APRIL 2020

EE Times EUROPE

SPECIAL REPORT:
Augmented/Virtual/Mixed Reality

PAGE 8

ALSO IN THIS ISSUE

USB Type-C™.
24 reasons to connect with us.

WÜRTH ELEKTRONIK

GND TX1+ TX1- +V CC1 D+ D- SBU1 +V RX2- RX2+ GND

GND RX1+ RX1- +V SBU2 D- D+ CC2 +V TX2- TX2+ GND

#USBTYPEC

WE speed up the future

PCIM Europe Hall 7 Booth 229

USB Type-C™

Würth Elektronik offers the complete product range from connectors, components for filtering of EMI noise, ESD surge protection as well as AC/DC to DC/DC power conversion for USB 3.1. The specially designed products are featured in the Würth Elektronik USB Type-C™ EMC Dongle to easily and quickly validate the EMC and Surge compliance of your application. Our products are referenced in the USB Type-C™ reference designs of leading IC manufacturers.

For further information, please visit: **www.we-online.com/typec**
App Note ANP007 – Effective USB 3.1 filtering and protection: **www.we-online.com/anp007**

- USB licensed and TID listed Type-C™ Connectors WR-COM
- 60 W and 100 W Vbus filter
- High efficiency molded power inductors WE-MAPI for Vbus filter
- Low parasitic capacitance WE-TVS diode
- High speed, low noise data transmission filter with WE-CNSW HF
- Pulse stable SMT Chip Bead Ferrites WE-MPSB for hot-pluggin

USB Type-C™ Connector

WR-COM

Shielded Power Inductor

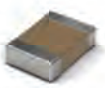
WE-MAPI

TVS Diode

WE-TVS

Common Mode Choke

WE-CNSW HF

Ceramic Capacitor

WCAP-CSGP

Ferrite Bead

WE-MPSB

EDITOR'S LETTER

The Show Goes On

Amid the panic wrought by COVID-19, it's easy to forget that the pandemic is only one of numerous challenges that enterprises and their employees must deal with.

The immediate concern, of course, is public health. Schools, museums, restaurants, and many other commercial and public services have been shuttered. China is slowly trying to return to normalcy — although at press time, Hubei province, the epicenter of the infection, remained in lockdown mode and the European Union had imposed a 30-day restriction on people's movements and activities.

But businesses are still in operation. Semiconductor manufacturers don't typically shut down fabs unless forced to do so by natural disasters or other overwhelming circumstances, because the cost of even a temporary shutdown is extremely high. Many of the countries and regions that have restrictions on movements have made it clear that they are not only allowing but also encouraging the unhindered flow of trade traffic. Shipments of essential goods have not been disrupted, and employees in health care and other essential sectors are on the job.

So it's not business as usual, but companies are still open for business. At AspenCore, we are still receiving press releases from electronics manufacturers, including companies that have asked many employees to work from home offices. At no other time has the internet been so critical to business activities. In fact, service providers are reportedly receiving requests for faster transmission-access lines to accommodate the new telecommuters.

As always, companies are shooting for resource optimization, demand visibility, supply velocity, financial efficiency, and effective communications. COVID-19 is just another challenge that companies must manage. Despite its disruptive effects, all the companies in the high-tech and electronics sectors still want to engage effectively with suppliers, customers, employees, shareholders, and other investors. In other words, they still want to get their stories out to the right audiences.

That's where AspenCore come in. We help tell your stories. We've got the audience you want to reach globally. We've got numerous channels: digital (e-books, websites, webinars, downloadable documents, and email campaigns, including newsletters and customized mailings), print magazines, and in-person exhibition events. And now, we're adding virtual exhibitions.

AspenCore will host its first virtual exhibition, the Power-Up Expo, in June. The event will offer companies in the power semiconductor and power systems sectors opportunities to showcase their designs, innovations, products, and technical capabilities. There will be networking opportunities online and chat rooms for engineers to engage with chipmakers and systems developers.

Things may look dire, but your business is still creating value, and that story must be told. EE Times editors are ready to help you tell it.

For more on the Power-Up Expo, visit *www.powerup-expo.com*. ■

— Bolaji Ojo, *global publisher and editor-in-chief at AspenCore Media*

VOL 3 | NO 3 | April 2020

EE Times EUROPE

SPECIAL REPORT: AUGMENTED/VIRTUAL/MIXED REALITY

Purveyors of VR and AR systems don't need to hype the technology anymore; they're hard at work making it practical. This Special Report investigates the technological status of VR/AR and the types of uses they're being applied to.

COVER IMAGE: ADOBE STOCK

boards & solutions + ECE

TECHNOLOGY TRENDS

Yole's CEO Identifies Five Investment-Worthy Trends

By Anne-Françoise Pelé

Jean-Christophe Eloy, Yole CEO

Behind every strategic decision lies a wealth of market research and analysis. Since its creation in 1998, market research firm Yole Développement (Lyon, France) has been committed to looking beyond the obvious to provide a reliable barometer of global electronics industry trends. Acquisitions, partnerships, product launches, economic downturns or upturns, even specific customer requests — all raise questions and open new fields of investigation.

In an interview with EE Times Europe, Yole CEO Jean-Christophe Eloy outlined five technology trends worthy of investigation and investment in the months and years to come.

HIGH-PERFORMANCE COMPUTING

Quantum computing could open possibilities we never even considered in chemical research, financial services, health care, life sciences, manufacturing, and defense.

Quantum computing is one of the key trends likely to have a long-term impact on the industry, and Yole's customers are eagerly asking for guidance. What is the expected silicon content in a quantum computer? How will the new paradigm affect business in the next five to 10 years? Will it change the structure of the industry?

"In quantum computers, there is a lot of coding, cryogenics is huge, and silicon is nothing," said Eloy. "The value is not in the silicon but in the system generating the cryogenics."

Many questions also surround neuromorphic imaging and computing, "which change the way we get and compute information," he said. Neuromorphic approaches are not only game-changers technologically but could also change the value structure of the market. Yole is "trying to understand how the value flow is changing the supply chain and who will be able to do something in that area," said Eloy.

Knowles Electronics, for instance, identified a turning point in edge audio processing with the emergence of artificial intelligence and took actions to make its products more powerful and suitable for AI. With the Google Pixel 4 smartphone, the Illinois-based company saw an opportunity not only to place microphones but also to add a processing chip, increasing its value in Google phones by a factor of four or five.

"If you combine the microphone with the ability to process the data and do specific things around the data with added value for the mobile-phone or car maker, you move from the few-cent sensor business to the few-dollar business," said Eloy. Companies like Knowles then multiply the value of their selling and climb the value chain.

TWO TRACKS FOR AUTOMOTIVE DESIGN

Auto manufacturers are not born equal. Tesla, founded in 2003 by eclectic Elon Musk, calls for "disruptive ways of working." Because it has no business history, it can map out technical and supply chain strategies on a clean slate. As a result, Tesla "is able to jump on solutions that are years ahead compared with other companies," said Eloy. Tesla is agile; it moves and adapts quickly. In contrast, traditional automakers such as Toyota, Renault-Nissan, and BMW have years of investment in technologies and long-term relationships with suppliers, making it "almost impossible for them to restart from a white sheet."

INSIDE THE DATA CENTER

Driven by the adoption of cloud-based technologies, data centers are proliferating in size and number around the world. The total number of hyperscale data centers — housing data, enabling entertainment, connecting consumers everywhere, and powering and cooling infrastructures — now exceeds 500.

Given those trends, Yole has received repeated customer requests for more intelligence on data centers. But for now, "everybody is driving in the dark," said Eloy. "Data centers are giving us a headache."

Over the years, Yole's analysts have closely observed the data-center market, including ups and downs in investments, movements in the supply chain, and technology choices. But without access to the parts, analysts can't do the reverse engineering to understand what is really used. Such opacity is due to the direct relationships between the companies designing and ordering the data centers (e.g., Amazon, Google) and the companies manufacturing them. The absence of intermediaries in the chain makes it difficult for Yole to gather information and build its own analysis.

TOWARD CONTEXT AWARENESS

With context-aware technologies, personal devices such as smartphones and home assistants can now collect and process raw contextual data about their users' surroundings, preferences, and behaviors in situ.

Mobile context awareness enhances the quality of service by providing more relevant responses, but it also raises security and privacy concerns. A wealth of personal data and private information is accessible, and "nobody is really sure of the way these devices are managing the data," said Eloy.

"I am surprised to see, every quarter, problems of data leaks in all the big applications," he added. Sensitive data immediately becomes vulnerable to cybersecurity risks, but customers have had limited reactions. "This is not acceptable."

In 2018, Europe came up with the General Data Protection Regulation (GDPR), which is still considered the most comprehensive privacy law. On the other side of the Atlantic, Apple preaches privacy and has made it a powerful marketing asset.

Today, Eloy said, companies simply say, "'Sorry, we will do better next time,' [but] my feeling is that there will be an amplification of this movement in the next few years to make sure your credit card is not hacked just because you are ordering something from your mobile phone or your home assistant."

Data leakage over the cloud is also a major concern for corporations and governments. In response, said Eloy, "the movement to the edge is huge. It will never stop, which means there will be more and more closed systems to protect data as much as possible."

PROCESSING AT THE EDGE

The transition from the cloud to the edge is indeed accelerating. Despite always-on power consumption and data-processing constraints, edge computing is emerging, placing the calculation at the system level.

As Moore's Law fast approaches its physical limits, the industry is under pressure to develop a new type of architecture. The trend toward systems-on-chip is decreasing because SoCs "cost a lot of money," said Eloy.

Chip partitioning and chip stacking are ways to extend Moore's Law and yield power, performance, area, and cost benefits, he added. It is all about having more processing power without "moving the whole chip to the 5-nm node." ■

Anne-Françoise Pelé *is editor-in-chief of* eetimes.eu.

POWER ELECTRONICS NEWS

ASPENCORE

EE Times EUROPE

POWER UP EXPO
The Virtual Power Electronics
JUNE 16-19, 2020
www.powerup-expo.com

Event Cancelled?
We've Got the Solution.

ASPENCORE offers POWER-UP, the Online Virtual Expo for players in the power electronics industry to meet and explore business and design opportunities.

Showcase Your Company's Power Electronics Innovations & Products to a Global Audience. Engage Virtually With:

- Design Engineers
- Power and Power Train System Builders
- Software Developers
- Executives and other Decision-Makers
- Power Supply Chain Architects
- Component Suppliers & Distributors

For information on Virtual booths, keynote presentation opportunities, webinars, podcasts, technical papers, networking events and interviews with ASPENCORE editors go to

www.powerup-expo.com/engage

Principal Host
Maurizio Di Paolo Emilio
Physicist
EE Times Correspondent &
Power Electronics News Editor-in-Chief

www.powerup-expo.com

SPECIAL REPORT: AUGMENTED/VIRTUAL/MIXED REALITY

Reality on Display: VR, AR, and MR

By Brian Santo

Virtual reality (VR) had been on the radar for so long that the commercial VR now available already seems old hat. That says more about consumer psychology than it does about the technology. Perhaps also inevitable was that the market would get segmented — into VR, augmented reality (AR), mixed reality (MR), and a grab-bag category of XR.

The distinctions have become important, as they map well to sets of applications. VR has thus far proved most suitable for entertainment purposes, such as video games and virtual travel. AR, meanwhile, has proved useful for industrial and professional applications, such as providing interactive guidance in manufacturing; for reference in system repair; as an aid in medical training as well as in the practice of medicine; and in similar situations wherein a professional can benefit from access to all kinds of reference data. MR and XR, meanwhile, still tend to be invoked only in special cases.

The market for AR/VR gear is projected to grow to US$18.8 billion by the end of this year, according to Statista. That includes mostly headsets and glasses but also handhelds, wearables, and body cameras.

Research-And-Markets has a more expansive view of AR/VR. It takes into account enabling technologies such as 5G, artificial intelligence, edge computing, and robotics. The firm expects that advances in those areas will pave the way for AR/VR adoption in diverse markets, including media, gaming, telepresence, retail, medicine, and education.

The industrial sector may have more applications, but the big part of the market in terms of revenue will still be consumer products, such as Sony's Playstation Virtual Reality and Facebook's Oculus Rift and Quest. Microsoft and Intel have been investing heavily in AR/VR and have yet to commercialize their technologies.

Intel, for example, has been working on a set of technologies that capture live activity and then replay it in such a manner that the viewer can review the action from literally any angle. The demonstrations of True View (photo, below) are almost always of sporting events, and after several years, they are still astonishing. At CES, company executives said they might have the processing power necessary to commercialize the technology cost-effectively in two or three more microprocessor generations.

Virtual Reality (VR)	Full immersion — Everything in view is computer-generated (or virtual).
Augmented Reality (AR)	Partial immersion — Most of what is in view is the real world; virtual objects are overlaid or otherwise inserted. Virtual objects can even appear to interact with real objects.
Mixed Reality (MR)	Partial immersion — MR commonly describes when real objects are inserted into a VR environment. The term is also sometimes applied to special cases of AR.
XR	A generic term covering VR, AR, MR, and related technologies; a grab-bag category covering related technologies that don't cleanly fit the other descriptions, including head-up displays (HUDs) and some systems based on haptics.

In this Special Report, we've got an interview with one company that took a deep dive into haptic actuator technologies for VR and wearables, a story about how Lockheed Martin is using AR in its manufacturing operations, and an article that takes a look at how AR/VR is being adopted in the manufacturing industry. We're also pleased to include a contribution from Peter Hartwell of TDK InvenSense; Hartwell not only is a top-notch technologist, but he is prescient when imagining how technology might be implemented. ■

Brian Santo *is editor-in-chief of EE Times.*

Intel True View (Image: Intel)

SPECIAL REPORT: AUGMENTED/VIRTUAL/MIXED REALITY

Kemet's CTO Talks Virtual Reality

By Gina Roos

Kemet Corp. was acquired by Yageo Corp. in 2019, transforming the companies into a US$3 billion interconnect, passive, and electromechanical powerhouse in the electronics industry. The electronic components manufacturer, which celebrated its 100th anniversary last year, has been expanding its product portfolio over the past several years in order to grow and innovate for next-generation technologies.

Historically, Kemet has been known as a major capacitor manufacturer for high-reliability applications, such as automotive, medical, military and aerospace, industrial, and telecom. But in recent years, the emergence of technologies such as IoT, 5G, automotive electrification, wearables, and AR/VR has led the company to expand beyond passive components to play a bigger role in these growing markets and to continue on a growth path.

In order to enhance its market position, Kemet needed to develop new technologies and products. The company's move into new component segments started in 2013 with its alliance with NEC TOKIN. It acquired the company outright in 2017, gaining entry into a range of new products and solutions, including temperature sensors, current sensors, piezoelectric actuators, current transformers, inductors, filters, and supercapacitors.

One emerging technology that Kemet had its eye on was haptics, or tactile feedback, which improves the user's touch experience on all kinds of devices, from smartphones to headsets. With the addition of haptics and some software, OEM designers can create a different user experience while differentiating their products.

Traditionally, the two key technologies used for haptic actuators are eccentric rotating mass (ERM) motors and linear resonant actuators (LRAs), with more recent developments around piezoelectric actuators and drivers.

Thanks to the NEC TOKIN acquisition, Kemet already had entry into the piezoelectric market. In 2016, Kemet began a collaboration with Novasentis Inc., a developer of haptic and sensory feedback technology for wearable devices. Kemet acquired Novasentis in July 2019. The two companies are working to develop and commercialize electromechanical polymer (EMP)-based haptic actuators for wearable and VR devices.

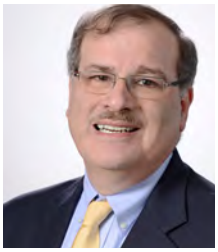
The collaboration leveraged Novasentis's haptic actuator film and core technology and Kemet's expertise in film-based capacitor manufacturing. The film-based haptic actuators offer an ultra-thin profile and small size for haptic feedback in VR and wearable applications, providing a variety of tactile sensations.

The thin, flexible actuators deliver piezoelectric effects at lower operating voltages compared with traditional piezoelectrics. They can also produce audible sound when they vibrate in response to audio signals, similar to speakers, simultaneously providing haptic and audio feedback. With the Novasentis LiveTouch haptic skin technology, designers can add touch to the surface of products, providing localized and natural authentic touch sensations, as well as programmable and customizable effects for a range of sensations.

In an interview with EE Times Europe sister publication Electronic Products, Philip Lessner, Ph.D., Kemet's senior vice president and CTO, said that he is starting to see real-world uses for VR in industrial settings, including Kemet's factory locations, where VR is helping to train workers on complex equipment. Excerpts appear below.

ELECTRONIC PRODUCTS: What are the basic building blocks, at the component level, for AR/VR devices?

Philip Lessner: I'll talk primarily about the two technologies that we have. One building block for these very advanced computers is coming up with the lightest power-conversion components. It really means capacitors, and in our case, it's polymer tantalum capacitors.



Philip Lessner

So it's the portability and wearability aspect of it — providing high-power computing in a very confined space.

The other is the sensory experience. We've had a long-term partnership with Novasentis to make flexible, lightweight, and miniaturized haptic devices that give you a sense of touch. And the first market penetration we're seeing for these devices — what we call polymer actuators — is in the VR/AR community for use in gloves and vests and in other applications in which they want to provide some type of touch or tactile feedback.

EP: What differentiates Novasentis's haptic technology from competitors?

Lessner: Most haptics are done via an electromagnetic coil or something similar, so they're fairly bulky and intrusive. We have a polymer film that when you apply a voltage signal to it, it vibrates and provides that haptics experience. How we're differentiated from traditional devices are the advantages of miniaturization and better integration into end products like a glove or vest.

The original work for the polymer film came out of Pennsylvania State University. Novasentis bought the patent, along with other add-on patents after that. We don't believe anybody else in the world can make a device quite like this.

EP: What makes polymer tantalum capacitors a good dielectric for these applications?

Lessner: We've done some designs into various headsets, and what makes it a good dielectric is the high capacitance and the retention of capacitance with bias. We're seeing a lot of use of the polymer tantalums on 3.3-V and 5-V rails where they need something either for filtering or for power holdup. There are a few applications and requests for 25-V and 35-V parts for the input power, similar to what you would see in high-end laptop computers. Apple, is, of course, a big user of polymer tantalum for the MacBooks. It's about getting the most capacitance into the smallest constrained space as possible.

EP: In terms of sensors — the second building block that you offer — do you think that sensors, in general, are the biggest area of innovation for VR applications?

Lessner: Yes, I think [they] will be. In general, you need a lot of sensors on these devices, and of course, a lot of them take advantage of the sensors that have been incorporated into smartphones and other devices, as part of continuing cost reductions. We believe the haptic ones, in particular, give a differentiated experience.

