor; ROHM; STMicroelectronics

Recommended Reading: "Market Trends: New Semiconductors Enable Compact and Energy-Efficient Power Converters"

"Why All the Investor Buzz Around Silicon Carbide and Gallium Nitride?"

mmWave

Analysis By: Bill Ray

Definition: mmWave is the frequency band between 30GHz and 300GHz, where the wavelength varies between 10 mm and 1 mm, respectively. However, nearby bands including 24 and 28GHz are also commonly referred to as "mmWave." The 5G mobile telephony standard extends into several mmWave bands, which are also used for point-to-point wireless communication and satellites links.

Position and Adoption Speed Justification: The addition of mmWave bands to the 5G standard has focused a great deal of attention, and investment, on what was (hitherto) a band used only by specialists and niche applications. The U.S., in particular, has been active in deploying 5G networks into mmWave and China has announced plans to allocate mmWave frequencies to its domestic operators. mmWave is attractive because it is largely unoccupied, so it can provide very high capacity, but it's poor penetration and propagation makes deployment of blanket coverage expensive and complicated.

The WiGig standard (802.11ad and 802.11ay) uses mmWave at 60GHz but has achieved little commercial success despite being available for more than a decade. The limited range, and the increasing speed of Wi-Fi, prevented WiGig from being adopted outside a few niche applications (such as wireless VR headsets). However, the adoption of mmWave in 5G has triggered investment in, and mass production of, mmWave antennas and transceivers, reducing the cost of equipment capable of operating in the band.

This innovation is further driven by the release of large blocks of mmWave into the public domain. In 2019 the U.K. regulator Ofcom has released 2.24GHz of spectrum around 24GHz for unlicensed use indoors, while the U.S. regulator Federal Communications Commission (FCC) released 14GHz of spectrum at 60GHz (including the band used by WiGig). These allocations are intended to spur development of new standards, and new applications, for high-speed, short-range, wireless connections.

Investment in 5G is creating new antenna designs and beamforming techniques which will benefit all applications of mmWave, increasing the utility while decreasing the cost of deployment. Phased-array mmWave antennas capable of accurately directing radio signals are a necessity for the next generation LEO satellite constellations. The technology can also be used for highly-accurate location tracking and for using reflected signals to improve the coverage of mmWave systems.

The publicity around 5G has driven mmWave to the top of the Hype Cycle very quickly, and it is now descending as 5G customers find that the bandwidth they get does not match with what was promised by network operator. This, combined with spurious concerns over the safety of mmWave, will push the technology into the trough of disillusion over the next year or two.

User Advice: Anyone looking for point-to-point, or point-to-multipoint, links should include mmWave in the technologies being evaluated. The limited range, and highly-directional, nature of mmWave means that license-free (or light license) systems can operate without fear of interference, and at much lower cost, than licensed systems. Campus networks should consider mmWave point-to-multipoint systems for backhaul of private LTE cells or Wi-Fi access points, especially where clear line of sight is available.

Equipment to run such links is already available but is also developing quickly as enhancements created for 5G networks filter into the rest of the industry. Users should expect to see costs falling rapidly over the next couple of years (following a similar path to the drop in pricing of equipment using microwave frequencies, once that was incorporated into the 4G standard).

Business Impact: The availability of mmWave bands provides plenty of bandwidth. In many instances mmWave will be able to replace wired connections, with ultra-reliable connectivity (using simultaneous connections to multiple access points) making it as reliable (if not more reliable) than wired connectivity. This will enable production lines to reduce the cable loom, increasing flexibility and reducing maintenance costs. Stationary equipment is ideally suited for highly directional mmWave signaling, but with beamforming and other optimization techniques mmWave rapidly becomes suitable for robots and other mobile endpoints.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Metawave; Qualcomm; Renesas Electronics; Texas Instruments

Recommended Reading: “Market Trends: mmWave Opportunities Outside Smartphones”

“Don’t Expect 5G to Replace Wired Access WANs Anytime Soon”

Sliding Into the Trough

Carbon-Based Transistors

Analysis By: Gaurav Gupta

Definition: Carbon-based transistors replace silicon in traditional transistors and offer an alternative solution for performance benefit as Si-based transistors reach practical limits. There are two examples of C-based transistors; graphene and carbon nanotubes. Graphene is a one-atom thick material of pure carbon, bonded together in a hexagonal honeycomb lattice. A carbon nanotube can

be thought of as a sheet of graphene rolled into a cylinder. The rolling-up direction of the graphene layers determines the electrical properties of the nanotubes.

Position and Adoption Speed Justification: Graphene is a hard material to create, as arranging carbon atoms in a two-dimensional hexagonal lattice at a fairly large scale is difficult. Material quality can drastically decrease with just one defect. Graphene field-effect transistors (GFETs) take the typical FET device and insert a graphene channel tens of microns in size between the source and drain. Graphene transistors have high device sensitivity and superior conductivity. Another issue is lack of band gap in graphene that makes it very hard to turn the current off once it starts flowing, a major roadblock for logic operations, which require on-off switching. Researchers have been working to find solutions to this problem, but compounded with a lack of fully integrated supply chain, graphene is far from commercial application.

Carbon nanotubes (CNTs) with semiconductor properties offer the promise of small transistors with high switching speeds in future semiconductor devices, while CNTs with metallic (conducting) properties hold the promise of low electrical resistance that can be applied to the interconnections within integrated circuits. Research indicates carbon nanotube FETs have properties that promise around 10 times the energy efficiency and far greater speeds compared to silicon. CNTs can be single or multiwalled depending on the number of graphene layers and as a result have different strength and efficiency. Currently, there are mixed opinions whether or not CNT transistors would maintain their impressive performance at extremely scaled lengths. But when fabricated at scale, the transistors often come with many defects that affect performance, so they remain impractical. Currently there is no technology for their mass fabrication and high production cost.

User Advice: Semiconductor opportunity will be available for next-generation transistors beyond 5 nm. C-based transistors have been moved forward in their position on the Hype Cycle toward the Trough of Disillusionment as they are past their peak of expectation, and now researchers and industry experts are facing reality. Target audiences that will require these semiconductors must continue to work on fabrication at scale to resolve issues with mass production. Additionally, alternative next-generation transistor solutions are evolving that can challenge their position.

Business Impact: There is potential for a huge impact, particularly when silicon devices reach their minimum size limits — expected during the next five to 10 years. Wireless communications is an area where these technologies will be really beneficial due to high current carrying capability in a small area. An example of current commercial application includes Nantero's NRAM, which leverages carbon nanotube technology.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Fujitsu; Graphenea; imec; IBM; Intel; Nano-C; Samsung Electronics; TSMC

Recommended Reading: "Emerging Technology Analysis: Carbon Nanotubes Will Drive the Next Generation of Semiconductor Devices"

“Emerging Technology Analysis: Carbon Nanotubes and Graphene Are Indispensable for Future Electronic Products, So Act Now”

“Emerging Technology Analysis: Graphene Just May Be the Material of the Future”

“Emerging Technology Analysis: Graphene Is the Catalyst to Extend Semiconductor and Electronics Innovation”

NRAM

Analysis By: Joseph Unsworth

Definition: Nano-RAM (NRAM) is a nonvolatile, random access memory technology, which works by changing the relative position of carbon nanotubes (CNTs) deposited on a substrate surface to switch the storage state of the memory cell. As CNTs stick together, the resistance of the memory cell is low, representing one state. Conversely, as the CNTs separate, the resistance of the memory cell is high, representing the alternative state.

Position and Adoption Speed Justification: NRAM has been touted as a “universal memory” because, in addition to its nonvolatility and byte addressability, it has the potential to be high endurance, high reliability, high speed, high density, low power and both 2D and 3D scalable. It therefore has the potential to compete on both cost and performance with the current mainstream memory solutions that utilize a combination of SRAM, DRAM and flash memory technologies.

Nantero is the inventor of NRAM technology, which it has been developing for almost 20 years. Commercialization of NRAM-based products seemed a step closer when, in 2016, Nantero licensed NRAM technology to Fujitsu Semiconductor and foundry Mie Fujitsu Semiconductor. Most recently, in mid-2018, Fujitsu announced that it would develop embedded NRAM-based logic chips first, to enrich its embedded solutions, with plans to begin volume production in 2019. Longer term, once it’s embedded NRAM has proven its capability, Fujitsu will reportedly consider producing stand-alone NRAM chips. However, so far there has been no update from the company on development progress and in its press releases, thus a degree of skepticism remains over this roadmap and time frame.

Nantero continues to receive strategic investment from its development partners and potential customers (including the investment arms of Dell Technologies and Cisco), and periodically makes new announcements about the progress of NRAM. However, NRAM products have yet to come to market and, until they do, the technology will head toward the Trough of Disillusionment on Gartner’s Hype Cycle. To compete with established and emerging memory technologies at scale, NRAM will have to move to a multilayer architecture. As an NRAM cell currently requires an associated switch transistor, such a development will take time.