I've done a few demonstrations using the Oculus headset. With our polymer haptic sensors, you can actually simulate different textures and feelings. For example, if you're in a VR environment and you're running your hand across a brick wall or you pick up a ball, you can actually feel not only the sensation of touching, but you can actually simulate the different textures, so it makes it very realistic.

Kemet's CTO Talks Virtual Reality

We're involved with haptics, but the other sensations of sight and sound and what you can do with the vision systems and the accelerometers and other technologies are equally important.

EP: Does the capability to feel different textures open up new opportunities?

Lessner: I think it does. Without a sense of touch, you're sort of just poking into the air.

We actually have a pilot program to implement some VR systems for manufacturing training to train technicians on how to set up some of Kemet's complex machines. There's no sense of touch yet; you press buttons or turn dials on the machine, but you're doing this in thin air, so you're missing part of the feedback experience.

I think having this sense of touch and some rich experience around that will greatly enhance the user's enjoyment of these systems, and in the more industrial case that we're working on, a sense of reality and feedback will help them learn how to set up these complex machines.

On the consumer side, I think it would open up more opportunities in gaming. Virtual reality will primarily be around gaming, while augmented reality has other applications, which will open up more realistic experiences.

The first generation of haptics is beginning to be incorporated into the next generation of game systems and VR systems. Based on press releases from Sony about its PS5 console, which will have some haptic devices in the game controller, users will have a greater sense of touch. They use the example of pulling back a bow. As you pull back a bow, there is more and more force exerted, and with the sensors and haptic controllers in the PS5, the user will be able to feel that, for a more realistic experience.

EP: What are the most important technology improvements/specifications for sensors in these applications?

Lessner: One of them is miniaturization. From personal experience, when we started out with the VR project for training, we had goggles and four set camera sensors on tripods placed around the room to get the realistic experience. As we've gone through this project, which we're doing in conjunction with the School of Computing at Clemson University, the headsets have advanced to the point at which we no longer need these bulky external cameras and sensors. It's really all in the headset now.

Thr Novasentis haptic actuators can be applied to equipment, the body, and gear to create very fine haptic feedback.
(Image: Kemet)

We're using headsets that still need to be connected to a computer, but I think the next generation will cut the cord. That's where 5G comes in, or some type of wireless connection between the computer and the headset.

The Clemson students and professors developed VR simulations of what they had seen a Kemet master technician perform in real life, using a game engine and other software, to develop the VR training module. The VR gloves transform the hand actions to digital inputs, allowing trainees to manipulate virtual controls, press buttons, and adjust calibrations to operate the virtual equipment.

I really see the continued untethering of the headset from a computer system and the continued miniaturization of the sensors, so everything is in the headset and nothing needs to be placed in the external environment.

I think we're probably about one or two years away from having VR systems where you can walk the street immersed in virtual reality and have a pretty realistic experience.

EP: Is Kemet primarily targeting industrial-type applications?

Lessner: We're targeting both consumer and industrial applications — across the board but more toward high-end headsets. We're engaged with a set of customers for both VR and XR. We've either been designed into or are sampling into products such as Oculus, Microsoft HoloLens, and Magic Leap. We've been working with Magic Leap for a while, which has advanced augmented reality.

The initial driving force will be more of an industrial application. There's a lot of activity in that area, especially around the high-end headsets for training industrial workers and also from an AR point of view for doing machine maintenance.

Everybody laughed a few years back with the Google Glass fiasco, but it looks like the whole idea of smart glasses is starting to come back. I think what's happened is the electronics has caught up, and miniaturization has caught up.

That's where very small-case-size ceramic and tantalum polymer parts and other miniaturized components are targeted.

EP: Designers tend to forget about the passives and the interconnects, but these components also have to shrink, so does there have to be innovation in these areas, too?

Lessner: You're right. With our acquisition of NEC TOKIN in 2017, we got the capability to do those small case sizes in polymer tantalum, and we've taken advantage of it in applications like these.

EP: Do you see any emerging applications for VR — outside of training or machine maintenance or virtual gaming?

Lessner: One is virtual travel. I was reading about some high-end VR system that was released the other day. The article on CNET said it felt like you were being transported to someplace else. They'll be increasingly used in those types of applications where people want an experience — sort of experiential instead of game-like applications.

EP: What do you think are some of the biggest technology challenges today associated with XR?

Lessner: Miniaturization and power consumption. In general, the amount of energy and power that's required to make these high-end computer and sensor-laden systems work. Power consumption has to come down. The other is compute power and really miniaturizing the electronics, such that they fit into the size of a smartphone.

And adding more sensory experiences: We're working on haptics, but there are other experiences with other simulated senses. As time goes on, we'll see more and more multi-sensory experiences to go with today's 3D vision-type experience.

We're in a unique position with a very thin and flexible sensor that is easily integrated into gloves or other wearable devices, but I think there will be other types of sensors that will be needed in the future. ■

Gina Roos *is editor-in-chief of Electronic Products.*

SPECIAL REPORT: AUGMENTED/VIRTUAL/MIXED REALITY

AR Is Propelling Space Manufacturing

By George Leopold

To hear the purveyors and early adopters of augmented reality platforms tell it, the technology is emerging as the ultimate "measure twice, cut once" reference check for an increasing number of mission-critical aerospace and military systems.

Last fall, we reported on Lockheed Martin Corp.'s embrace of AR technology as it builds NASA's Orion spacecraft. The world's largest defense contractor is working with large and small AR developers to speed the presentation of workflow data spanning assembly, test, and maintenance.

Among the aerospace giant's key AR suppliers are Microsoft, with its "mixed reality" HoloLens platform, and Scope AR, a software vendor specializing in AR training tools. Those companies are at the leading edge of efforts to push AR and MR tools into the mainstream.

Augmented reality tools are allowing companies like Lockheed Martin to complete production activities in half the time.

In November, Scope AR announced support for Microsoft HoloLens 2 on its flagship AR platform, called WorkLink.

That combination is aimed directly at large enterprises like Lockheed Martin that can afford to invest in AR tools for huge projects such as Orion while getting a return on investment in the form of manufacturing efficiencies. The contractor is using AR tools to assemble various Orion components, including the skeletal framework of the spacecraft's titanium heat shield.

Among the goals in adopting AR technology for aerospace projects with the strictest of tolerances is what Scope AR's Scott Montgomerie calls "real-time knowledge transfer."

On the factory floor, the AR specialist is working with Lockheed Martin to develop training manuals that include animations for assembling spacecraft components. Training time for Orion technicians was reduced by 85%, according to Shelley Peterson, Lockheed Martin's augmented technology project leader.

The resulting workflows have been used to reduce touch labor for Orion spacecraft components, including fasteners and accelerometers with narrow tolerances.

The company estimates that AR technology has reduced touch labor for drilling by 45% and for torque applications by 50%. The task of tightening Orion's fasteners was reduced from six weeks to two weeks using AR.

Peterson said that AR and MR tools have also streamlined the translation and presentation of workflow data ranging from assembly and manufacturing to test and maintenance steps. AR software can also be used to add part identifiers or color coding of components. Assembly steps can then be animated.

In one example, Peterson said that technicians don virtual reality goggles that display the precise spot where a spacecraft component should be attached. For volume parts like connectors, technicians previously had to measure manually before installing individual components.

Lockheed Martin, which pioneered the use of disruptive technologies like quantum computing to help debug mission-critical code, is among a growing list of manufacturers embracing AR technologies. Others include Boeing, Siemens, and Toyota.

Indeed, market watchers say that manufacturers are embracing augmented reality for a range of applications. In a recent AR adoption survey, business consultancy PwC reported that product design and development are the most popular AR applications, followed closely by safety and skills training along with maintenance and equipment operation.

"What we're seeing, then, is [virtual reality and] AR as an advanced manufacturing technology tool — just like robotics, 3D printing, and the internet of things," the PwC survey concluded.

The bottom line for aerospace companies applying AR, said Peterson, is a 50% reduction in the time needed to complete manufacturing tasks.

"Augmented reality is allowing us to complete space manufacturing activities in half the time," she said. "It's becoming a reality today." ■

George Leopold *is a contributing writer to EE Times Europe.*

SPECIAL REPORT: AUGMENTED/VIRTUAL/MIXED REALITY

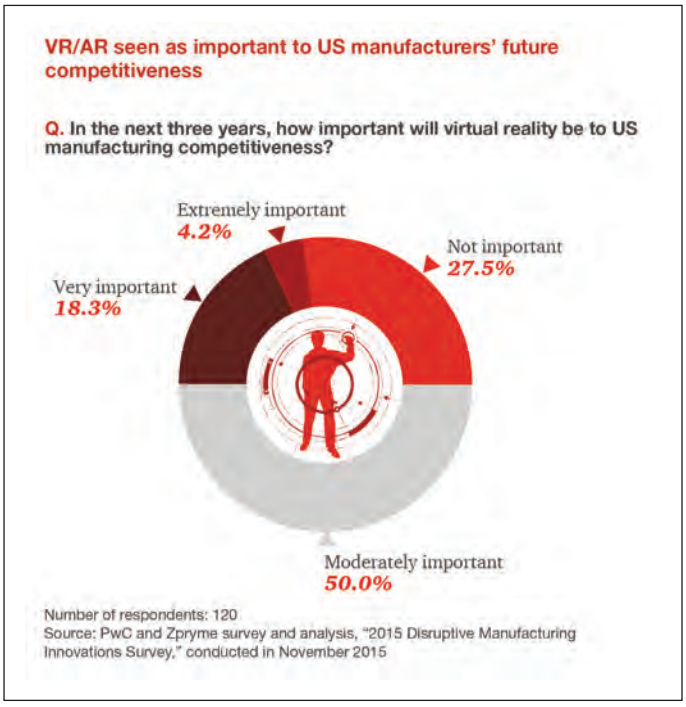
Flex Visualizes AR/VR in Manufacturing's Future

By Barbara Jorgensen

As AR/VR technologies come of age, manufacturers are finding ways to adopt these technologies on a variety of fronts, including product development, training, maintenance, repair, and worker safety. Flex, a US$25 billion electronics design and manufacturing services (EMS) provider, envisions expanding that universe even further.

AR/VR is so important to the manufacturing industry that Flex has adopted it as one of six key pillars of its Industry 4.0 strategy. The company's focus includes enhancing its global manufacturing processes and delivering customer solutions through M2M communications, smart automation and robotics, augmented/virtual reality, 3D manufacturing, simulation and visualization, and business intelligence.

"[AR/VR] is something that will not only impact Flex but the industry in general," said Zohair Mehkri, XR and simulation engineering manager at Flex. "It has a place in design, usability, and product life-cycle management. The use case depends on where the product is in its life cycle. AR and VR are usually lumped together, but they are inherently different in the development of the products."



EMS companies have a unique position in the electronics supply chain; Flex's 100 facilities worldwide are outsourced to original equipment and original design manufacturers (OEMs/ODMs). Manufacturing services alone yield razor-thin profit margins, so EMS providers add value through design assistance, process improvement, and after-market services.

If AR/VR technologies catch on as predicted, combined sales are forecast to hit US$150 billion by 2020, according to a Manatt Digital Media estimate — with AR alone accounting for about US$120 billion.

CONNECTIVITY IS KEY

Connectivity is driving the manufacturing industry's interest in AR/VR. "You hear all the buzzwords [such as] IoT and M2M, and one way to think about this is connecting people — the way they move, the way they communicate — to each other and with machines," Mehkri said. Machines are digitized and people are analog, so AR/VR adds a layer of human-device connectivity.

"We see uses such as remote assistance, or you can collaborate with people anywhere in the world on a design and draw something in real time," he said. Users can send out information and share what they're doing as they're doing it.

Most enterprises are adopting AR for their internal use and benefit, marking improvements in operational efficiency or lower costs as main value drivers, according to research by PTC, a digital solutions provider. To achieve these benefits, use cases are being developed that leverage AR to provide instruction or guidance. The interact capability, or using AR to manipulate digital graphics or interface with a smart, connected product, is an emerging capability that will grow as AR becomes integrated with more business.

The way in which manufacturing information has been documented, maintained, and shared has historically been slow and costly. Flex sees applications for institutional knowledge.

"This kind of experience is usually shared on-site or through people talking with one another," Mehkri said. "If it is archived, it's a bunch of work instructions followed by a process-definer that outlines the product." That doesn't capture the art of making the product.

"The way a designer places their finger or bends a wire — that's a level of experience that isn't always captured," he said. "AR could be used to show the best way to make something and having that information on record to be studied and archived. You have a library of best practices."

WOWING THE CUSTOMER

Opportunities to leverage AR/VR capabilities across the value chain are diverse, with the heaviest concentrations in design, manufacturing, service, and training.

Design use cases typically drive the design-for-excellence (DFX) value propositions, wherein designers visualizing products at scale can engineer based on manufacturing and service efficiencies, sustainability, cost, or ergonomics. Productivity is gained by a reduction in design iterations and by working those downstream efficiencies into the design requirements more efficiently.

Even the most experienced designer doesn't always account for how a product is physically built. "We could use AR to collaborate with designers and provide remote assistance," Mehkri said. "Now, a customer-support check could mean flying halfway around the world. Telecommunications and video are solving some of those problems, but content is lacking."

Augmented reality can display annotations on a project or visually demonstrate how to complete a task. "With AR, you can draw on a field of vision, insert bubbles, or have manuals pop up for the application," Mehkri said. "You get closer to your expertise, and that can be used for both customers and suppliers."

Virtual reality can show customers how their end product will look. Drawings or PowerPoint presentations are two-dimensional, whereas VR immerses users in an environment.

"Customers ask us how a production line is laid out, and with a VR headset, we can 'walk' around the line, or [you can] 'poke' your head into each machine," said Mehkri. "You can also show customers what works or doesn't work in their design or where [the designers might] need to make a design change."

During product design, Mehkri added, customers may unintentionally

(Image: PTC)

skip a step that's crucial to manufacturing. "We can show customers that process in the 'real world' and capture that for future use cases."

For manufacturers, use cases are most often utilizing step-by-step instructions to speed the completion of tasks and improve work quality and collaboration among workers. As consumer expectations rise, these types of initiatives not only yield internal workforce benefits but ultimately improve relationships with customers and supply chain partners, providing competitive advantages. Industrial enterprises are focusing on harnessing AR capabilities across their value chains.

BUILDING BETTER PRODUCTS

Flex foresees AR/VR uses for both employee training and product development. Onboarding a worker in a factory setting is typically loud and distracting. Rather than train workers on the factory floor, VR can immerse them in a virtual environment. For industrial enterprises offering AR experiences to enhance their workforce, 69% of use cases are focused on benefiting their internal workers in engineering, manufacturing, service, and training, according to PTC.

For EMS companies and their customers, relationships largely revolve around hardware. Here, Flex envisions breakthroughs in the design and production of the headsets required for AR/VR experiences. Current headsets are bulky, awkward, and slow, and they provide a limited perspective, yet some employees may be required to use them at work.

"Your workforce are also consumers," Mehkri noted. "So we ask ourselves, would they go out and purchase this hardware and wear it all day? The answer is no. AR/VR hardware in general requires a lot of improvement for it to be made, accepted, and used."

This is technology that Flex has been looking at for years, he added. "We see it becoming more prevalent in industry. Penetrating consumer lifestyle with breakthroughs from the manufacturing sector is really what drives the industry.

"This is where a lot of things come together to drive technology forward." ■

Barbara Jorgensen *is managing editor of EPSNews.*

SPECIAL REPORT: AUGMENTED/VIRTUAL/MIXED REALITY

VR Could Be the Last Nail in Privacy's Coffin. Unless…

By Peter Hartwell

Emergence is one of the most interesting concepts in the study of complex systems. Emergence occurs when an entity is observed to have properties that its parts do not have on their own. Experts in cyber-physical systems study these unintended consequences, often trying to predict failure points in fields as diverse as global electronic banking and air traffic control. But emergence can also describe cultural shifts caused by widespread adoption of a new technology.

One of the unexpected consequences I have come across in my creation of 360° video content for consumption in a VR headset is that nothing escapes the camera lens. I learned this lesson early on as I was viewing a clip for the 10th time, at yet another angle, and noticed someone (who thought no one was looking) trying to remove something from between their teeth. It's such a common thing, but something we reserve for a private moment. And yet there it was, preserved forever, and it got me thinking.

Peter Hartwell

The 360° camera is the last step in the loss of our privacy, but it makes the deep fake harder to pull off. That could make us better people.

The 360° camera is the last step in the loss of our privacy. From the first time we went online to the first time we carried a device that pinged a fine-grained network of towers for connectivity, we have slowly given ourselves away for the convenience of being in touch and accessing information. So now, with the possibility of these cameras anywhere and everywhere in public, we will always be on stage, with the chance to be viewed by others.

Spooky? Perhaps.

Yet emergence is about unintended consequences. For example, the cellphone network has the "correct time," so it's easy to keep phone clocks synchronized to the world's definition of now. If you are late to a meeting, you can no longer blame the clock for being slow or having stopped. Being late is now, at its simplest, a demonstration that something else was more important. However, I do believe people today are more on time than they were in the past.

Now, imagine that we have extended our cars' dashcams to 360°. It's no longer a matter of catching the right angle for a traffic collision; the dashcam will capture all angles — your camera and mine. If I was distracted and caused that minor fender bender, I won't be able to jump out of the car and start ranting about you this and you that. Rather, because it's captured as it happened, I will have no choice but to own up to what I have done. There is no discussion, no he said/she said, just the data.

And maybe this will make us better people.

The 360° camera also has the potential to expand journalism. A scoop will be the person with a 360° camera who is at the right place at the right time, so we can all share the experience. No more dispute about how many people were at the U.S. presidential inauguration; just pop on your glasses, stand on the National Mall, and experience as little or as much of it as you like for yourself.

Today, we are at the technology maturity point at which we can use computers to alter or create realistic video images to entertain us with exotic stories. The potential also exists to create new or different versions of the news — the deep fake. But a 360° video, with the studio, production crew, or other reporters who are all clearly visible, is much harder to recreate. Therefore, 360° cameras add a much higher barrier to the integrity of the digital media, one that may be too hard to overcome, as humans are actually really good at being bothered when something doesn't look quite right.

So maybe the emergence of capturing our full experiences for VR is the beginning of the end of the loss of our privacy.