User Advice: NRAM has the benefits of nonvolatility, low power consumption, high speed and potential low costs. It could be used widely to replace SRAM, DRAM and flash memory, initially in applications that require relatively low density but fast access. However, the development of any emerging memory takes time and its advocates must be able to prove that it can be manufactured

reliably and in high volume and that it performs as promised, before it can be adopted widely. This is especially true when new materials are involved and if new process equipment is required.

Technology leaders should monitor the progress of NRAM technology development and commercialization before including NRAM in their product development roadmaps. A key indicator of possible future traction in the market will be strategic investments by Tier 1 semiconductor vendors, especially memory specialists. Until that happens, NRAM is likely to remain a specialist niche technology.

Business Impact: Currently, the most likely use of NRAM technology in the near term is as an embedded memory in markets such as automotive, industrial and IoT applications.

Longer term, NRAM technology could prove disruptive in the established markets for stand-alone SRAM, DRAM and flash memory but the economics of these high volumes, commodity memory markets make the barriers to entry for a new technology extremely high.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Fujitsu Semiconductor; Nantero; United Microelectronics Corporation (UMC)

Recommended Reading: "Market Definitions and Methodology: Semiconductor Devices and Applications"

"Market Trends: New Memory Technologies Enhance Energy Savings, Security and Functionality of IoT Devices"

FPGA Accelerators

Analysis By: Alan Priestley

Definition: Field-programmable gate array (FPGA) accelerators are server-based, reconfigurable computing accelerators that deliver extremely high performance by enabling programmable hardware-level application acceleration.

Position and Adoption Speed Justification: FPGA accelerators feature a large array of programmable logic blocks, reconfigurable interconnects and memory subsystems that can be configured to accelerate specific algorithmic functions. This allows FPGAs to offload tasks from the main system processor. In the data center, FPGAs can be used in a range of use cases that require applying consistent processing operations to large volumes of data, such as high-frequency trading (HFT), hyperscale search, video analytics and DNA sequencing. For example, Microsoft is leveraging FPGAs for search analytics and networks, and Illumina's FPGA-based DRAGEN Bio-IT Platform enables high-performance genome-sequencing workflows.

FPGAs are typically configured using hardware programming languages, such as register transfer level (RTL) and VHSIC Hardware Description Language (VHDL), that are very complex to use; this has held back widespread adoption. However, major FPGA vendors (Intel and Xilinx), along with a number of startups such as Mipsology and Swarm64, are working to address this with libraries and toolsets that enable FPGAs to be configured using software-centric programming models.

Adoption is also becoming easier, helped by frameworks such as OpenCL that lower the time and skills required to use FPGAs. Workloads like deep learning (inference) and easier access to development platforms, exemplified by Amazon Web Services' FPGA-enabled instance types, are also driving adoption of FPGAs within the data center.

Today, the biggest growth opportunity for FPGAs in the data center is in the inference portion of deep-learning workloads. Given the evolving nature of this new use case and the maturing of the surrounding software ecosystem, FPGA accelerators have been positioned pretrough.

User Advice: FPGA accelerators can enable dramatic performance improvements within significantly smaller energy consumption footprints than comparable commodity technologies. I&O leaders need to evaluate applicability of FPGA accelerators by:

- Identifying application subsets that can be meaningfully impacted using FPGAs.
- Evaluating the availability of FPGA-based hardware for use in data center server deployments, such as FPGA-based PCIe add-in cards.
- Outlining costs associated with necessary skill set and programming challenges of FPGAs and the maturity of the software-centric programming toolsets.
- Leveraging cloud-based FPGA services to accelerate development.
- I&O leaders should use FPGA accelerators when:
- Preconfigured solutions exist that can help dramatically transform key workloads (e.g., financial trading analytics, genome sequencing, etc.).
- Algorithms will evolve requiring frequent updates at the silicon level that can be utilized by broader applications.

Business Impact: FPGAs can deliver extreme performance and power efficiency for a growing number of workloads. They are well-suited for AI inference workloads as they excel in low-precision (8 bit and 16 bit) processing capabilities in energy-efficient footprints. While programmability continues to be a major challenge, limiting broader adoption of FPGAs, I&O leaders should evaluate FPGA-based solutions for genome sequencing, real-time trading, video processing and deep learning (inference). I&O leaders can further insulate against risks by utilizing cloud-based infrastructures for provisioning FPGAs (e.g., Amazon EC2 F1 instances, Microsoft Azure, Baidu cloud).

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; Baidu; Intel; Microsoft Azure; Xilinx

Recommended Reading: "Forecast Database, AI Neural Network Processing Semiconductors, 1Q20"

"Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide"

"Forecast Analysis: Data Center Workload Accelerators, Worldwide"

"Top 10 Technologies That Will Drive the Future of Infrastructure and Operations"

SDR

Analysis By: Bill Ray

Definition: Software-defined radio (SDR) moves the decoding of received signals from traditional hardware into software running on generic microcontroller units (MCUs) or graphics processing units (GPUs). This allows devices to switch between radio protocols. SDR is most attractive where standards are changing and uncertain, or where multiple standards are required. Smart antennas are an important part of SDR and are typically implemented as an array of elements with programmable characteristics, such as field width, frequency and waveform shape.

Position and Adoption Speed Justification: SDR has great potential to reduce hardware complexity, as less radio hardware is required to support multiple protocols, as well as future proofing existing designs. SDR is a standard feature of testing equipment and products designed to educate users about radio technologies. The rapidly-evolving 5G standards has pushed companies including Qualcomm and NXP to expand their interest in SDR.

SDR can enable a regular PC or an inexpensive, specialized platform (such as the LimeSDR) can interact with a vast array of radio systems. Alternatively, SDR can be used to enable in-field upgrades to custom semiconductors, ensuring they can adapt to changing standards. However, software processing is inherently less efficient than dedicated hardware, so SDR makes most sense where support for many different wireless standards and protocols is required, or where support for a specific standard is necessary but will rarely be used.

Practical SDR implementation relies on a high-performance digital signal processor (DSP) technology or DSP functions implemented in GPU cores. Current DSP-based processors consume too much power for use in battery-powered handsets, but improved software and hardware is steadily reducing the power consumption.

In October 2019, Vodafone started deploying SDR base stations supporting Open RAN in the U.K. These base stations support 2G, 3G and 4G, and are providing rural coverage at significantly lower cost than products using specialist hardware. Qualcomm has also adopted SDR in its small-cell radio product, the FSM100xx, to provide future proofing against variations in the 5G standard.

Given these developments we have moved the technology toward the Trough of Disillusionment, and reduced the time to plateau.

User Advice: Evaluate SDR technology as part of your network migration plans for all kinds of wireless infrastructure, because it will, in the short term, reduce network complexity and lower overall costs. Manufacturers of mobile devices and communications equipment that include multiple radios should pilot SDR technology as it becomes available. Adopt it when the economic case is proven and when the performance and power consumption are sufficient.

Business Impact: SDR has the potential to reduce the production cost of mobile devices and will allow them to be brought to market more rapidly. Support for legacy standards, such as 2G (GSM) and 3G (W-CDMA), will be required in most telephone handsets and base stations for many years, but will rarely be used, so could usefully be migrated into software. SDR will have an important role in the Internet of Things (IoT) when costs become low enough. Some devices and "things" will have a required lifetime of a decade, and support for particular cellular and other wireless standards will change considerably over this period. SDR will allow embedded things to be updated with the latest frequencies or other standards to guarantee a long operating life in the field.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Lime Microsystems; NVIDIA; NXP; Qualcomm; Vodafone

Recommended Reading: "The Top 10 Wireless Technologies and Trends That Will Drive Innovation"

"3 Technologies Critical to 6G Network Product Roadmaps"

STT-MRAM

Analysis By: Joseph Unsworth

Definition: Spin-transfer torque magnetic random-access memory (STT-MRAM) is a type of nonvolatile memory (NVM) that uses the magnetic properties of controlled electrons to store data. The electrons sit in a magnetic layer, where a property described as spin can be manipulated and read.

Position and Adoption Speed Justification: STT-MRAM products are commercially available in both stand-alone and embedded form and have realized some commercial success in the industrial, military, automotive, enterprise storage and IoT markets.

Everspin Technologies is currently the only vendor engaged in volume production of high-density stand-alone STT-MRAM; as of 2020, its 1Gbit density is in mass production along with its 256Mbit density. The target market for Everspin's 256Mbit and 1Gbit is as a DRAM/NAND flash replacement/ supplement in enterprise storage. Everspin has collaborated with GLOBALFOUNDRIES to produce

an embedded STT-MRAM process on 300 mm wafers. Everspin, and thus discrete STT-MRAM sales, struggled in 2019 due to the loss of a key automotive customer. However, Everspin continues to diversify its customer base and gain more traction in server and storage OEMs.

Avalanche Technology also has stand-alone products available, but these are relatively lower density, 1Mbit to 32Mbit, devices targeted as NOR flash replacements. Avalanche has foundry manufacturing agreements with Sony for its stand-alone products and with United Microelectronics Corporation (UMC) for embedded STT-MRAM.

There has recently been an uptick in activity to develop an embedded STT-MRAM ecosystem. Spin Memory (formerly known as Spin Transfer Technologies) has reinvented itself as an STT-MRAM IP provider, and manufacturers such as TSMC, Samsung, Intel and Semiconductor Manufacturing International Corporation (SMIC) are working on providing embedded STT-MRAM offerings.