InvenSense's MEMS solution, ICM-42605 (Image: InvenSense)

Or maybe — just maybe — it's the dawning of a new era, one in which we take responsibility for our actions, knowing there's the potential for anyone and everyone to see what we're doing. If that happens, we may emerge as kinder, gentler neighbors, treating others as we would have them treat us. ■

Peter Hartwell *is CTO of InvenSense Inc., a TDK group company.*

IMAGES: UNSPLASH (TOP RIGHT); INVENSENSE

AUTOMOTIVE

EC to Bet on Hydrogen Fuel-Cell Vehicles

By Maurizio Di Paolo Emilio

In its push to become climate-neutral by 2050, the European Commission has unveiled an industrial strategy called the Clean Hydrogen Alliance. The public-private partnership among the EC, Europe's fuel-cell and hydrogen industry, and research organizations will drive the region's efforts to ensure Europe's energy independence and develop zero-emission cars.

FROM ELECTRIC VEHICLES TO HYDROGEN-POWERED CARS?

Next-generation vehicles that produce less harmful substances and less noise are universal goals, and as a result, the whole world is experimenting with electric vehicles (EVs). But a drawback of EVs, which belong to the electro-mobility sector, is their long charging times.

That's a problem that might be solved by hydrogen fueling, also called cell fueling. Hydrogen cells exploit the same energy used by spacecraft to reach Earth's orbit. But how do hydrogen cars work? What are the strengths and weaknesses?

An environmental argument for hydrogen-powered vehicles is that they do not emit greenhouse gases. But hydrogen is not exactly a renewable resource; it must be produced.

THE HARDWARE OF A HYDROGEN VEHICLE

Hydrogen vehicles convert chemical energy into mechanics. A hydrogen internal combustion engine vehicle (HICEV) uses a hydrogen-fueled version of the traditional internal combustion engine. Alternatively, hydrogen can be used to cause a reaction with oxygen in a fuel cell, thus producing electricity. This type of vehicle is called a fuel-cell electric vehicle (FCEV), and in recent years, FCEVs have drawn the attention of many manufacturers. FCEVs are powered by an electric engine and have an on-board power plant to allow the production and management of hydrogen.

An environmental argument for hydrogen-powered vehicles is that they do not generate polluting emissions; rather than greenhouse gases and fine particles, they emit water vapor (**Figures 1** and **2**). However, the overall environmental impact of hydrogen mobility depends on the energy source used to produce it. Hydrogen is the most abundant chemical element in the universe, but not in natural form. For this reason, it is not exactly a renewable resource and must be produced. If you use renewable energy sources to produce hydrogen, the environmental impact is minimal. Conversely, if fossil sources are used, the environmental impact is much higher.

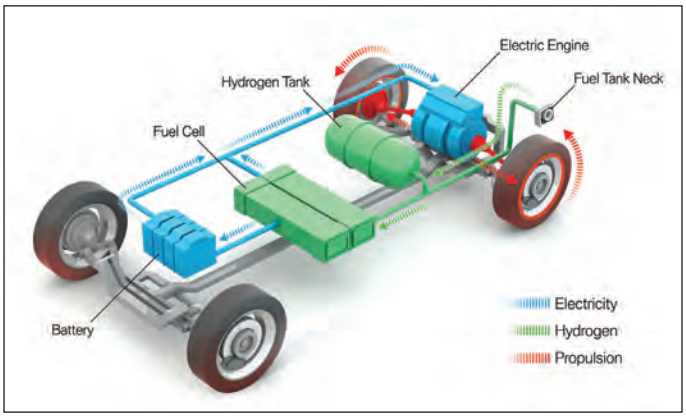


Figure 1: *Block diagram of a hydrogen car. In the fuel cell, hydrogen and oxygen generate electricity, which is conveyed to the electric motor and/or battery.* (Image: BMW)

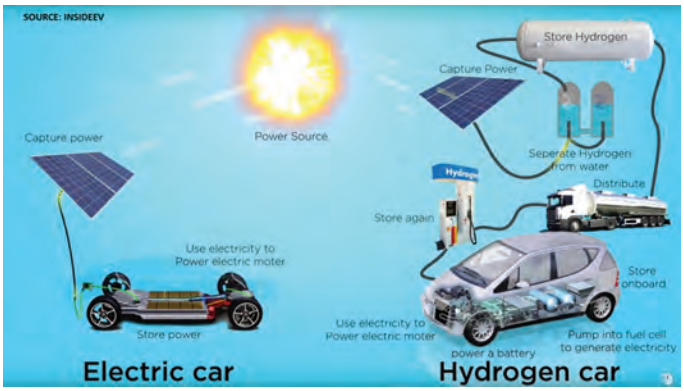


Figure 2: *Electric and hydrogen vehicle in comparison* (Image: InsideEVs.com)

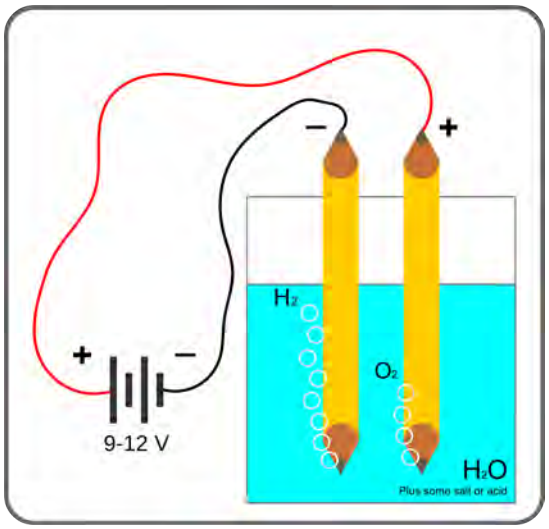


Figure 3: *General block diagram of electrolysis*

Hydrogen can be produced through two production processes and technologies: reforming and electrolysis. Reforming has a higher environmental impact because it involves the extraction of crude oil, transportation, and refining. Electrolysis is the process of splitting H_2O water molecules into individual hydrogen and oxygen atoms via the chemical reaction induced by electricity. This production process does not emit polluting gases but requires a large amount of power. The electrolysis process consists of a low-voltage current that flows through the water to release oxygen and hydrogen in a gaseous form.

One of the qualities of hydrogen is its very high specific energy density, 40,000 Wh/kg, or 236× the specific energy of lithium-ion batteries. This means that hydrogen-powered vehicles are lighter than battery-powered vehicles and have a more extended range. Also, hydrogen refueling takes just a few minutes, compared with several hours for battery-powered cars.

Figure 1 shows a general layout of a hydrogen vehicle. Reverse electrolysis takes place in the fuel cell: Hydrogen comes from a tank, air from the surrounding environment. Fuel cells do not provide thermal combustion but an electric current that produces non-altered water as the only waste.

Fuel cells receive two incoming flows: hydrogen from the negative pole and oxygen from the positive pole. The catalyst contained in the hydrogen engine causes the electrons to separate from the nucleus, and this reaction releases electricity. The electrons move to the positive pole and join the oxygen atoms, which receive a negative charge.

The union of hydrogen with oxygen produces a chemical reaction whose final product is water. What is emitted is water vapor, which can be released directly into the atmosphere by hydrogen-fueled cars (**Figures 3** and **4**).

The current generated in the fuel cell can directly power the vehicle or charge a battery (smaller than a conventional battery) that acts as an intermediate accumulator. As in other hybrid vehicles, hydrogen-powered vehicles also use energy harvesting to recharge the battery. The battery is used to cover possible engine energy demands and to recover braking energy, as in electric and hybrid cars. The voltage produced by fuel cells must be up to 600 V.

At low speeds, the car moves and obtains energy only from the battery. At higher speeds, such as on a highway, the fuel cell brings extra power to the engine while recharging the battery at the same time.

ADVANTAGES AND DISADVANTAGES

The advantages that hydrogen fueling can bring are maximum reduction of emissions (water is the only waste product), speed of refueling, reduction of consumption, and prolonged autonomy. But there are also disadvantages related to the problematic management of hydrogen, from electrolysis techniques that require electricity to storage in filling/charging stations. The losses caused by electrolysis pose problems for hydrogen production.

On the downside, you cannot refuel/recharge anywhere; in fact, many of the problems are due to the difficult storage of hydrogen, which, as a result, can count very few distributors today. And, as mentioned above, the process of separating hydrogen can be energy-intensive.

The storage of hydrogen requires a lot of space compared with that needed for petrol or diesel. The hydrogen tank must be strong enough to withstand pressure of about 700 bars. Hydrogen is stored in gaseous or liquid form at –253°C.

The refueling infrastructure, mostly non-existent today, would require significant economic investment. Furthermore, the production costs of a fuel-cell vehicle are considerably higher than the costs for manufacturing a conventional car.

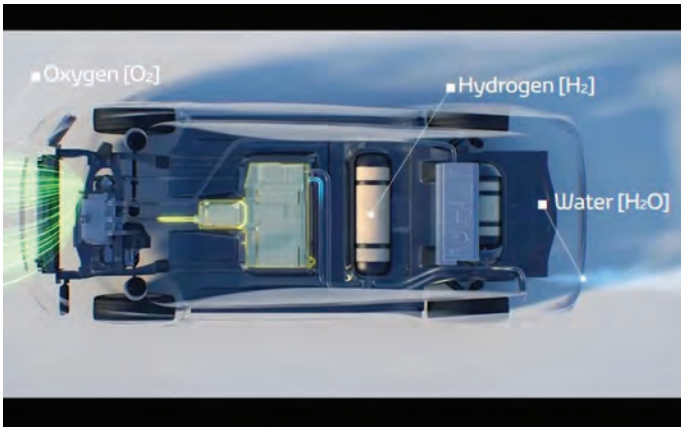


Figure 4: *Internal structure of a hydrogen car*

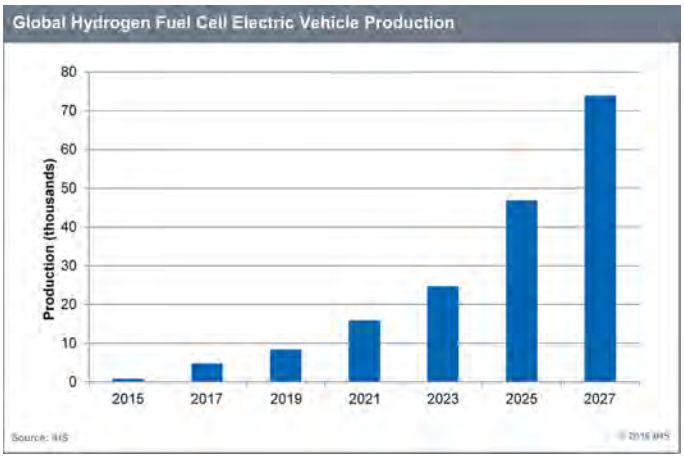


Figure 5: *Global hydrogen fuel-cell electric vehicle production* (Image: HIS)

Hydrogen cars have greater driving range than battery-powered vehicles. A tank full of hydrogen is sufficient to cover a distance of about 500 km. Cars powered by a battery reach this value only if they are equipped with huge batteries, but this will result in higher vehicle weight and longer charging times.

The refueling infrastructure, mostly non-existent today, requires significant investment. Germany has taken a lead; by December 2019, it had built about 80 stations.

MARKET EXAMPLES

Germany is the leading country in Europe that is focusing on infrastructure for hydrogen-powered vehicles. In December 2019, in fact, there were about 80 refueling stations for hydrogen cars in Germany.

Numerous manufacturers, such as Ford and Renault, have ventured into making hydrogen-powered cars but have preferred to direct their resources toward electric vehicles. In contrast, others, such as Toyota, have continued to develop the technology and have seen their hydrogen vehicle designs all the way through to production and marketing.

The Toyota Mirai uses the Toyota Fuel Cell System (TFCS) and offers both fuel-cell technology and the one that is typical of hybrid electric cars. Another hydrogen car is the Honda Clarity Fuel Cell. The Japanese sedan can cover a distance of 650 km and a maximum speed of 165 km/h, thanks to the 177 horsepower of the innovative engine. Audi H Tron, BMW i Hydrogen Next, and Hyundai NEXO are of particular interest among hydrogen cars.

Hydrogen cars consume an average of 1 kg of fuel per 100 km. The average price of hydrogen is €10 per kg: It can be assumed that a full tank can cost about €50, while the average driving range on a full tank is comparable to that of a petrol car (**Figure 5**).

Hydrogen makes an essential contribution to the sustainable mobility of the future, alongside cars powered by a battery. FCEVs can represent zero-emission technology and allow users to maintain their flexible driving habits. Hydrogen can be up to twice as efficient as gasoline. But we must improve hydrogen production through renewable solutions. It is essential to reduce infrastructure costs for transportation and distribution. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

AUTOMOTIVE

Maxim LiDAR Module Revs Self-Driving Safe Speeds

By Maurizio Di Paolo Emilio

The advent of self-driving has decisively expanded the presence of laser-imaging detection and ranging (LiDAR) sensors in the automotive-electronics platform. LiDAR works according to the radar principle but uses light pulses emitted by an infrared laser diode.

Maxim Integrated's new MAX40026 high-speed comparator and MAX40660/MAX40661 high-bandwidth transimpedance amplifiers (TIAs) enable 15 km/h faster autonomous driving at highway speed by doubling the bandwidth and adding 32 channels (for a total of 128 instead of 96) in a LiDAR module of the same size.

WHAT'S LIDAR?

Along with artificial intelligence, cameras, and radar, sensors are indispensable to assisted and autonomous driving. Because they can provide accurate measurements of objects and detect obstacles on the road — fallen tree limbs, other cars, or even a child who darts out into traffic — LiDAR sensors have helped advance the adoption of advanced driver-assistance systems (ADAS) and are critical to autonomous-vehicle (AV) development. An AV's perception of the surrounding environment must be extremely precise, which is why experimental robo-cars are full of sensors. The use of a laser lighting system allows self-driven cars to be operated under low- or no-visibility conditions and even in the absence of road markings.

"LiDAR sensors are playing an increasing role in the fusion of vehicle sensors for their ability to provide accurate distance measurement of objects," said Maurizio Gavardoni, principal member of the technical staff at Maxim Integrated. "A typical LiDAR sensor sends light pulses that, reflected by objects and detected adequately by photodiodes, allow you to map the surrounding environment."

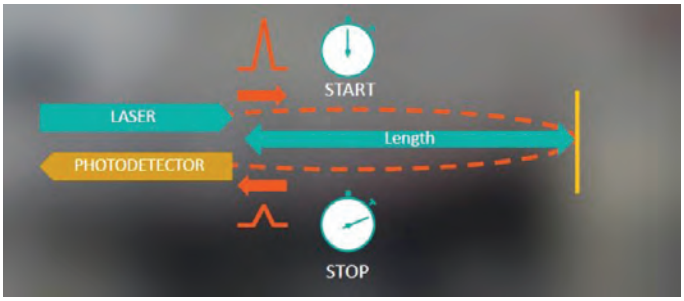


Figure 1: *Time-of-flight functional diagram* (Image: Maxim Integrated)

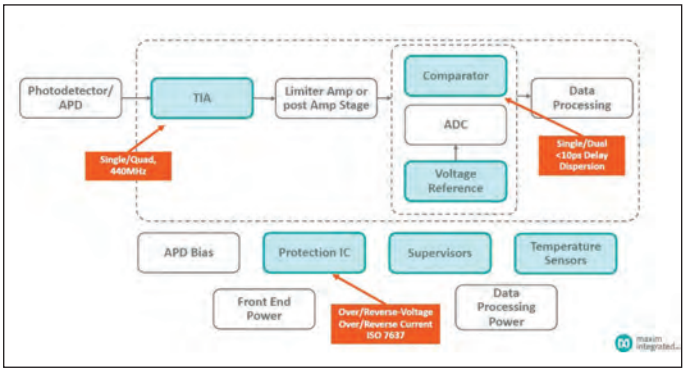


Figure 2: *General layout of a LiDAR sensor with the main electronic parts shown* (Image: Maxim Integrated)

Maxim Integrated's Maurizio Gavardoni demonstrates the evaluation board for a four-channel LiDAR receiving system. It includes optical photodiodes from First Sensor and Maxim's newly launched TIA and high-speed comparator. (Image: Maxim Integrated)

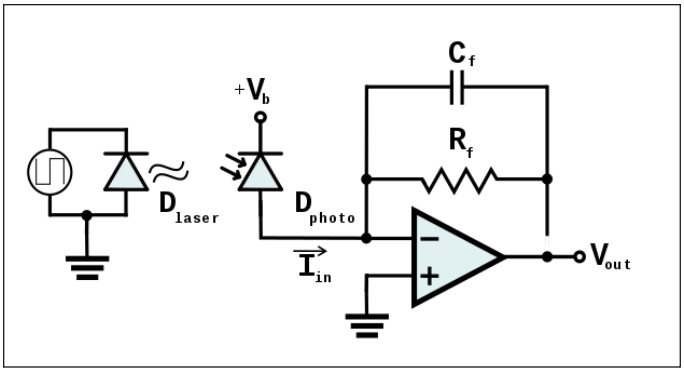


Figure 3: *General layout of a TIA with a reverse polarization photodiode* (Image: Wikipedia)

Figure 1 shows the basic layout of a LiDAR sensor. There are two basic types of LiDAR systems: micropulse LiDAR and high-energy. Micropulse systems have been developed as a result of the ever-increasing computing power available and advances in laser technology. These new systems use very low power, in the order of 1 W, and are entirely safe for most applications. High-energy LiDAR, on the other hand, is common in atmospheric monitoring systems, where the sensors are used to detect atmospheric parameters such as height, stratification, and cloud density.

"Automotive self-driving systems are evolving from 35 mph to 65 mph and beyond, but faster autonomous self-driving systems are essential," said Gavardoni. "The challenges in meeting these demands [translate] into high-precision distance measurements of objects, [requiring] more accuracy, more channels to fit in space-constrained platforms, [and compliance with] stringent safety requirements."

LIDAR HARDWARE

In a LiDAR project, the transimpedance amplifier is the most critical part of an electronic layout. Low noise, high gain, low group delay, and fast recovery from overload make the new Maxim TIAs ideal for distance-measurement applications.

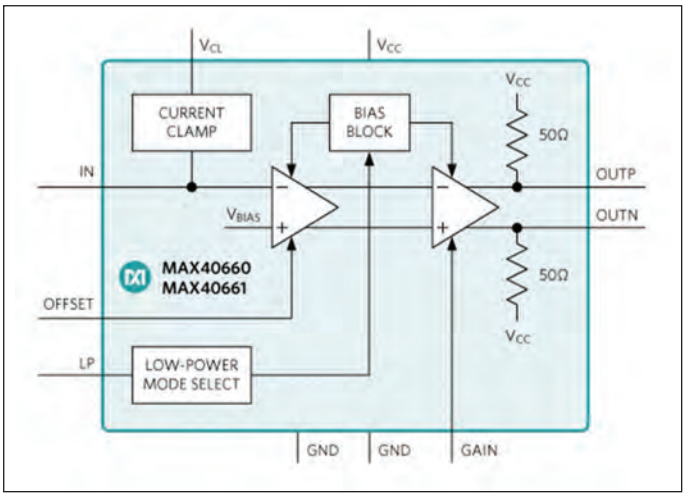


Figure 4: *MAX40660 block diagram* (Image: Maxim Integrated)

TIA circuits are often used in applications that share the need for circuitry to buffer and scale the output of electro-optical solutions to achieve high speed and high dynamic range. TIA is a current-to-voltage converter, almost exclusively implemented with one or more operational amplifiers (**Figure 3**).