There are two main applications for embedded MRAM:

- Replacing embedded flash in microcontrollers (MCUs) because flash is hard to shrink past the 14 nm process node.
- Replacing embedded SRAM in high-performance processors as second- or even third-tier cache memory.

Gartner has positioned STT-MRAM on the Hype Cycle closer to the Trough of Disillusionment. Whether STT-MRAM can quickly move through the trough will depend on its ability to migrate from a logic to a memory manufacturing process, drive 4Gbit and higher densities as well as attract other high-volume memory manufacturers.

User Advice: Technology leaders at processor or MCU vendors should collaborate with STT-MRAM vendors to evaluate the possibility of incorporating STT-MRAM as embedded memory to improve the performance of their products. Technology leaders at storage vendors should consider complementing QLC NAND flash with discrete STT-MRAM as a means to improve performance and reliability.

Business Impact: Stand-alone STT-MRAM could be used as an SRAM, DRAM and low-density flash memory replacement. It can also be used as embedded NVM to replace embedded SRAM, embedded DRAM and embedded flash memory.

Stand-alone STT-MRAM is a potentially good replacement for low-power SRAM and EEPROM because it consumes less power and has a lower-cost cell structure. Although STT-MRAM is nonvolatile, so consumes less power than DRAM at system level, its density is relatively low and so it cannot compete on cost as a stand-alone memory technology. In practice, the initial application of stand-alone STT-MRAM will be as a replacement for lower density, nonvolatile DRAM (nvDRAM) in industrial or military solid-state drives (SSDs). These are niche markets and relatively small.

STT-MRAM has little chance to replace NAND flash as a mass storage memory because of the latter memory's cost-efficiency. Currently, 1Gbit MRAM will see more use across industrial, automotive

and storage applications. It's also expected to see initial applications in the battery-backed IoT as code or data storage.

Embedded STT-MRAM is a good replacement for embedded NOR (eFlash), eSRAM and eDRAM because of its faster writing speed and lower power consumption. MCUs/application processors with embedded STT-MRAM can have lower power consumption, higher cache density, better performance and smaller chip sizes than conventional processors.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Avalanche Technology; Everspin Technologies; GLOBALFOUNDRIES; Intel; Samsung; SK hynix; Spin Memory; Toshiba; TSMC; United Microelectronics Corporation (UMC)

Recommended Reading: "Emerging Technology Analysis: The Future and Opportunities for Next-Generation Memory"

"Market Trends: New Memory Technologies Enhance Energy Savings, Security and Functionality of IoT Devices"

Phase Change Memory

Analysis By: Joseph Unsworth

Definition: Phase change memory (PCM) is a type of semiconductor memory based on the conductive properties of a material in its amorphous or crystalline states. The state is changed by means of a pulse that applies a thermal profile to the material.

Position and Adoption Speed Justification: Phase change memory is capable of competing with 2D NAND flash memory in terms of cost and density; however, when compared against 3D NAND, it is roughly one-eighth the maximum density and about 10 times more expensive per GB. PCM is built in the metal layers and not on the surface of the silicon, freeing up the silicon for control, management and fast local cache. After five years of commercialization, the technology is currently only in its first generation, with Intel announcing that the second generation is due later in 2020.

PCM is in volume production as 3D XPoint by Intel uses a first-generation 20nm, two-layer stack manufacturing process to achieve a density of 128Gbit per die. The technology was co-developed with Micron Technology, with a key feature being the addition of a selector diode to the memory cell, eliminating a major barrier to scaling the technology. Intel has branded the technology Optane, with its use as an SSD or persistent-memory DIMMs, while Micron debuted its first product as a NVMe PCIe SSD in 4Q19. The Intel and Micron joint venture has largely ended, with the companies no longer cooperating on the technology after the second generation, although the foundry manufacturing relationship for Intel was extended in 1Q20.

In terms of performance, 3D XPoint is superior to NAND flash in read/write speeds, latency, input/output operations per second (IOPS) and endurance. It also has a very consistent, sustained performance, which does not degrade over time. Compared to DRAM, 3D XPoint is slower but has lower power consumption and at least half the price per GB given its larger densities. For these reasons, 3D XPoint initially appeared in cache memory in servers and NVMe PCIe SSD storage. Intel is now shipping its DC persistent-memory DIMMs that operate as server main memory, albeit requiring substantial software and platform enhancements only available in Xeon second generation of processors. These systems are capable of managing and processing large databases stored entirely in memory in a less expensive, small footprint than previously possible.

The use of 3D XPoint in servers as NVDIMMs expands the directly addressable physical nonvolatile memory and reduces the amount of DRAM needed in the system. 3D XPoint can also be used to complement NAND-based SSDs in data centers acting as a cache for metadata or as a tier of high-performance storage. In PCs, it is available as a low-density cache module to complement HDD or integrated with lower-cost QLC NAND flash as a high-performance PC SSD.

3D XPoint is transitioning from being classed as an emerging memory to become a more mainstream memory technology, thanks to successful commercialization of the technology by Intel in data center applications. Although it is still in Generation 1 and provided by only two suppliers, several companies such as Alibaba Group, Dell Technologies, Google, Hewlett Packard Enterprise, Lenovo, Oracle and SAP, have announced that they have commercialized or are optimizing for the technology. Gartner has therefore moved PCM's location on the Hype Cycle past the trough.

User Advice: Storage, server and independent software vendors must evaluate the technology's scalability and impact on infrastructure in order to justify investment. To fully exploit the benefits of the technology will require ecosystem optimization and a consistent effort to educate the end customers on its value; this will require considerable effort and time.

PC vendors need to align and coordinate product development efforts with Intel's roadmap for PC products with demand for high performance.

Business Impact: PCM, specifically 3D XPoint technology, could alter price, performance and density of current storage and compute platforms. Initially, 3D XPoint technology will be used as server cache memory and NVMe PCIe SSD storage and potentially as a directly addressable persistent DIMM for use in servers with corresponding Intel chipset support. This new tier of high-performance storage and/or new tier of dense, persistent memory can have compelling benefits for applications such as:

- SQL/NoSQL data stores — SQL-based in-memory databases, such as SAP HANA; NoSQL-based in-memory data grids, such as Redis and memcached; and document and table style databases, such as MongoDB and Apache Cassandra
- High-performance computing (HPC) workloads — HPC technical and commercial workloads in academia, research, hyperscale cloud and enterprise organizations
- Big data applications — Apache Spark-based or Apache Storm-based applications for real-time analytics processing

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Intel; Micron Technology

Recommended Reading: "Determining the Data Center Opportunity Created for 3D XPoint Persistent Memory"

"Top 10 Data and Analytics Technology Trends That Will Change Your Business"

"2019 Strategic Roadmap for Storage"

"The Future of High-Performance Storage"

"Emerging Technology Analysis: The Future and Opportunities for Next-Generation Memory"

Climbing the Slope

Sensor Fusion

Analysis By: Amy Teng

Definition: Sensor fusion is a process that aggregates and "fuses" many disparate sensor inputs to increase sensor data accuracy and/or sensing coverage for the system to develop insights and decisions. A sensor fusion solution usually includes a set of sensors, a hardware sensor hub, a fusion engine and a software sensor fusion stack.

Position and Adoption Speed Justification: Sensor fusion for automotive and industrial systems has been prevalent for decades. Sensor fusion also exists in smartphones and tablets where the sensor data has been interpreted by OS allowing app developers access through APIs.

During the past few years, sensor fusion has evolved to include lidar, radar and visual sensing for autonomous cars, SLAM (simultaneous localization and mapping) for drones and robots, and six degrees of freedom (6DoF) visual and 3D audio immersion for (head mounted displays) HMDs.

Technology evolution of autonomous things and Internet of Things (IoT) will keep on pushing the number and the diversity of sensor and sensing technology to a higher level, driving continuous enhancement of sensor fusion technology. Sensor fusion algorithms' development has been benefited by the growing leverages of artificial intelligence/machine learning (AI/ML) technology. When this combined with improved sensor accuracy and advanced computational power of sensor engine, sensor fusion can analyze faster and better than before. Therefore, we moved its position a bit forward

User Advice: Sensor fusion software stacks can be derived from three resources: open source, sensor companies and software companies who focus on serving customers who have sensor

fusion requirement. Sensor fusion software can be ported to a variety of different hardware platforms, ranging from application processors, general purpose microprocessors and microcontrollers, programmable logic, integrated combo sensors, and purpose-built devices like Microsoft's customized holographic processing unit (HPU) for HoloLens. Consider discrete, low-power sensor hub solution to offload the process of main processor when designing for power constraint and long battery life applications.

Build a central platform to manage sensor fusion algorithms across different products and applications to gain reuse efficiency. For example, a smartwatch and wristband can share same algorithm for detecting a wearer's gesture and posture, the two devices can also share the same voice actuated algorithm with smart speakers.

Leverage AI/ML to accelerate the pattern recognition of sensor signals and shorten sensor fusion development time.

Sensor fusion can also offer reduced size and weight, which is particularly important in drone applications.