Phototransistors and photodiodes are closely related and convert incident laser light into electrical current. To achieve maximum performance from these devices, designers must pay special attention to interface circuits, wavelengths, and optical-mechanical alignment.

The MAX40660/MAX40661 transimpedance amplifiers enable much faster self-driving systems using high resolution. The TIAs reduce current consumption by more than 80% in low-power mode.

Maxim's TIAs support 128 channels with a bandwidth of 490 MHz in the case of the MAX40660 and 2.1-pA/√Hz noise density to provide greater measurement accuracy (**Figure 4**).

The MAX40026, meanwhile, is a single-supply, high-speed comparator for TOF distance measurement applications. Its low, 10-picosecond propagation delay dispersion contributes to the accurate detection of fixed and moving objects.

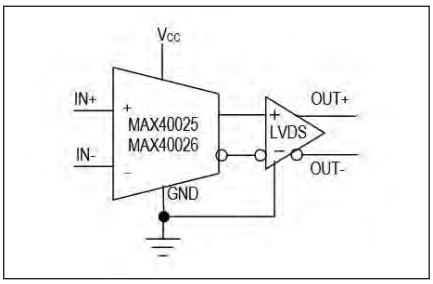


Figure 5: *MAX40026 functional diagram* (Image: Maxim Integrated)

"Lower dispersion delay and more channels per system enable more precise timing measurement, thus enhancing system resolution and enabling higher driving speed," said Gavardoni.

The MAX40026 has an input common-mode range of 1.5 V to VDD + 0.1 V, compatible with the output swings of several widely used high-speed TIAs. The low-voltage differential signaling (LVDS) output stage minimizes power dissipation and interfaces directly with many FPGAs and CPUs (**Figure 5**).

The size of the new solutions is further reduced, allowing many more channels to be inserted into vehicle platforms with limited space. These integrated circuits meet the automotive industry's most stringent safety requirements, with AEC-Q100 qualification, improved electrostatic discharge (ESD) performance, and effects and diagnostic analysis (FMEDA), in order to support ISO 26262 certification at the system level. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

NEWS

Europe Unveils Hydrogen Alliance to Combat Climate Change

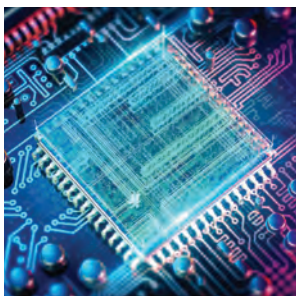
The European Commission has unveiled an industrial strategy that sees hydrogen playing a crucial role in Europe's transition to a sustainable future. Until now, the biggest challenge for hydrogen energy has been its cost. According to a recent Hydrogen Council report, "Path to Hydrogen Competitiveness: A Cost Perspective," scaling up hydrogen production and distribution, along with related equipment and component manufacturing, could decrease the cost of hydrogen solutions by up to 50% by 2030.

Europe is in the early stages of its energy transition to decarbonize all sectors. With the newly launched Clean Hydrogen Alliance, a key component of a vast Industrial Strategy, the EC aims to accelerate the decarbonization of industry and maintain its industrial leadership. Over the past decade, the EC has promoted several research and innovation projects on hydrogen within the Horizon 2020 framework. The Clean Hydrogen Alliance is expected to be modeled on the European Battery Alliance, which gathers more than 200 companies, governments, and research organizations around battery manufacturing. ■

First IBM Quantum Computer in Europe to Be Installed in Early 2021

IBM is slated to install its first quantum computer system in Germany as part of a collaboration agreement with Fraunhofer-Gesellschaft to provide European companies and research institutions with access to the technology for exploring application scenarios. It will also allow development of application-oriented quantum-computing strategies under the complete data sovereignty of European law.

As part of the collaboration, an IBM Q System One quantum computer will be installed in an IBM computer center near Stuttgart. The system is scheduled to go into operation in early 2021 and will be the first of its kind in Europe.



Fraunhofer plans to bring together established partners from research and industry under the umbrella of a research infrastructure of Fraunhofer institutes, which will work together in a centrally coordinated national Fraunhofer competence network for quantum computing. The network will initially be represented by competence centers in six German states.

Currently, more than 10 Fraunhofer Institutes are already working on various aspects of quantum technology. ■

NEWS IMAGES: SHUTTERSTOCK

POWER ELECTRONICS

High Power with SiC and GaN

By Giovanni Di Maria

The wide-bandgap (WBG) semiconductor materials silicon carbide (SiC) and gallium nitride (GaN) offer better thermal conductivity, higher switching speeds, and physically smaller devices than traditional silicon. The poor parasitic-diode characteristics of silicon MOSFETs produce high current peaks and high electromagnetic interference (EMI). The WBG materials have about 10× better conduction and switching properties than Si. Consequently, WBG technology is a natural fit for power electronics, particularly for electric cars, because the SiC and GaN components are smaller, faster, and more efficient than their silicon counterparts.

INCREASINGLY POWERFUL COMPONENTS

Among the positive aspects and improvements of SiC and GaN semiconductors over Si-based MOSFETs and IGBTs, the materials ensure lower losses, work with higher switching frequencies, endure much higher operating temperatures, are more robust in difficult environments, and offer higher breakdown voltages.

The electronics sector is moving toward larger high-voltage batteries with shorter charging times and reduced losses. The new materials are therefore very useful.

SiC, GaN, OR SILICON?

Wide-bandgap power devices (**Figure 1**) are expensive, and in some designs, the cost/performance considerations will not work in WBG's favor. Designers must weigh cost and performance compromises and, in some cases, evaluate substrates against each other.

The first SiC devices to be made available were simple diodes, but the material technology has since improved to allow the production of JFETs, MOSFETs, and bipolar transistors. GaN came after SiC and, in theory, is faster than SiC and allows higher switching speeds. But GaN adoption has been slow because of the material's high cost and reliability problems.

GaN voltages are currently limited to about 650 V. SiC voltages are commonly from about 650 V to 1,200 V but can range higher. SiC is widely used in the production of components and is cheaper, stronger, and more reliable than GaN. From a packaging point of view, SiC devices are available in TO-247 and TO-220 formats. This allows a quick and simple replacement of components, even in existing projects, with many immediate advantages. GaN devices use surface-mount packages, with consequential limits of use. One factor that gives SiC an advantage in industrial systems is its high reliability in overvoltage conditions. Conversely, the maximum voltage should not be exceeded for GaN devices.

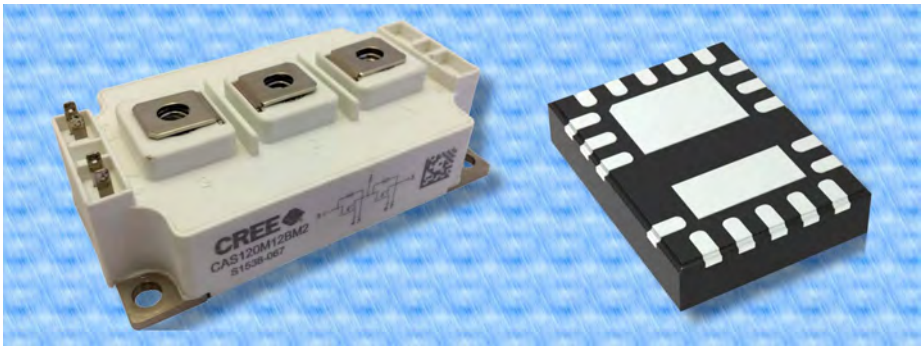


Figure 1: *Half-bridge SiC (left) and GaN devices* (Image: Wolfspeed)

Figure 2: *A silicon carbide inverter* (Image: PED-Board)

APPLICATIONS

Wide-bandgap devices work smoothly at high temperatures, high switching speeds, and low losses. For this reason, they are ideal for military and industrial applications. Their main use is with bridge circuits for high power, used in inverters (**Figure 2**), Class D audio amplifiers, and more. For high-power applications, robustness against short-circuit transients and surges is a critical consideration.

The inverter that controls the motor in an electric vehicle (EV) is an example of a system that can take advantage of WBG devices. The main function of the inverter is to convert a DC voltage to a three-phase AC waveform in order to operate the car engine. Because the inverter converts battery energy into alternating current, the lower the losses during this conversion, the more efficient the system will be. The higher conductivity and higher switching frequency of SiC devices compared with silicon reduce power loss because less energy is dissipated as heat. Ultimately, the increased efficiency of SiC-based inverters will result in greater EV autonomy.

A key element that acts as an interface between the controller and the power device is the gate driver. Gate driver design is always problematic for electronics designers who adopt new devices, and it is important to understand how to drive SiC and GaN power devices. The SiC MOSFET transistor must be operated with a higher gate voltage and must exhibit efficient voltage derivation over time (dV/dt) to achieve fast switching times. DC/DC converters also need to be designed to accommodate new components, such as SiC MOSFETs. They must have asymmetrical outputs for controlling SiC drivers. Insulation and parasitic capacitance are also important factors to consider in the design.

SPICE MODELS

Electronic components with SiC and GaN technology are increasingly popular, on both an industrial and a commercial level. For this reason, SPICE models for electronic simulations are proliferating on the internet.

Figure 3 presents a schematic of a test circuit for the UF3C065080T3S SiC FET, produced by UnitedSiC. The electrical characteristics of this component are truly stunning:

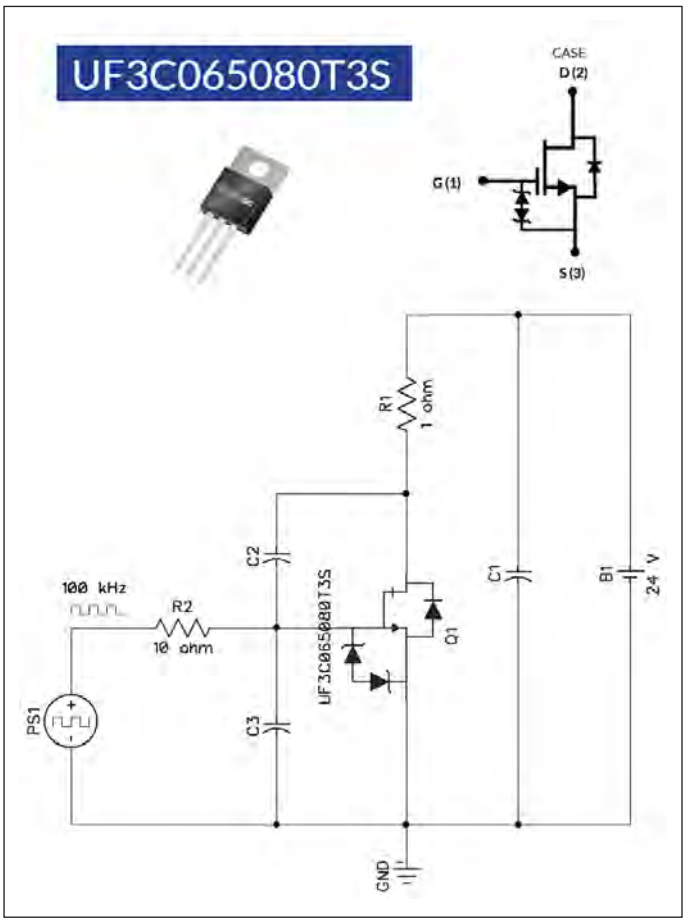


Figure 3: *Test circuit for the UF3C065080T3S, a MOSFET SiC JFET* (Image: UnitedSiC)

drain-source voltage (V_{DS}) of 650 V; gate-source voltage (V_{GS}) of –25 V to 25 V; continuous drain current (I_D) of 31 A at TC = 25°C, 23 A at TC = 100°C; pulsed drain current (I_{DM}) of 65 A; power dissipation (P_{TOT}) of 190 W; maximum junction temperature (T_{Jmax}) of 175°C; drain-source on-resistance ($R_{DS(on)}$) of 80 mΩ; and gate resistance (R_G) of 4.5 Ω. Typical applications are EV charging, photovoltaic (PV) inverters, switched-mode power supplies, power-factor correction (PFC) modules, motor drives, and induction heating.

The SPICE test is very intensive. As shown in the plots of **Figure 4**, the component switches its state at 100 kHz. The dissipated power is very high, but it works without any problem. You can easily find the SPICE model on the internet, enclosed in the statements ".subckt UF3C065080T3S nd ng ns" and ".ENDS".

As SiC and GaN technologies become increasingly popular, the products and devices available on the market are rising in number and performance. Let's examine some of them.

UF3SC120009K4S SiC FET

This UnitedSiC device is based on a unique "cascode" circuit configuration, in which a normally on SiC JFET is co-packaged with a silicon MOSFET to produce a normally off SiC FET (**Figure 5** shows an example of UnitedSiC's SiC FETs). The device's standard gate-drive characteristics make it a true drop-in replacement for Si-based IGBTs, FETs, MOSFETs, and superjunction

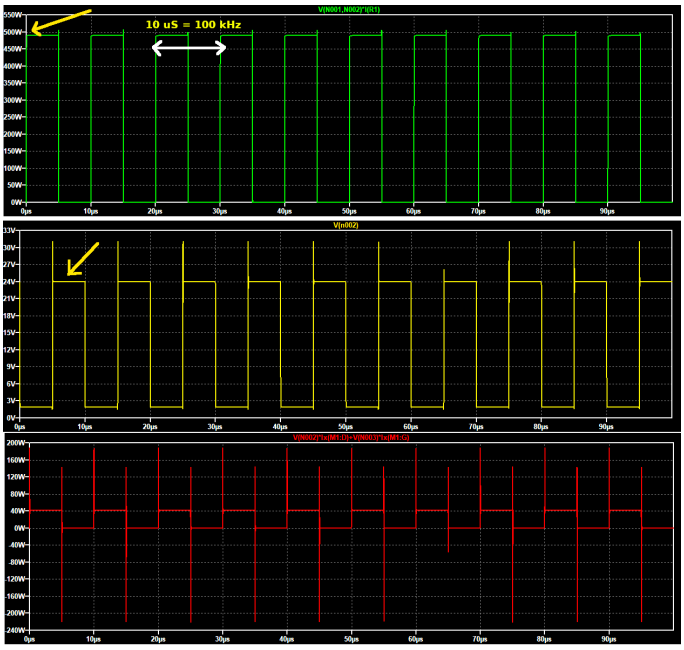


Figure 4: *Plots of the SPICE simulation at different points of the circuit* (Image: UnitedSiC)

devices. Packaged in the TO-247-4L format (resulting in faster switching and clean gate waveforms), the UF3SC120009K4S exhibits ultra-low gate charge and exceptional reverse-recovery characteristics, making it ideal for switching inductive loads and for any application requiring standard gate drive.

Notable characteristics include $R_{DS(on)}$ (typ.) of 8.6 mW, 175°C maximum operating temperature, excellent reverse recovery, low gate charge, low intrinsic capacitance, and HBM Class 2 electrostatic discharge (ESD) protection. Specifically, the impressive electrical characteristics include 1,200-V V_{DS}, –20-V to 20-V V_{GS}, 120-A I_D, 550-A I_{DM}, and 789-W P_{TOT}. Typical applications are EV charging, PV inverters, switched-mode power supplies, PFC modules, motor drives, and induction heating.

HALF-BRIDGE MOTOR DRIVERS

Power Integrations' BridgeSwitch (**Figure 6**) is a family of high-voltage, self-powered, half-bridge motor drivers with integrated device protection and system monitoring. The family's integrated half-bridges dramatically simplify development and production of high-voltage inverter-driven two- or three-phase permanent magnet

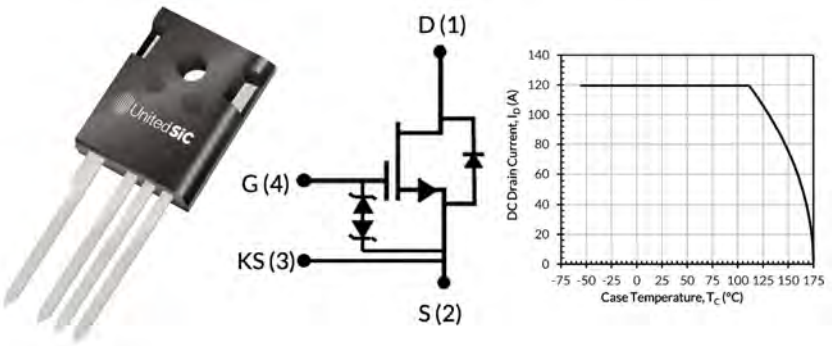


Figure 5: *UF3C065080T3S G3 SiC JFET* (Image: UnitedSiC)

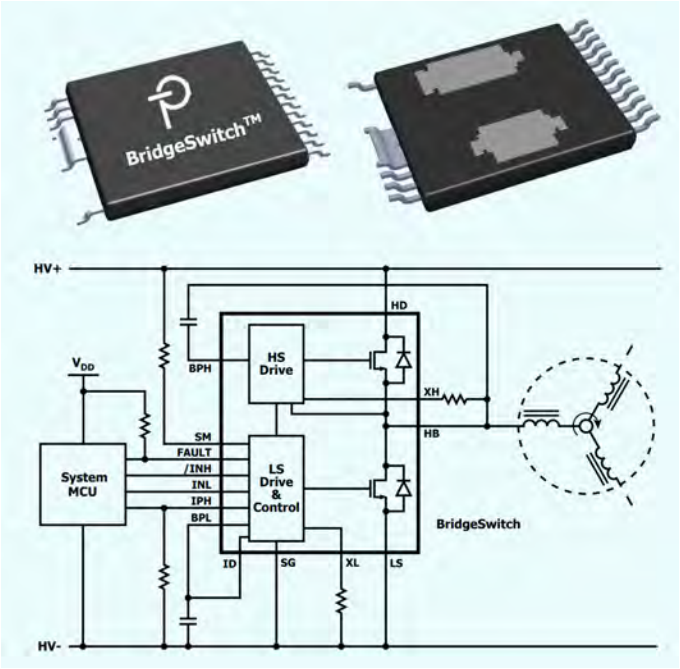


Figure 6: *BlueSwitch half-bridge motor driver*
(Image: Power Integrations)

Figure 7:
A GS66508T 650-V GaN transistor and its evaluation board
(Image: GaN Systems)

(PM) or brushless DC (BLDC) motor drives. The BridgeSwitch driver incorporates two high-voltage N-channel power FREDFETs with low- and high-side drivers in a single small-outline package. The internal FREDFETs offer ultra-soft and ultra-fast diodes ideally suited for hard-switched inverter drives. Both drivers are self-supplied, thus eliminating the need for an external auxiliary power supply.

BridgeSwitch provides a unique instantaneous phase current output signal, simplifying implementation of sensorless control schemes. The low-profile, compact-footprint surface-mount package offers extended creepage distances and allows heat-sinking of both power FREDFETs through the PCB. The driver offers internal fault-protection functions and external system-level monitoring. Internal fault protection includes cycle-by-cycle current limit for both FREDFETs and two-level thermal overload protection. External system-level monitoring includes DC bus sensing with four undervoltage levels and one overvoltage level, as well as driving external sensors. The bidirectional bused single-wire status interface reports observed status changes.

These are the absolute maximum ratings: HD pin voltage, –1.3 V to 600 V; HB pin voltage, –15 V to 600 V; DC output current, 1 A for the BRD1X60C, 1.7 A for the BRD1X61C, 3 A for the BRD1X63C, 5.5 A for the BRD1X65C; and junction temperature, –40°C to 150°C.

650-V ENHANCEMENT-MODE GaN TRANSISTOR

GaN Systems' GS66508T (**Figure 7**) is an enhancement-mode GaN-on-Si power transistor. The properties of GaN allow for high-current, high-voltage breakdown and high switching frequency. GaN Systems implements patented Island Technology cell layout for high-current die performance and yield. GaNPX packaging enables low inductance and low thermal resistance in a small package.

The GS66508T is a top-side-cooled transistor that offers very low junction-to-case thermal resistance for demanding high-power applications. Its features combine to provide very high-efficiency power switching. Its numerous applications include high-efficiency power conversion, high-density power conversion, AC/DC converters, bridgeless totem pole PFC, zero-voltage switching (ZVS) phase-shifted full-bridge/half-bridge topologies, synchronous buck or boost, uninterruptible power supplies, industrial motor drives, single inverter legs, solar and wind power, fast battery charging, Class D audio amplifiers, 400-V input DC/DC converters, on-board battery chargers, and traction drives.

To evaluate the entire GaN Systems 650-V family of GaN E-HEMT products, designers can use the universal motherboard with daughter-cards (sold separately). ■

Giovanni Di Maria *is a programmer, educator, technical writer, and EEWeb.com contributor.*

POWER ELECTRONICS

Wide-Bandgap Application in EV Chargers

By Maurizio Di Paolo Emilio

Electric vehicles (EVs) represent a fundamental factor for the success of e-mobility, thanks to their reduced environmental impact and lower operating costs compared with traditional internal combustion engine vehicles. While waiting for the EV charging network to reach a capillarity similar to that of common gas stations, electric vehicles must be equipped with on-board charging circuits that ensure high efficiency and long range.

The recharging of electric batteries requires, first of all, a conversion of the electric power source from alternating current (available on the electricity distribution network) to direct current. The circuit topologies used to perform this energy conversion are quite standard, including half-and full-bridge rectifier circuits and the classic "totem pole" configuration.

ACHIEVING HIGH EFFICIENCY

A classic EV charger circuit comprises a current rectification stage followed by a DC/DC converter stage. The rectifier circuit, composed of diodes with nonlinear characteristics, has a rather low power factor and a large number of undesired harmonic components. A high level of efficiency can be achieved only through careful design of the power factor correction (PFC) circuit.

To improve the power factor and reduce harmonic distortion, a solution based on active power factor correction (APFC) is commonly adopted. APFC is essential for an active switching circuit that receives a rectified voltage at its input and boosts that value until it reaches a DC set value, checking that the line current maintains the desired sinusoidal waveform. In principle, in an ideal PFC circuit, the input current "follows" the input voltage, behaving like a pure resistor and without manifesting harmonics in the input current.

In high-power devices like EV chargers, capable of handling the power of several kilowatts, active PFC is implemented using boost converter circuits. The boost converter, shown in **Figure 1**, causes the input current to be stored in an inductor for a certain time interval. Subsequently, when the switch S opens, the energy can reach the C0 capacitor passing through the D diode. The inductor behaves like a current source in series with the input current, and therefore, the output voltage is always higher than the input: with 220- to 240-VAC input, more than 340 V is obtained in output (380 to 400 V is commonly used worldwide). Note that the PFC stage is always followed by a DC/DC converter, with isolation of the output from the input.

WIDE-BANDGAP DEVICES

The circuit in **Figure 1** can be improved by replacing diodes with MOSFETs, each of them acting both as boost switch and synchronous rectifier. However, high-voltage MOSFETs usually have poor body-diode reverse-recovery characteristics; therefore, bridgeless totem-pole circuits have not been very common so far. The recent market of devices based on wide-bandgap (WBG) semiconductor materials, such as silicon carbide (SiC) and gallium nitride (GaN), has allowed the adoption of circuits for the implementation of an EV charger.

With a forbidden band two or three times greater than that of silicon-based components, WBG devices can withstand voltages and electric fields of higher intensity (electrons require two or three times more energy to pass from the interdiction zone to the conduction zone). As a consequence, the breakdown voltage of WBG devices is much higher, while the on-resistance is much lower. In high-power electronic circuits, such as EV chargers, a high breakdown voltage simplifies the design and improves efficiency. A reduced on-resistance value represents a further advantage for high-voltage circuits because it allows both switching and power losses to be reduced, enabling a particularly compact footprint.

A further advantage of WBG devices is their ability to generate lower temperatures than silicon-based devices operating under the same conditions. In a circuit for high-voltage applications, a SiC component can withstand junction temperatures higher than 200°C, compared with about 150°C for a silicon equivalent. The use of WBG devices in EV chargers enables higher switching rates and better energy efficiency, which, in turn, translate into more compact modules that are simpler to cool.

Analog Devices offers a wide selection of small-form-factor isolated gate drivers designed for the higher switching speeds and system size constraints required by power switching technologies, such as SiC and

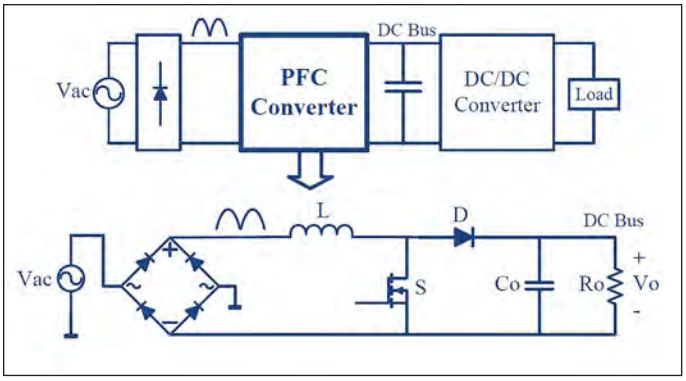


Figure 1: *Block diagram of a typical PFC boost converter* (Image: Analog Devices)

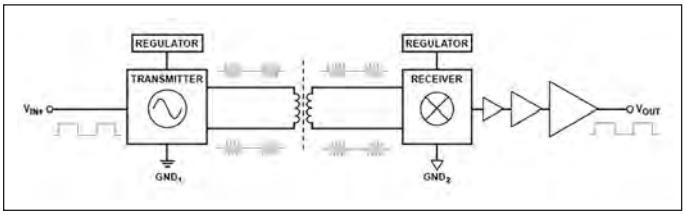


Figure 2: *Operational block diagram of the ADum4122 isolation feature* (Image: Analog Devices)

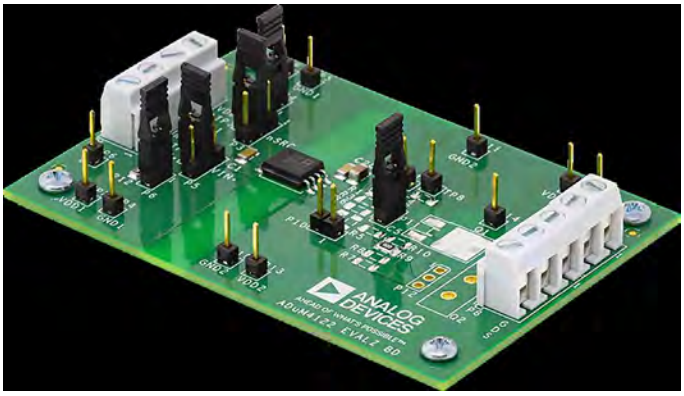


Figure 3: *The EVAL-ADuM4122EBZ evaluation board* (Image: Analog Devices)

plexim
electrical engineering software

RT BOX 1:
THE ORIGINAL

RT BOX 2:
MULTI-CORE

RT BOX 3:
HIGH I/O COUNT

THE REAL-TIME FAMILY HAS GROWN

Building blocks for HIL simulation
and rapid control prototyping

Find specs and pricing at
www.plexim.com

GaN. These isolated gate drivers are based on ADI's proven iCoupler isolation technology, combined with high-speed CMOS and monolithic transformer technology. The ADuM4122 device, an isolated dual-output driver, provides 5-kV RMS true galvanic isolation between the input and output regions.

Gate drivers are required for achieving fast rise times for switching device gates. As shown in **Figure 2**, ADuM4122 achieves isolation between the control side and the output side of the gate driver by using a high-frequency carrier that transmits data across the isolation barrier with iCoupler chip-scale transformer coils separated by layers of polyimide isolation.

The EVAL-ADuM4122EBZ evaluation board (**Figure 3**) supports the ADuM4122 isolated gate driver with slew rate control. The evaluation board supplies jumpers and screw terminals to configure different drive conditions, accepting both square waves and DC values on the V_{IN+} and SRC pins. The SRC pin controls whether the V_{OUT_SRC} pin is either set to high-Z or follows the logic of the user-supplied pulse-width modulation (PWM) input at V_{IN+}. When the external series gate resistors combine the outputs from the V_{OUT} and V_{OUT_SRC} pins, one isolated gate driver is able to have two easily selectable slew rates.

The EVAL-ADuM4122EBZ board tests the propagation delay, drive strength, slew rate selection, and input logic of the device.

WIRELESS CHARGING

Wireless EV charging uses an inductor, usually placed under the asphalt, and a receiver onboard the vehicle. Charging occurs automatically through magnetic plates that continuously recharge the batteries, whether the vehicle is standing still or in motion.

In the area of wireless power transfer, engineers need solutions with high-power and high-efficiency GaN-based approaches. GaN Systems' broad portfolio of transistors provides high-power wireless charging solutions to design smaller, cheaper, and more efficient power systems for demanding applications. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

POWER ELECTRONICS

Wide-Bandgap Semiconductors Find Homes in Space

By Maurizio Di Paolo Emilio

Wide-bandgap (WBG) semiconductors, such as gallium nitride (GaN) and silicon carbide (SiC), are proving to be the most promising materials in the field of power electronics since silicon was introduced. These materials have several advantages over traditional silicon-based technology, such as the ability to manage high power levels, insensitivity to radiation, high-temperature operation, high switching frequencies, low noise, low power losses, and high efficiency.

WBG semiconductors are of strategic importance to the development of next-generation space-borne systems. Gallium nitride, in its enhanced-mode version (eGaN), is widely used in the development of FETs and HEMTs for space applications.

EFFECTS OF RADIATION ON POWER DEVICES

The space environment has particular conditions that can influence and, in some cases, degrade the mechanical characteristics of space-based materials, which can negatively influence the overall behavior of a system's operation.

Space radiation consists primarily of 85% protons and 15% heavy nuclei. The effects of radiation can lead to degradation, interruptions, and discontinuities in the performance of devices.

The main requirement for space-qualified components is the ability to ensure reliable long-term operation. A radiation-hardened, or rad-hard, design determines the requirements of an electronic component to withstand the effects of radiation. It can be one of the most expensive and time-consuming approaches, but it is sometimes the only solution for electronic components in order to protect human lives or safeguard important orbital missions in space.

Electronic components used in space-borne applications are primarily subjected to space radiation, known as single-event effect, or SEE, caused by electrons and protons trapped in Earth's magnetic field. Another important effect of space radiation is the total ionizing dose (TID). The difference between the two concepts is very simple: SEE

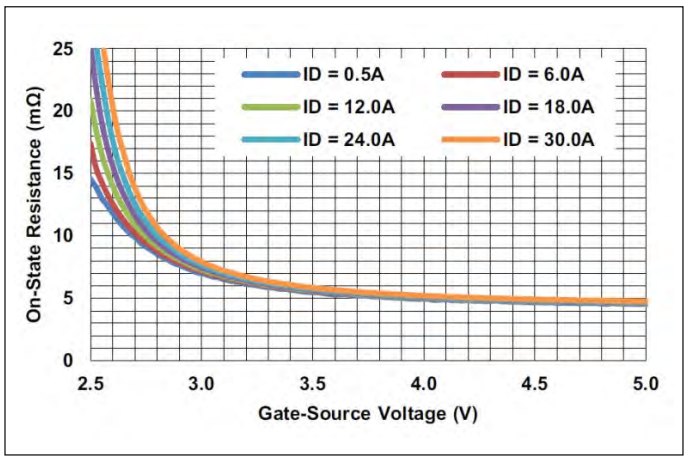


Figure 1: *Low on-state resistance of the Renesas ISL70023SEH*
(Image: Renesas Electronics)

is the result produced by a single high-energy particle that hits the device, while TID measures the effects produced by prolonged exposure to ionizing radiation.

TID exposure, measured in radiation-absorbed doses (rads), quantifies the total exposure of a material to radiation. Given a specific device, the total dose radiation threshold is the minimum rad level that will cause device failure. Most rad-hard commercial devices can withstand up to 5 krads before functional failure occurs. The SEE indicator becomes particularly significant in applications such as satellites and spacecraft. The high density of protons and ions present in the environment in which these systems operate can cause a series of different SEEs in electronic circuits, including single-event upset (SEU), single-event transient (SET), single-event functional interrupt (SEFI), single-event gate rupture (SEGR), and single-event burnout (SEB).

SEE events can cause a degradation of system performance, up to total destruction. In order to ensure a high degree of reliability, it is

necessary to select components in which the effects produced by radiation have been measured and declared.

WBG ADVANTAGES IN SPACE-BORNE SYSTEMS

Reduced weight and size, together with high efficiency and reliability, are fundamental requirements for components intended for use on spacecraft. GaN power devices provide the highest level of efficiency in the smallest footprint available today. Gallium nitride also has excellent characteristics in terms of electromagnetic compatibility (EMC): The reduced parasitic capacitance decreases the energy stored and released during the switching cycles, while the reduced footprint improves the loop inductance, particularly insidious as it acts as a transceiver antenna.

Power devices used in critical applications such as space missions, high-altitude flights, or strategic military applications must be resistant to failures and malfunctions caused by ionizing radiation. Commercial GaN power devices offer significantly higher performance than traditional rad-hard devices based on silicon technology. This allows the implementation of innovative architectures for applications in satellites, data transmission, drones, robotics, and spacecraft.

ENHANCED GaN HEMT

Rad-hard MOSFETs have reached their technology limits with large die sizes and a performance figure of merit (FoM), expressed by the formula $FoM = R_{DS(ON)} \times C_{iss}$, which is much higher than that of an eGaN transistor. The FoM is a very important parameter: The smaller the value, the better the efficiency of the system.

In addition, eGaN HEMTs are easier to drive, as they require 10× to 40× less gate charge than the best rad-hard MOSFETs. GaN devices can also be mounted directly on the ceramic substrate without requiring any external package. It is possible to eliminate wire bonds and related inductance, enabling very high switching rates. The eGaN switching speeds are determined only by the resistance and capacitance of the gate and drain nodes.

Switching times can easily reach sub-nanosecond levels, so particular attention should be paid to both the design and PCB layout phases of development when using these high-performance devices.

RAD-HARD GaN SOLUTIONS

Renesas Electronics, a leading supplier of advanced semiconductor solutions, has developed the industry's first rad-hard 100-V and 200-V GaN FET power solutions, suitable for enabling primary and secondary DC/DC converter power supplies in space-borne systems. These GaN FETs have been characterized for destructive single-event effects and tested for TID radiation. The ISL7023SEH 100-V, 60-A GaN FET and ISL70024SEH 200-V, 7.5-A GaN FET provide up to 10-orders-of-magnitude-better performance than silicon MOSFETs while reducing package size by 50%.

Figure 3: *Freebird's GaN adapter module* (Image: Freebird Semiconductor)

They also reduce power supply weight and achieve higher power efficiency with less switching power loss. At 5-mΩ $R_{DS(ON)}$ and 14 nC (QG), the ISL70023SEH enables the industry's best figure of merit. **Figure 1** shows the very low $R_{DS(ON)}$.

VPT Inc. offers the SGRB series of DC/DC converters, specifically designed for harsh radiation environments in space applications. Based on advanced GaN technology, the SGRB series provides high efficiency, resulting in reduced system size, weight, and cost.

With up to 95% efficiency, the series' GaN technology results in greater efficiency compared with traditional radiation-hardened silicon products. It has been designed specifically for space-borne telecommunications in which high efficiency, low noise, and radiation tolerance are imperative (**Figure 2**).

Freebird Semiconductor offers a wide selection of high-reliability GaN HEMT discrete devices integrated into GaN adapter modules (GAMs), creating the patented circuitry found in its multifunction power module series. These universal GaN adapter modules (**Figure 3**) incorporate eGaN switching power HEMTs with GaN-based high-speed gate drive circuits for use in commercial satellites.

The rad-hard FBS-GAM01-P-C50 single low-side power development driver module incorporates GaN switching power HEMTs in a nine-pin SMT overmolded epoxy package. Integrated devices include Freebird's FDA10N30X output power eGaN HEMT switch and an output clamp Schottky diode, optimally driven by high-speed gate drive circuitry consisting entirely of eGaN switching elements. It also includes 5-V input V_{BIAS} overvoltage clamping protection with V_{BIAS} undervoltage driver disable and reporting. The SMT overmolded epoxy package provides an engineering development platform for the FBS-GAM01-P-R50 flight unit version.

Figure 2: *VPT's SGRB series of DC/DC converters* (Image: VPT Inc.)

CONCLUSION

A reliable, continuous power supply is essential to the success of a space mission. In real-world applications, the main advantage of switching to SiC- or GaN-based broadband semiconductors is the increased power conversion efficiency.

The ability of SiC- or GaN-based broadband semiconductors to operate at high temperatures also has significant advantages. Not only can these devices be used in higher heat environments, they require less overall cooling, reducing the space and cost of cooling components in the power converter. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

Power Electronics Conference

Technical Trends with Wide Bandgap Devices

December 14-15 2020
Hilton, Munich Airport

Power Electronics is rapidly moving towards Wide Bandgap Semiconductors, as the key for the next essential step in energy efficiency lies in the use of new materials, such as GaN (gallium nitride) and SiC (silicon carbide) which allow for greater power efficiency, smaller size, lighter weight and lower overall cost.

Wide Bandgap Semiconductors are transforming power electronics designs across many applications including data centers, renewable energy and automotive as well as many others.

Our technical conference will explain why, how and where this is happening. Conference delegates will be provided with the knowledge necessary to make their decisions on where, how and which Wide Bandgap platforms and devices can play a role in current or upcoming designs.

POWER-CONFERENCE.COM

Bodo´s Power Systems®

POWER ELECTRONICS NEWS

ASPENCORE

INTERNET OF THINGS

IoT Security Requires 'Multi-Tiered Approach,' Says ST's Scarlatella

By Maurizio Di Paolo Emilio

The IoT ecosystem has undergone explosive growth, thanks to the rapid proliferation of voice assistants and connected devices (such as intelligent thermostats and security cameras) and even intelligent infrastructure. Today, connected devices are integrated into critical management systems for air traffic control, energy networks, environmental controls, and many other high-value systems.

As connected devices on the internet of things become even more widespread, safety becomes a central issue. This will inevitably lead regulators to insist that developers respond accordingly with appropriate security measures for a consumer audience increasingly aware of the positive and negative aspects of ubiquitous connectivity.

IoT devices are used to acquire, process, and transmit data and to control processes. They are designed for everyday use by consumers, who, attracted by the innovation and many features of these products, often undervalue the security aspects. This becomes worrisome for an industrial vision in which security is important both for the physical safety of workers and for safeguarding production against tampering. With the ubiquitous connectivity of the IoT, the concern over potential breaches extends beyond data collection, data sharing, and tampering by third parties to the possibility that malicious parties can "take control" of connected objects.

"Security is now, and has always been, a moving target," said Michele Scarlatella, director at Strategy, Technology, & Systems Architecture in the Microcontrollers and Digital ICs Group at STMicroelectronics. "As the industry develops more and better ways to protect data and systems using offensive, defensive, and observational measures, the bad guys continue to find ways to bypass or break the protections. While we've developed the capabilities for an incredible level of protection, that level presents challenges in convenience, usability, and cost.

"Ultimately, the key to a truly secure industrial IoT is developing a multi-tiered approach where all parts of the system are watching and protecting each other from attempts to break the system.

"As a chip company, ST has gotten pretty good at securing the information on, and the operation of, our chips. As developers introduce their own programs and start connecting these chips to others, the potential for new vulnerabilities increases. Therefore, as a system is only as good as its weakest security, it is critical that developers recognize the importance of security, build their systems with secure components, and follow industry best practices."

But the rise of the IoT has also been accompanied by a dizzying increase in cyberattacks, with malicious users trying to exploit vulnerabilities in devices manufactured with little or no cybersecurity protection. Increasing the number of devices connected to the network also increases the number of access points for a possible attack on the corporate information system.

"An important security factor in accelerating the growth of the IIoT is giving the ultimate customer the confidence that their systems and valuable user data are extremely well-protected from intrusion," said Scarlatella. "With our Secure GP MCUs and Secure Elements and the ecosystem around them, we give designers and our customers a set of tools they can use to build these kinds of systems.

"With the STM32Trust initiative, we provide a complete set of cryptographic libraries — reference components to build a secure software update system. We pay strong attention to certification, such as the Arm PSA [Platform Security Architecture] and SESIP [the Security Evaluation Standard for IoT Platforms].

"That said, while we continue to reduce the risks, security comes at a price, whether that price is in convenience and usability or in cost," he added. "And unfortunately, the truth is that a secure system almost always has a higher upfront cost than one without security — until you've had a serious problem. In the longer term, the reduced liability and secure product upgrades will bring their benefits."

The IoT market represents an extraordinary opportunity for all products that integrate a sensor, processor, and connectivity functions: portable or mobile devices, along with solutions for entertainment, automotive, agriculture, energy, health care, industry, and much more. The type of processing used within the devices is strongly influenced by the feedback needed for the target application. For example, some smart processors perform a limited amount of processing on simple data sets, such as temperature, humidity, or pressure. Others, more complex, are required to handle high-resolution audio or video data, to correlate environmental information with self-learning databases, or to communicate within distributed mission-critical networks.

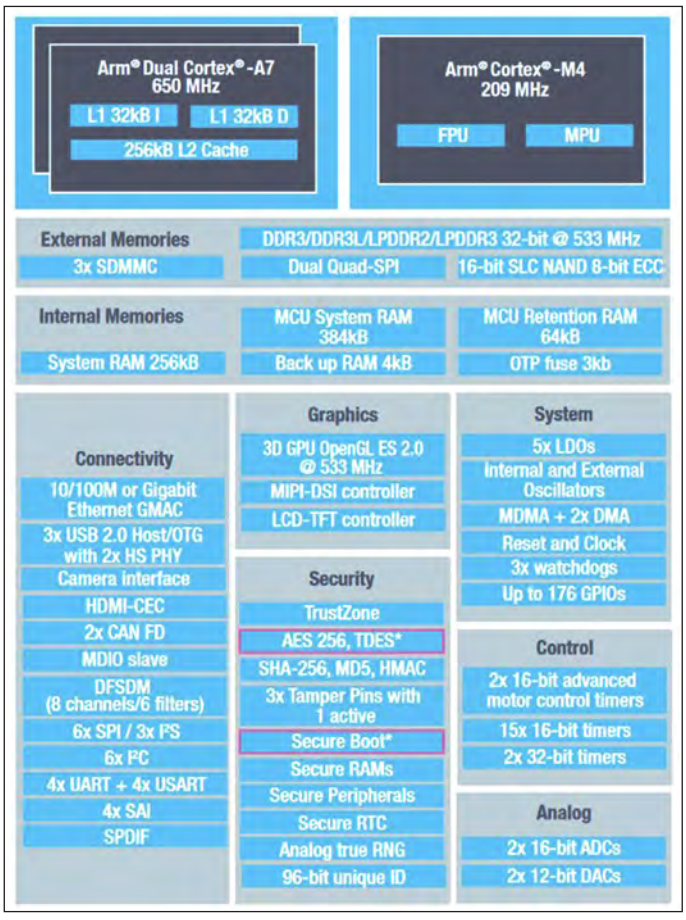


Figure 1: *STM32MP1 block diagram* (Image: STMicroelectronics)

ST's Scarlatella on IoT Security

STMicroelectronics' STM32MP1 multicore processors integrate a dual-core application processor, a GPU for graphics, and an MCU core for power-efficient hard real-time tasks. This multicore approach, with high functional integration, will facilitate the development of high-performance solutions for industrial, consumer, smart-home, and health and wellness applications. ST's release of OpenSTLinux as a mainlined and open-source Linux distribution for the STM32MP series will facilitate the developer's job in meeting essential customer needs for applications in the IoT market (**Figure 1**).

The STM32MP1 has a TrustZone that can be used to build secure products, with valuable system integrity checks. The processor has a secure boot ROM code that can be used as the root of trust of the platform. It can be securely provisioned by the customer using the STM32Trust ecosystem solution. When activated, the STM32MP1 can initialize its TrustZone and its Trusted Platform Module (TPM) environments using authentication software. External secure elements such as TPM offer additional secure services, including application secure-key storage, chain of trust boot authentication, and secure cryptographic functions.

"The industry's reception to the STM32MP1 microprocessor has been terrific," said ST's Scarlatella. "Since launching the product at embedded world last year, we've now got 24 Authorized ST Partners in our partner program offering a range of services and products; we've won several designs and are being evaluated in many, many others.

"In fact, we've just launched several new STM32MP1 MPUs with better performance and security, including trusted boot authentication, available OTP [one-time programmable] fuses for customers, and a secure operating system. The enhanced STM32MP1 toolset also allows secure provisioning of customer code into the chip. So we're constantly moving forward in making the SoCs more protected."

Hardware protection that starts at the individual component level allows for the long-term protection of sensitive data and applications, even when threatened by attacks that are increasingly difficult to detect and contain. Intelligent sensors are essentially connected microcontrollers that integrate several analog interfaces for detection. Manufacturers of microelectronic solutions are engaged in the difficult task of creating versatile, scalable solutions designed to address a variety of IoT devices and systems. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

OPINION

Smart-Device Success Hinges on Unclonable Security

By Maurizio Di Paolo Emilio

The internet of things brings together billions of devices every day, and the IoT's exponential growth has transformed the way we go about our daily activities at home, on the job, and on the move. In many ways, our lives have improved as a result. But like any good thing, there is a downside to IoT: It is becoming an increasingly attractive target for cybercriminals, with far too many IoT devices left vulnerable to cyberattacks.

IoT technologies now let manufacturers obtain information of value that can be exploited in person, transferred to the consumer as added value after the purchase of the product, or monetized through the provision of new services. New developments will allow vehicles to be connected with smart-city infrastructure to create an entirely different ecosystem for the driver. And connected health devices can represent a new, more economical business model for the health-care industry, with a strongly positive impact on medical outcomes and overall health.

The implementation of IoT technologies provides an infrastructure that can be articulated over several networks and systems. Each part hides critical issues that must be assessed to avoid risks to security and network efficiency and to prevent uncontrolled growth of the data collected and the resulting processing and storage costs. And as IoT applications proliferate, more devices are being deployed in uncontrolled and hostile areas that render them vulnerable to physical attacks. These attacks are more sophisticated than software threats, which are characterized by poor cryptographic implementations or the exploiting of predefined passwords. Designers need advanced system defenses for their critical applications, in which exposure to secret keys could destroy networks, ruin reputations, disrupt businesses, and even put lives at risk.

The success of any application depends on providing robust solutions with security features that offer reliable protection.

Attacks to IoT systems are becoming more advanced. Devices are strongly at risk for hacker attacks such as compromises to information and customizations. With the growing trend of "everything connected," techniques such as secure boot, secure key saving, encryption, and authentication are essential. While designers understand the importance of preventing device-level security breaches, they often struggle with compromises between adding advanced security features and minimizing board space, all while balancing sophisticated design and time-to-market demands.

The success of any application depends on gaining and keeping users' trust by providing robust and easy-to-use solutions with security features that offer reliable protection. The large number of devices and their wide distribution create a huge attachment area that is difficult to monitor. Some sensors and communication protocols are designed to be used on closed private networks and either address cybersecurity unsatisfactorily or ignore it altogether. A hardware-based approach is more robust than software-based protection and encryption strategies because it is more difficult to alter a physical protection system.

At embedded world, Maxim Integrated introduced the MAX32520 ChipDNA Secure Arm Cortex-M4 microcontroller, a device that integrates physically unclonable functionality (PUF) technology for multiple levels of protection in IoT, health care, industrial, and IT systems. The main encryption key created by the authenticator's PUF circuit is generated at the precise moment that it's needed, exploiting the intrinsic uniqueness of the analog characteristics of each MOSFET circuit. The PUF-based device does not store any keys in non-volatile memory.

That's a boon for connected-device security. You can't steal a key that isn't there. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

INTERNET OF THINGS

Semtech's LoRa Edge Speeds IoT Applications

By Maurizio Di Paolo Emilio

The development of the IoT brings with it design challenges that many companies are trying to solve by providing designers with relevant tools. Over the next decade, more than 500 billion devices will be connected to the internet, and a robust corporate strategy will be needed to provide devices with location capability and proper energy management.

IoT will affect all aspects of life, covering applications from wearable connected devices to asset management, monitoring, and sensors that communicate with one or more gateways over distances of up to several kilometers. These devices are usually intended for service of up to eight years, so it necessary to choose a network connectivity approach that will be supported during this period. Typical IoT applications require data transmission in certain situations only, depending on the parameters to be measured, and do not require high bit rates. These characteristics are met by low-power wide-area network (LPWAN) technologies, which demand low transmission power and provide excellent coverage and scalability.

Semtech's LoRa Edge platform enables sensor management in a wide range of applications. Shown at embedded world, the platform integrates an ultra-low-power LoRa (long range) transceiver, GNSS, and Wi-Fi scanning technologies. The combination of LoRa technology and cloud solutions with geolocation provides the right balance between location accuracy and low power consumption, optimizing system cost and complexity. By eliminating the need for incremental GNSS and Wi-Fi components, LoRa Edge reduces BOM costs.

Figure 1: *LoRa Edge geolocation solution* (Image: Semtech)

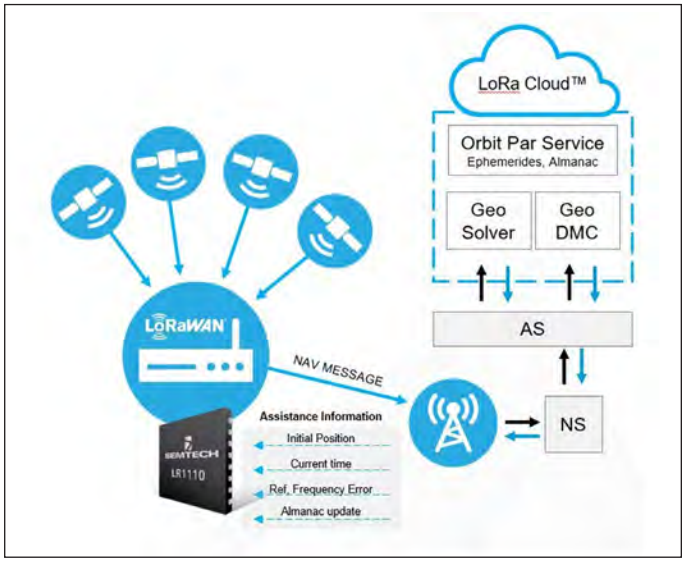


Figure 2: *GNSS scanning principle, assist mode* (Image: Semtech)

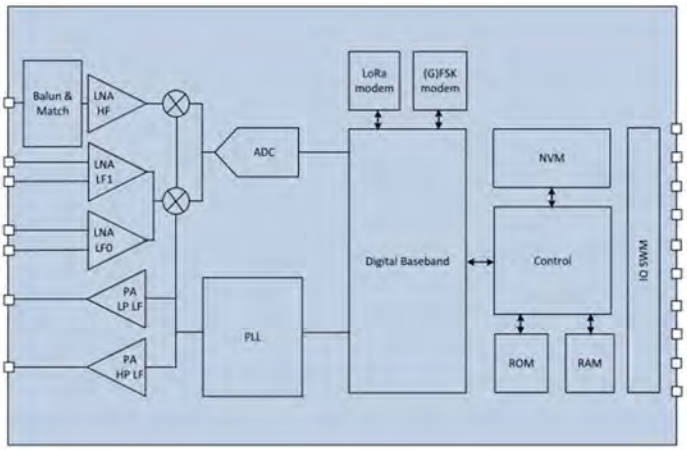


Figure 3: *LR1110's block diagram* (Image: Semtech)

LoRa has significant technical credentials and is already in use in applications that require reliable communications capability over distances of several kilometers, such as wireless instrument reading and street lighting control systems. This sub-gigahertz technology supports data rates from 0.3 Kbps to 50 Kbps, depending on distance and message duration. Transmission distance can be up to 15 to 20 km. Even in a high-density urban environment, communication distances of more than 2 km are achievable.

"LoRa is long-range, [so] you don't need a lot of base station repeaters," said Olivier Beaujard, senior director, LoRa Ecosystem at Semtech. "People can have [a] public [or] private network solution based on LoRa, which is already in use for tracking applications. But most importantly, LoRa is extremely low-power. And what does [that] mean? People can do [a] very small tracker" with a small battery.

LoRa's low-power capabilities were "proven with a lot of use cases in utility, water, and gas," he added. "In asset management, it makes a difference in many situations."

The LoRa technology stack comprises two layers: the physical layer, using a proprietary modulation derived from chirp spread spectrum, and the protocol for the media access control sublayer called LoRaWAN. A LoRa network includes gateways to connect to the central network server. Endpoints communicate with a star network topology via a single-hop wireless link to gateways with the ability to connect to multiple gateways to ensure connection redundancy (**Figures 1** and **2**).

Semtech's LR1110, the first LoRa Edge chipset targeted with geolocation, captures a portion of a satellite broadcast signal. The signals are aggregated to NAV message and sent to the cloud server. GNSS scanner supports GPS L1 + GPS geostationary, SBAS: EGNOS + WAAS, and BeiDou B1 + BeiDou geostationary GEO/IGSO (**Figure 3**).

LoRa Edge is an all-in-one chipset for indoor/outdoor applications. It allows customers to further manage the total cost of ownership, paying only when an asset needs to be identified. Improved key provisioning at the point of production and a secure join process simplify IoT solution development. LoRa Edge and LoRa Cloud geolocation services offer unique system architecture to reduce power consumption. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

INTERNET OF THINGS

Efficiency and Energy Savings with Networked Lighting Control

By Maurizio Di Paolo Emilio

The main functionality of a lighting system is to provide a space with the amount and quality of illumination appropriate for the specific context while minimizing energy consumption. Lighting control strategies that sense and respond to the presence of occupants, and that adjust the output of artificial lighting as daylight conditions change, can substantially reduce electricity use compared with traditional lighting systems.

The pairing of LED lighting with wired or wireless network communication has allowed the creation of intelligent lighting control systems that integrate sensors, advanced information processing, and connectivity. Networked lighting systems enable advanced features such as presence detection, flexible programming, networking to the remote system, daylight adaptability, and real-time monitoring. These features help building operators reduce the maintenance costs and increase the efficiency of their structures, improve the ambient environment for occupants, and exploit the new capabilities of IoT infrastructure.

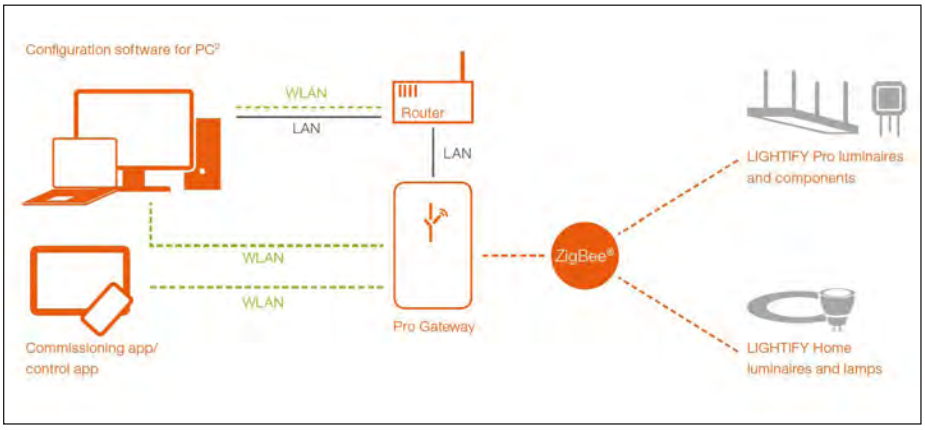


Figure 1: *Possible configuration of the LIGHTIFY system* (Image: OSRAM)

After the adoption of LEDs, the advent of networked lighting control is regarded as the second revolution in the lighting sector. Networked lighting control systems offer a great opportunity for cost savings and improved environmental conditions in new structures and renovated buildings alike.

Wireless network technology, for example, leads to a reduction in wiring with a consequent decrease in installation and maintenance costs, combined with the possibility of remotely controlling and programming the brightness of different light points. LEDs require relatively simple control circuits, do not lose efficiency even at low brightness, and withstand a large number of on/off switching cycles. In addition to adjusting the light intensity, LED dimming lets occupants and building managers dynamically vary the gradation of light in order to meet different lighting needs.

Networked lighting combines LED technology with intelligent control systems capable of maximizing energy savings through flexible and dynamic programming of light points divided into zones, remote control of operations via apps, and measurement and monitoring in real time. Features such as high data-processing capacity, wireless connectivity, and the possibility of installing additional sensors in the LED light points allow deployment of networked lighting systems suitable for supporting the building internet of things (B-IoT). In a recent study based on the detection of electricity consumption in 100 contexts, the DesignLights Consortium (DLC) found that networked lighting control systems can yield savings of close to 50% in an installation's electricity use.

COMPONENTS OF A NETWORKED LIGHTING SYSTEM

A networked lighting control system consists of dimmable drivers, ballasts, intelligent light and temperature sensors, manual switches, power supplies, and all the hardware and software components necessary to perform the control, configuration, and programming of the system, including storage of server or cloud information. Sensors and control systems are connected via wired or wireless connection to form a network in which the addressable devices can be configured and programmed individually or grouped into zones.

Depending on the degree of complexity, the control system can be confined to a single device or to a room, a building, or an entire urban area. As for the individual devices, some manufacturers offer solutions consisting of an LED driver that can be controlled remotely (via app and Wi-Fi network, or by means of a special remote control) and sensors integrated in the same device.

In terms of performance and benefits, whole-building lighting control systems are the most efficient solutions. They enable accurate multilevel control strategies, collecting consumption data for subsequent processing and analysis. This data proves particularly useful for supporting maintenance activities. Although the whole-building solution is technically complex, it offers the greatest potential to reduce energy consumption and provide the best possible lighting conditions for occupants.

The advantages offered by network lighting can be summarized as follows:

Reduction of electricity consumption and costs associated with lighting systems

All companies look to minimize costs in order to maximize profits. A lighting control system can contain costs by ensuring that the correct type of lighting is deployed and activated where and when it is required. Lighting costs can account for up to half of a building's total costs for electricity consumption. By using programmable timers, motion sensors (capable of detecting when occupants enter or leave a room), and brightness sensors (capable of balancing artificially produced light with daylight), it is possible to reduce lighting costs drastically.

Improved maintenance

Remote monitoring of devices connected to the network lets the system automatically send alarm notifications via email or SMS when problems arise. It is also possible to produce daily reports indicating which devices require maintenance or replacement.

More efficient use of lighting within a space

Presence detection allows optimized electricity consumption, automatically dimming or switching off the lighting in unoccupied meeting rooms and offices.

Figure 2: *A Dynalite sensor for wall mounting* (Image: Philips)

Optimal, customized lighting characteristics
With the ability to control the light level, spatial brightness, and hue of the lighting in a space, users are able to set their own local lighting preferences.

Support for the B-IoT infrastructure
Potentially, each lighting device can integrate a series of sensors that can monitor different physical aspects in real time.

NETWORKED LIGHTING SOLUTIONS

OSRAM's LIGHTIFY solution is an innovative intelligent lighting control system with wireless functionality. Operators and occupants can use a smartphone or tablet to program LIGHTIFY ceiling lights as well as lamps, switches, pushbutton couplers, sensors, or entire LIGHTIFY systems via the network. As shown in **Figure 1**, the wireless interface used by the devices is based on the Zigbee standard, and the control systems connect to them via a gateway.

Zigbee is a standard for wireless data transmission developed specifically to manage short messages, such as those required by smart-home and smart-building applications. Zigbee technology ensures low energy consumption and high system longevity without maintenance.

LIGHTIFY Pro targets professional applications with installations of up to 100 light points, such as offices. The Home version is for domestic applications with up to 50 light points.

Philips also offers solutions for both the home and professional sectors. Its wide portfolio of products for lighting control ranges from simple wall dimmers to intelligent sensors and the most advanced management and lighting control software. The Philips Dynalite series is among the most widely deployed, with thousands of installations worldwide. It includes a complete set of lighting control devices and is available in a wireless version, based on the Zigbee protocol, that offers all the benefits of an intelligent control system without the need for extra wiring.

Figure 2 shows a completely remotely controllable sensor in the Dynalite series. Dynalite sensors combine motion detection, light level detection, and IR reception in a single unit controlled by a microcontroller.

Lutron Electronics, another leader in lighting control systems, offers products ranging from dimmers and switches to turnkey solutions for managing entire buildings. The Lutron LCP128 lighting control system integrates the control of all lighting circuits, both on/off and dimmable, both internal and external, in a single, simple operating system. The circuits can be activated automatically based on daily programs or manually through intuitive wall devices.

The LCP128 system is ideal for small business installations such as restaurants, shops, spas, and recreation centers. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

NEWS

Open-Source Initiative Targets Energy Transition

Growing and supporting renewable energy through a truly collaborative open-source initiative is the objective of LF Energy. LF Energy is a nonprofit, vendor-neutral initiative from The Linux Foundation with an action plan to modernize electrical systems worldwide through open frameworks, reference architectures, and a support ecosystem of complementary projects.

"Our mission is to accelerate the energy transition by hosting, building, facilitating, and enabling the distributive computing paradigm as it relates to distributed energy resources," said Shuli Goodman, executive director of LF Energy. "That [covers] everything from control infrastructure to the foundations for new markets, micro-transactions, the edge devices — the whole thing."

LF Energy takes care of reusable components, open APIs, and interfaces through projects to adopt in platforms based on open-source code libraries. Systems integrators, vendors, developers, and end users can solve complex and interconnected problems with secure and flexible open-source software. ■

ARTIFICIAL INTELLIGENCE

Top 10 Processors for AI Acceleration at the Endpoint

By Sally Ward-Foxton

While the acceleration of artificial-intelligence and machine-learning applications is still a relatively new field, there is a variety of processors springing up to accelerate almost any neural network workload. From the processor giants down to some of the newest startups in the industry, all offer something different — whether that's targeting different vertical markets, application areas, power budgets, or price points. Here is a snapshot of what's on the market today.

APPLICATION PROCESSORS

Intel Movidius Myriad X

Developed by Movidius, the Irish startup that was bought by Intel in 2016, the Myriad X is the company's third-generation vision-processing unit and the first to feature a dedicated neural network compute engine, offering 1 tera-operations per second (TOPS) of dedicated deep neural network (DNN) compute. The neural compute engine directly interfaces with a high-throughput intelligent memory fabric to avoid any memory bottleneck when transferring data. It supports FP16 and INT8 calculations. The Myriad X also features a cluster of 16 proprietary SHAVE cores and upgraded and expanded vision accelerators.

The Myriad X is available in Intel's Neural Compute Stick 2, effectively an evaluation platform in the form of a USB thumb drive. It can be plugged into any workstation to allow AI and computer-vision applications to be up and running on the dedicated Movidius hardware very quickly.

NXP Semiconductors i.MX 8M Plus

The i.MX 8M Plus is a heterogeneous application processor featuring dedicated neural network accelerator IP from VeriSilicon (Vivante VIP8000). It offers 2.3 TOPS of acceleration for inference in endpoint devices in the consumer and industrial internet of things, enough for multiple object identification, speech recognition of 40,000 words, or even medical imaging (MobileNet v1 at 500 images per second).

In addition to the neural network processor, the i.MX 8M Plus features a quad-core Arm Cortex-A53 subsystem running at 2 GHz, plus a Cortex-M7 real-time subsystem. For vision applications, there are two image signal processors that support two high-definition cameras for stereo vision or a single 12-megapixel (MP) camera. For voice, the device includes an 800-MHz HiFi4 audio digital signal processor (DSP) for pre- and post-processing of voice data.

XMOS xcore.ai

The xcore.ai is designed to enable voice control in artificial intelligence of things (AIoT) applications. A crossover processor (with the performance of an application processor and the low-power, real-time operation of a microcontroller), this device is designed for machine-learning inference on voice signals.

It is based on XMOS's proprietary Xcore architecture, itself built on building blocks called logical cores that can be used for I/O, DSP, control functions, or AI acceleration. There are 16 of these cores on each xcore.ai chip, and designers can choose how many to allocate to each function. Mapping different functions to the logical cores in firmware allows the creation of a "virtual SoC," entirely written in software. XMOS has added vector pipeline capability to the Xcore for machine-learning workloads.

The xcore.ai supports 32-bit, 16-bit, 8-bit, and 1-bit (binarized) networks, delivering 3,200 MIPS, 51.2 GMACCs, and 1,600 MFLOPS. It has 1 Mbyte of embedded SRAM plus a low-power DDR interface for expansion.

XMOS's xcore.ai is based on a proprietary architecture and is designed specifically for AI workloads in voice-processing applications. (Image: XMOS)

AUTOMOTIVE SOC

Texas Instruments Inc. TDA4VM

Part of the Jacinto 7 series for automotive advanced driver-assistance systems (ADAS), the TDA4VM is TI's first system-on-chip (SoC) with a dedicated deep-learning accelerator on-chip. This block is based on the C7x DSP plus an in-house developed matrix multiply accelerator

NXP's i.MX 8M Plus is the company's first application processor with a dedicated neural network accelerator. It's designed for IoT applications. (Image: NXP Semiconductors)

TI's TDA4VM is intended for complex advanced driver-assistance systems that allow vehicles to perceive their environments. (Image: Texas Instruments Inc.)

(MMA), which can achieve 8 TOPS.

The SoC can handle a video stream from a front-mounted camera at up to 8 MP or a combination of four to six 3-MP cameras plus radar, LiDAR, and ultrasonic sensors. The MMA might be used to perform sensor fusion on these inputs in an automated valet parking system, for example.

The TDA4VM is designed for ADAS designs between 5 and 20 W. The device is in pre-production, but development kits are available now.

GPU

Nvidia Corp. Jetson Nano

Nvidia's well-known Jetson Nano is a small but powerful graphics processing unit (GPU) module for AI applications in endpoint devices. Built on the same Maxwell architecture as larger members of the Jetson family (AGX Xavier and TX2), the GPU on the Nano module has 128 cores and is capable of 0.5 TFLOPS, enough to run multiple neural networks on several streams of data from high-resolution image sensors, according to the company. It consumes as little as 5 W when in use. The module also features a quad-core Arm Cortex-A57 CPU.

Nvidia's Jetson Nano module houses a powerful GPU with 128 cores for AI at the edge. (Image: Nvidia Corp.)

Like other parts in Nvidia's range, the Jetson Nano uses CUDA X, Nvidia's collection of acceleration libraries for neural networks. Inexpensive Jetson Nano development kits are widely available.

CONSUMER CO-PROCESSORS

Kneron Inc. KL520

The first offering from American-Taiwanese startup Kneron is the KL520 neural network processor, designed for image processing and facial recognition in applications such as smart homes, security systems, and mobile devices. It's optimized to run convolutional neural networks (CNNs), the type commonly used in image processing today.

The KL520 can run 0.3 TOPS and consumes 0.5 W (equivalent to 0.6 TOPS/W), which the company said is sufficient for accurate facial recognition, given that the chip's MAC efficiency is high (over 90%). The chip architecture is reconfigurable and can be tailored to different CNN models. The company's complementary compiler also uses compression techniques in order to run bigger models within the chip's resources to save power and cost.

Kneron's KL520 uses a reconfigurable architecture and clever compression to run image processing in mobile and consumer devices. (Image: Kneron Inc.)

The KL520 is available now and can also be found on an accelerator card from manufacturer AAEON (the M2AI-2280-520).

Gyrfalcon Lightspeeur 5801

Designed for the consumer electronics market, Gyrfalcon's Lightspeeur 5801 offers 2.8 TOPS at 224-mW power consumption (the equivalent of 12.6 TOPS/W) with 4-ms latency. Gyrfalcon uses a processor-in-memory technique that is particularly power-efficient compared with other architectures. Power consumption can actually be traded off with clock speed by varying the clock speed between 50 and 200 MHz. Lightspeeur 5801 contains 10 MB of memory, so entire models can fit on the chip.

The part is the company's fourth production chip and is already found in LG's Q70 mid-range smartphone, where it handles inference for camera effects. A USB thumb drive development kit, the 5801 Plai Plug, is available now.

ULTRA-LOW-POWER

Eta Compute ECM3532

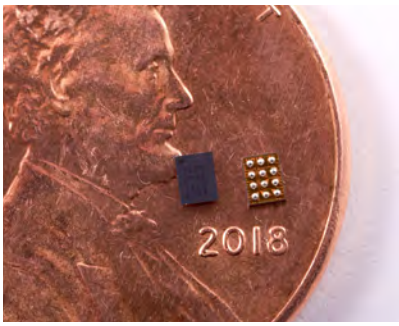
Eta Compute's first production product, the ECM3532, is designed for AI acceleration in battery-powered or energy-harvesting designs for IoT. Always-on applications in image processing and sensor fusion can be achieved with a power budget as low as 100 µW.

The chip has two cores — an Arm Cortex-M3 microcontroller core and an NXP CoolFlux DSP. The company uses a proprietary voltage and frequency scaling technique, which adjusts every clock cycle, to wring every last drop of power out of both cores. Machine-learning workloads can be processed by either core (some voice workloads, for example, are better-suited to the DSP).

Samples of the ECM3532 are available now, and mass production is expected to start in the second quarter.

Syntiant Corp. NDP100

U.S. startup Syntiant's NDP100 processor is designed for machine-learning inference on voice commands in applications in which power is tight. Its processor-in-memory–based silicon consumes less than 140 µW of active power and can run models for keyword spotting, wake word detection, speaker identification, or event classification.



Syntiant's NDP100 is designed for voice processing in ultra-low-power applications. (Image: Syntiant Corp.)

The product will be used to enable hands-free operation of consumer devices such as earbuds, hearing aids, smartwatches, and remote controls, according to Syntiant.

Development kits are available now.

GreenWaves Technologies GAP9

GAP9, the first ultra-low-power application processor from French startup GreenWaves, has a powerful compute cluster of nine RISC-V cores whose instruction set has been heavily customized to optimize the power consumed. It features bidirectional multi-channel audio interfaces and 1.6 MB of internal RAM.

GAP9 can handle neural network workloads for images, sounds, and vibration sensing in battery-powered IoT devices. GreenWaves' figures have GAP9 running MobileNet V1 on 160 × 160 images, with a channel scaling of 0.25 in just 12 ms and with a power consumption of 806 µW/frame/second. ■

Sally Ward-Foxton *is a staff correspondent at AspenCore.*

MEDICAL

Addressing the Challenges of Designing for Medical Markets

By Maurizio Di Paolo Emilio

Progress in electronics technology has made it possible to create increasingly safe, compact medical devices with the lowest degree of invasiveness, but challenges remain in meeting strict safety standards, providing wireless connectivity, and operating within highly constricted power budgets.

An electromedical device is defined as an electronic device with a part that can be applied directly to the subject or used to transfer energy to or from the patient. Applications include diagnosis, treatment, and monitoring of the patient's health conditions, as well as alleviating or even eliminating pain.

Safety, both for the patient and health-care professionals, is the first requirement with which an electromedical device must comply. The IEC 60601 family of standards sets the requirements for safety, performance, and electromagnetic compatibility of electromedical equipment. Compliance with IEC 60601 standards can be achieved only through careful evaluation of all phases of the product development cycle, starting from the selection of the components.

Whether a device is entirely new or a retrofit to which connectivity has been added, it must undergo a rigorous certification process.

Electrical insulation is directly connected to the issue involving safety. To meet the stringent requirements imposed by regulations, it is necessary to guarantee high galvanic isolation of the circuits in the equipment, using barriers or other protection solutions such as optocouplers or transformers. In addition, leakage currents must be minimized — and, better yet, eliminated.

Class III electromedical devices and active implantable devices such as pacemakers and artificial hearts present the highest level of risk. Medical electronic devices that fall under the lower risk classes (IIa, IIb, and I) of IEC 60601 include X-ray diagnostic equipment, surgical lasers, ultrasound equipment, and digital clinical thermometers.

The spread of the internet and IoT infrastructure has paved the way for devices that monitor patients' vital parameters or schedule or control the administration of drug doses. Solutions of this type have made it possible to achieve considerable savings on health-related expenses and to improve the efficiency and quality of the therapies provided. The key factors for the success of these innovative systems are connectivity and wearability.

WEARABLE DEVICES

A wearable electromedical device is an autonomous, non-invasive device capable of performing a specific medical function (such as monitoring or dosing drugs) over an extended period. "Wearable" implies that the equipment must be supported directly by the human body or by clothing.

Wearable devices, which have had exponential growth in the last decade, can be grouped into three categories: monitoring devices, rehabilitation devices, and wearable medical aids. The last category includes all devices designed to provide long-term care to patients with temporary or permanent disabilities.

Measurements and vital signs that can be monitored include cardiac electrical activity (electrocardiogram [ECG]), respiration, body temperature, heart rate, blood pressure, blood oxygenation, and body movements. Wearables that measure these vital signs make it possible to provide timely assistance if any of the parameters are abnormal, helping to save lives.

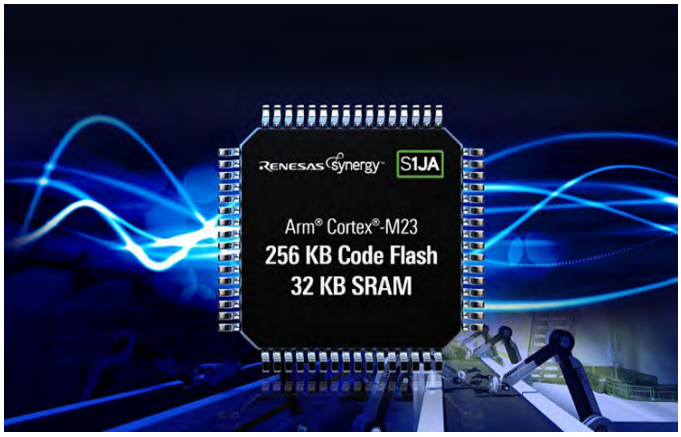


Figure 1: *The Synergy S1 MCU series* (Image: Renesas Electronics)

Wearable medical devices represent important challenges for electronic designers, however.

Power absorption must be kept to a minimum, guaranteeing long battery life to the battery-powered device. This factor influences the choice of components, orienting toward low-power solutions with the possibility of operating in low-absorption sleep mode. The mechanical constraints, related to size and weight, also affect the design activity by favoring the use of low-profile miniaturized components mounted on extremely compact PCBs.

Another important requirement for wearable electromedical devices is connectivity: Through a Bluetooth interface (typically Bluetooth Low Energy [BLE]), mobile network, or Wi-Fi, it is possible to connect to an application or gateway that can acquire sensor measurements and remotely control device operation.

CONNECTED DEVICES

Connectivity is not the exclusive domain of wearable devices but concerns a wider category of electromedical equipment. The market for connected electromedical devices is booming, but that underscores the need to increase efficiency, reduce the costs of therapeutic treatment, and improve the treatments provided to patients.

Electromedical devices relying on IoT infrastructure allow patients to be discharged more quickly or even avoid hospitalization, reducing health-care costs. Health-care devices such as insulin pumps, defibrillators, CPAPs, cardiac monitoring devices, and oxygen cylinders can now integrate remote-monitoring functionality, providing patients and caregivers with valuable real-time information without being tied to a hospital or health-care facility.

Connected medical devices must be able to connect to the cloud infrastructure of the health-care system and must therefore be equipped with a reliable antenna and network interface. A critical factor, often underestimated by companies that intend to place connected electromedical devices on the market, concerns the certification processes of wireless devices. In North America, for example, certification is separate from FDA tests and is required for all wireless devices.

Regardless of whether a device is totally new or is the retrofit of an existing device to which connectivity has been added, it is necessary to go through a rigorous certification process not only with the FCC

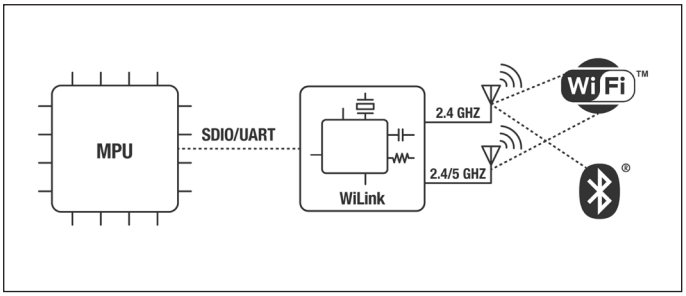


Figure 2: *Block diagram of a WiLink combo device*
(Image: Texas Instruments)

and CE but also mobile network operator, if the solution uses the cellular network.

COMPONENTS FOR MEDICAL APPS

Even if each electromedical application has different requirements, there is a common need to use microcontrollers capable of providing high performance in terms of processing times, reliability, safety, absorption, and connectivity.

The explosion of connected devices favors the introduction of cybersecurity mechanisms implemented at the chip level. The demand for ultra-low-power microcontrollers equipped with analog peripherals has been growing. Benefits include high reliability, low latency, reduced noise, and lower costs, particularly significant in devices such as glucometers, heart-rate monitors, and implantable devices.

An example of a low-power microcontroller with integrated analog peripherals is Renesas Electronics' Synergy S1 series of MCUs. The Renesas devices (**Figure 1**) include a 48-MHz Arm Cortex-M23 core with safety functionality, equipped with programmable analog peripherals for signal acquisition and conditioning.

Texas Instruments offers a wide portfolio of low-absorption solutions with analog front ends that are particularly suitable for electromedical applications. An example is its combined high-performance Wi-Fi/Bluetooth module, based on proprietary WiLink technology. This solution allows you to add Wi-Fi (2.4 and 5 GHz) and dual-mode Bluetooth 4.2 to an electromedical device.

WiLink series modules can be directly connected to a host microprocessor (MPU) via the serial interface (**Figure 2**). The WL1837MOD device from the WiLink series offers the highest level of safety available thanks to its FIPS 140-2 Level 1 certification, which allows it to be used on hospital electromedical devices.

Another fundamental component of electromedical devices is represented by the DC/DC converter, which must have high-insulation characteristics. An example is given by the converters of the REMOM series of RECOM Power, equipped with 250-VAC insulation with continuous operation and a creepage/clearance distance greater than 8 mm. REMxE converters are equipped with single or dual output, as well as protection from short-circuits, overcurrent, overvoltage, and undervoltage lockout (UVLO). Leakage current of only 1 µA makes these devices ideal for electromedical applications.

The primary constraint of thin, small, and lightweight wearable devices will always be battery life. Conventional batteries such as lithium-ion (Li-ion) batteries may be suitable for such devices.

However, there are also other challenges, notably safety and security. Your internet connection will expose data collected by your medical device for a potential breach. This problem becomes serious in the medical device industry, as the data may represent a privacy violation or an impediment to correct measurement and misanalysis. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

MEDICAL

Smartphone Sensor Gauges Blood Pressure with Clinical Accuracy

By Maurizio Di Paolo Emilio

Leman Micro Devices (LMD) has a patented sensor-based solution that can reside on a smartphone and take blood pressure measurements that are clinically accurate.

For nearly 120 years, there has been only one accurate, non-invasive means of measuring blood pressure that does not require constant recalibration: the Riva-Rocci technique, which was devised by Scipione Riva-Rocci in 1896 and employs the use of physical pressure, typically in the form of an inflatable arm cuff. LMD said it has developed a variation of the Riva-Rocci technique that makes use of sensors that LMD proposes can be easily built directly into smartphones.

LMD's technology can be used to gauge a person's diastolic and systolic blood pressure as well as measure other vital medical parameters through its V-Sensor. The product is part of a coming wave of medical devices integrated with phones — not just wearable devices, but standardized medical solutions that read vital signs with clinical precision. Instead of measuring what is easy, they will measure what is useful. Instead of seeking a medical use for what can be measured, they will make measurements that are medically meaningful.

The smartphone will be even more integral in the IoT ecosystem when it becomes a critical tool for the medical sector.

One goal is to measure blood pressure quickly and easily, without any personal calibration, in accordance with the ISO standard.

LMD's V-Sensor literally puts critical information in the hands of people who have never measured, and have no idea of, their blood pressure. Better blood pressure monitoring outside of a clinical environment could bring about qualitative progress toward more effective prevention and treatment of hypertension.

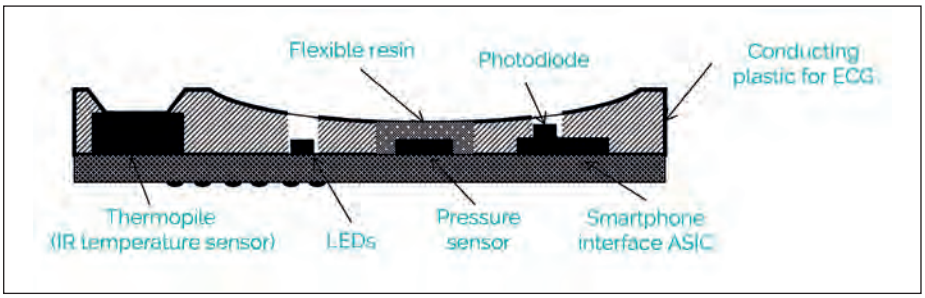
Temperature is another frequently measured parameter. A sensor on the smartphone will be able to perform this required function for medical purposes as well as in various industrial circumstances.

Pulse rate, respiration rate, and blood oxygen level are other parameters. Together, they represent a powerful diagnostic tool that can be combined with other sensors to monitor a wide range of cardiovascular conditions.

THE PHYSICS BEHIND THE MEASUREMENTS

Technologies are transforming the health industry, particularly through the continuous monitoring of people, at home, during leisure time, and even at work, recording events that

Smartphone Sensor Gauges Blood Pressure with Clinical Accuracy



Hardware setup (Image: Leman Micro Devices)

happen during the day. According to proponents of the emerging technologies, the continuous monitoring of biomedical parameters would encourage people to adopt and sustain healthy behaviors, with a view to promoting health and preventing disease.

The solution offered by LMD uses innovative sensor technology to measure blood pressure, temperature, and pulse and even perform ECG functions, recording the electrical activity of the heart. The blood pressure measurement uses the Riva-Rocci procedure.

Blood pressure is the pressure exerted by the blood, pumped from the heart, on the wall of the arteries that distribute blood in the body. The blood pressure in the arteries can be measured directly by introducing a small tube (catheter) into the artery itself and connecting it to a measuring device (pressure transducer). In practice, this happens only in special circumstances, such as during surgery.

The common blood pressure measurement, on the other hand, is carried out indirectly, using special devices that are able to assess the blood pressure from outside, in a non-invasive way. Of these devices, the most precise and commonly used so far is the one designed by Riva-Rocci just over a century ago. It consists of a rubber bracelet connected on one side to a small hand pump and on the other to a mercury column pressure gauge. Because the instrument was equipped with a glass column containing mercury and bearing a gauge indicating the distance expressed in millimeters, the unit of measurement with which the maximum and minimum blood pressure values are reported is millimeters of mercury (mmHg).

Body temperature, meanwhile, is traditionally measured using an infrared forehead scan with environmental conditions, as first used in 1835. ECG information can be obtained from a transducer placed between the hands. Finally, traditional measurements of pulse rate, blood oxygen (SpO_2), and respiratory rate are all based on established pulse oximetry transducers with established technology, in use since 1972. Respiratory rate is derived from heart rate and amplitude modulation (respiratory sinus arrhythmia).

"We have combined established, mature science with modern technology," said Chris Elliott, founder and director of Leman Micro Devices SA. "That gives users and their doctors measurements that they understand and know that they can trust with the convenience of a medical-grade device with them at all times.

"For example, we measure blood pressure using exactly the same science as a cuff around your arm, but we make it easier, smaller, and cheaper by using the fingertip and by getting the user to press harder or softer rather than having a heavy and expensive pump and cuff," Elliott added.

He explained that high blood pressure (hypertension) is a "silent killer" with no symptoms, so putting a cuffless and calibration-free blood pressure meter into every phone could save the lives of many of the 9 million people who die from it every year.

THE TECHNOLOGY

LMD's technology consists of its V-Sensor hardware module and e-Checkup software algorithm to instruct the user and analyze the data and diagnostic support on a server to store the data. All three parts will comply with all CE, FDA, and CFDA regulatory requirements for a Class IIA/Class II medical device as well as all relevant privacy regulations.

The V-Sensor has a pressure sensor embedded in a soft, flat epoxy resin that is pressed against the fingertip. LEDs and a photodiode form a conventional pulse oximeter that illuminates the tip of the finger. A thermopile measures heat radiation to estimate body temperature. The application-specific integrated circuit designed by LMD controls the LEDs; captures and digitizes data from the photodiode, pressure sensor, and thermopile; and communicates with the mobile-phone processor.

The V-Sensor can integrate into the back of a smartphone and is curved to fit a fingertip. It offers a comfortable and ergonomic user experience. To measure blood pressure, the user simply holds the phone in a natural grip, with the index finger on the V-Sensor recessed on the back of the phone and the thumb on the front screen. The e-Checkup app guides the user to play with the display by pressing harder or softer. A measurement takes approximately one minute. No external devices are required; an e-Checkup–enabled smartphone provides the complete solution.

Smartphone with V-sensor (Image: Leman Micro Devices)

The app is the interface between the phone's operating system and LMD's V-Sensor. It takes care of all the processing, including obtaining the V-Sensor features from LMD's database the first time it is used. Blood pressure accuracy is within the scope of ISO 81060-2, which for the main criterion is an average error of <5 mmHg and <8-mm standard deviation.

All measurements can be taken anywhere, anytime, in less than 60 seconds by following the on-screen instructions and without the need for special accessories or prior calibration.

The sensor in question (a tiny, 15-mm-long device to be placed on the back of a smartphone) not only allows you to measure blood pressure but, according to LMD, allows you to obtain accurate measurements of body temperature, oxygen saturation (a measure of pulmonary function), heart rate, and breathing speed.

LMD's research and development have shown that V-Sensor data can be combined with accelerometer data in the phone when held against the chest to find the timing of the heart function: left ventricular expulsion time (LVET), aortic and mitral valve timing, and many others. Further measurements may combine images from the phone's camera.

A health economist developed a Markov model using empirical data to estimate the health impact of buying a phone that measures blood pressure. He showed that, on average, a young user would gain about six months of life, while an older user would gain about three months of life. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

boards & solutions + ECE

Technical Magazine for the Embedded Systems Designer

April 2020

SPECIAL REPORT:
AI at the Edge

PAGE 38

SPECIAL REPORT: AI AT THE EDGE

A Much Smarter Edge for IoT

By Nitin Dahad

The internet of things has been touted for several years as the answer to many challenges. Connected IoT devices can improve efficiencies and productivity in industrial systems, provide valuable feedback mechanisms for connected health-care systems and wearables, and provide a vast array of capabilities to improve driver assistance and enable a path toward autonomous vehicles.

Fulfilling the promise of IoT has relied to a large extent on sensors in a network collecting data that is transmitted via a gateway to the cloud, processed, analyzed, and returned to the local system or sensor as actionable feedback.

Over the past couple of years, however, developers and systems integrators have come to realize that this approach poses issues around latency, data security, and bandwidth cost. Adding more intelligence at the edge can mitigate those challenges by speeding response times, keeping data secure and private, and minimizing data communication costs. So putting intelligence at the edge is a no-brainer, right?

It is — but there are, of course, practical limitations.

As edge intelligence has become the buzz phrase of the embedded-systems industry, its definition has gotten rather fuzzy and now spans a wider swath of the network. Some define the edge as anything not in the cloud; but even within the edge, are you at the endpoint or the edge of the gateway? These are some of the questions we asked for this special report as we looked for points of consensus among the vendors supplying the chips and systems.

We also explore how much intelligence should be added to the edge and what the practicalities and limitations are for doing so. And we look at some of the top edge-AI chip startups and make some predictions for the key trends coalescing around AI inferencing. ■

Nitin Dahad *is a staff correspondent at AspenCore.*

SPECIAL REPORT: AI AT THE EDGE

Let's Talk Edge Intelligence

By Nitin Dahad

What's the difference between edge and endpoint? It depends on your perspective.

When new industry buzzwords come up, the challenge for people like us who write about the topic is figuring out what exactly a company means, especially when it uses the term to fit its own marketing objective. The latest buzzword is actually a phrase: edge AI.

Because of the proliferation of the internet of things and the ability to add a fair amount of compute power or processing to enable intelligence within IoT devices, the "edge" can be quite wide and could mean anything from the edge of a gateway to an endpoint.

We set out in search of industry consensus on the definitions of edge and endpoint, who might want to add artificial intelligence at the edge, and how much "smartness" could be added.

We discovered that the answers depend on your perspective. But, for starters, essentially anything not in the cloud can be defined as the edge.

MANY 'EDGES,' BUT JUST ONE ENDPOINT

Probably the clearest definitions of edge and endpoint came from Wolfgang Furtner, Infineon Technologies' senior principal for concept and system engineering. "The term 'edge AI' inherits its vagueness from the term 'edge' itself," he said. "Some people call a car an edge device, and others are using the term for a small energy-harvesting sensor with low-power wireless connectivity. Edge is used in relative ways and distinguishes the more local from the more central.

Wolfgang Furtner, Infineon

"But indeed, there is a need to distinguish between the various kinds of things that you find at the edge. Sometimes, you hear terms like 'edge of the edge' or 'leaf nodes.' Edge AI can be many things, including a compute server in a car."

The key, he said, is that "endpoint AI resides at the location where the virtual world of the network hits the real world, where sensors and actuators are close."

It's all about semantics and where you draw the boundary, according to Markus Levy, director of machine-learning technologies at NXP Semiconductors. "Edge machine learning [ML] is the same as an 'endpoint' machine learning, except edge ML can also include ML that takes place in a gateway or even fog compute environment," said Levy. "Endpoint ML is typically related to distributed systems — for example, where our customers are adding intelligence even down to the sensor level. Another example is a home automation system, where there are 'satellite' devices, such as thermostat, doorbell camera, security cameras, or other

types of connected devices. While these can independently perform machine-learning functions, they might also feed into a gateway where more advanced ML processing occurs."

Chris Bergey, general manager and vice president of the infrastructure business at Arm, had a somewhat different perspective, citing the increasing levels of intelligence in both edge servers and endpoints. "Basic devices such as network bridges and switches have given way to powerful edge servers that add data-center–level hardware into the gateway between endpoint and cloud," he said. "Those powerful new edge servers making their way into 5G base stations are plenty powerful enough to perform sophisticated AI processing — not only ML inference but training, too."

Chris Bergey, Arm

How is that different from endpoint AI? Bergey explained by offering an example: "Due to their powerful internal hardware, smartphones have long been a fertile test-bed for endpoint AI. As the IoT intersects with AI advancements and the rollout of 5G, more on-device intelligence means that smaller, cost-sensitive devices can be smarter and more capable while benefiting from greater privacy and reliability due to less reliance on the cloud or internet.

"As this evolution of bringing more intelligence to endpoints continues, the boundaries of where exactly the intelligence takes place will also begin to blend from endpoint to edge, stressing the need for a heterogeneous compute infrastructure."

There are others for whom edge is everything that's not in the cloud. For example, Jeff Bier, founder of the Edge AI and Vision Alliance, said that the group defines edge AI "as any AI that is implemented — in whole or in part — outside the data center. The intelligence might be right next to the sensor — for example, in a smart camera — or a bit farther away, such as an equipment closet in a grocery store, or even farther away, such as in a cellular base station. Or [it might be in] some combination or variation of these."

Xilinx takes a similar position. "Edge AI is basically a self-sufficient intelligence deployed in the field without reliance on a data center," said Nick Ni, the company's director of product marketing for AI, software and ecosystem. "It is essential for applications that require real-time response, security — for example, not sending confidential data to the data center — and low power consumption, which is most of the devices out there. Just as humans don't rely on a data center to make countless decisions daily, edge AI will dominate the market in applications like semi-autonomous cars and smart-retail systems in coming years."

Nick Ni, Xilinx

Andrew Grant, senior director for artificial intelligence at Imagination Technologies, affirmed that idea. "It's all edge as far as we are concerned; it's the customer who decides where it goes," he said. "We'll see very much a hybrid approach, and there's absolutely a role for the cloud and data centers in this, too."

Grant added that "the speed with which the market is moving [to the edge] is phenomenal. There's been a wave of movement to the edge, but for many applications, it takes time for the silicon to materialize. We were talking to a traffic management company in China; they are moving data back and forth from the cloud. When I explained to them what we do, they immediately saw the benefit of not having to take the data to the cloud if the traffic lights themselves can