Business Impact: Sensor fusion technology is a key element of enabling real-time contextual analysis because it comprehends the status of a system. Additionally, leverages AI/ML and clever manipulation of different types of input sensors to train model will explore new algorithms, resulting in new insights or new applications unseen before. For example, fusion of sound waves and surface vibration of sound source can create a new type of noise-cancellation device.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: CEVA (Hillcrest Labs); CyweeMotion; Kionix; Knowles; NXP; Qualcomm; QuickLogic; Renesas; STMicroelectronics; TDK (InvenSense)

Recommended Reading: "Cool Vendors in Novel Sensors"

"Market Trends: Supplying Intelligent Sensors for the IoT Takes Cooperation, Integration and Software"

Entering the Plateau

Cognitive Radio

Analysis By: Bill Ray

Definition: Cognitive radio allows wireless systems to dynamically negotiate spectrum to use, either by detecting locally congested bands or with recourse to a cloud database, and to automatically switch frequencies, protocols and modulations.

Position and Adoption Speed Justification: Cognitive radio is easy to implement in a limited form. The venerable Digital Enhanced Cordless Telecommunications (DECT) standard for wireless telephones has been monitoring available frequencies to find a suitably vacant band since 1982. Modern Wi-Fi routers also search the available bands to ensure their chosen frequency isn't already occupied. However, these systems can only detect those using the same standard, so their cognition is severely limited. Detecting any radio user in the same band is much more complicated, and it suffers from physical limitations when an existing user may be hidden behind the terrain. Scanning for frequencies requires complex hardware and increases cost. Evolving cognitive radio systems have focused on the database approach using detect-and-avoid as an additional mechanism and to enhance the database with crowdsourced feedback.

These systems require legislative changes permitting shared access to radio frequencies. The first systems were used to avoid interference with TV transmissions (White Space networks), but the launch of CBRS in the U.S. (in September 2019) now permits private LTE and 5G networks to be deployed without a spectrum license.

Other countries have not, so far, copied this approach (generally adopting a light-licensing regime instead) but as CBRS grows (and demonstrates the utility of the technique) we expect to see database-driven cognitive radio systems proliferating. Therefore we have moved the technology along the Hype Cycle and reduced the time to plateau below two years.

User Advice: CBRS systems are already being deployed by enterprises, and mobile operators have expressed their plans to make use of the technique (and band) for extending their public network coverage.

Include cognitive radio in long-term spectrum resource planning, as it can enable a better use of finite radio resources and improve user experience (depending on what networks are available to the user). Consider existing solutions for campuswide or citywide deployments and point-to-point connections in rural areas, where the delay in setup and overhead of frequency monitoring will have a limited impact. Cognitive radio systems have to change frequencies on demand, requiring greater complexity in design and operation, and higher cost.

Business Impact: End users will benefit from a seamless wireless experience as they will be less likely to ever be completely out of radio coverage. For communications service providers (CSPs), this could enable less costly radio coverage, as multiple legacy access networks could be leveraged to achieve full coverage and capacity. Note that nonexclusive ownership of spectrum is likely to reduce the perceived value and will result in much renegotiation of license fees. Different countries will move at different speeds when implementing the required legislation. These business and legislative issues are likely to become the trickiest and most time-consuming problems in the implementation of cognitive radio.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Adolescent

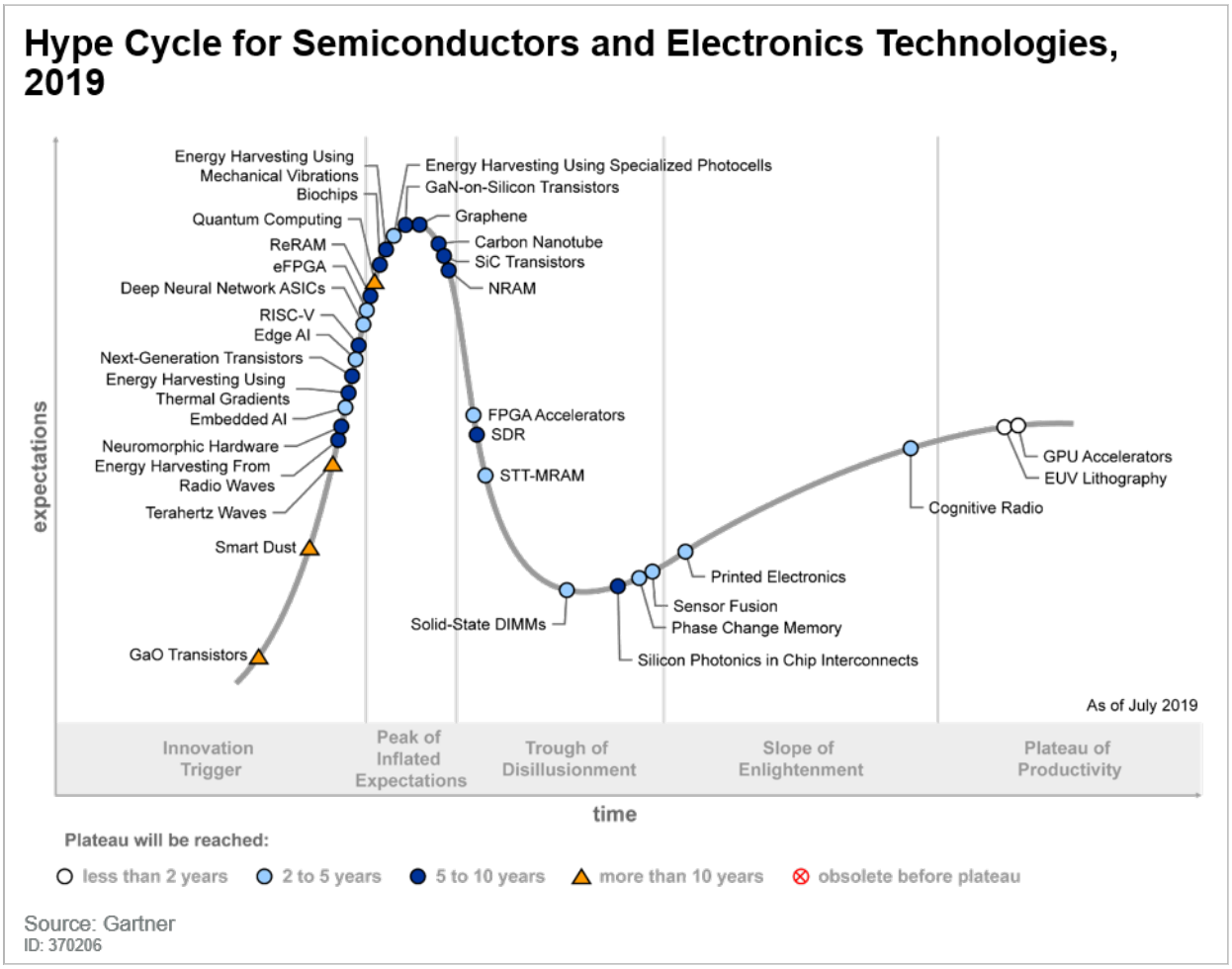
Sample Vendors: Carlson Wireless Technologies; KTS Wireless; Microsoft

Recommended Reading: "Market Trends: New Opportunities and Business Models Will Arise From the Growth of Dynamic and Shared Radio Spectrum"

"Innovation Opportunities Will Be Enabled as 5G Evolves Through 2025"

Appendixes

Figure 3. Hype Cycle for Semiconductors and Electronics Technologies, 2019



Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 1. Hype Cycle Phases

Phase	Definition
Innovation Trigger	A breakthrough public demonstration, product launch or other event generates significant press and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the technology is pushed to its limits. The only enterprises making money are conference organizers and magazine publishers.
Trough of Disillusionment	Because the technology does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the technology's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the technology are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the technology to reach the Plateau of Productivity.

Source: Gartner (July 2020)

Table 2. Benefit Ratings

Benefit Rating	Definition
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2020)

Table 3. Maturity Levels

Maturity Level	Status	Products/Vendors
Embryonic	■ In labs	■ None
Emerging	■ Commercialization by vendors ■ Pilots and deployments by industry leaders	■ First generation ■ High price ■ Much customization
Adolescent	■ Maturing technology capabilities and process understanding ■ Uptake beyond early adopters	■ Second generation ■ Less customization
Early mainstream	■ Proven technology ■ Vendors, technology and adoption rapidly evolving	■ Third generation ■ More out-of-box methodologies
Mature mainstream	■ Robust technology ■ Not much evolution in vendors or technology	■ Several dominant vendors
Legacy	■ Not appropriate for new developments ■ Cost of migration constrains replacement	■ Maintenance revenue focus
Obsolete	■ Rarely used	■ Used/resale market only

Source: Gartner (July 2020)

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

Understanding Gartner's Hype Cycles

Predicts 2020: Semiconductor Technology in 2030

Cool Vendors in Autonomous Vehicle Systems

Market Trends: Top 5 Semiconductor Industry Strategic Technology Trends for 2020

Cool Vendors in AI Semiconductors

Cool Vendors in IoT Thingification

Cool Vendors in Networking Semiconductors

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp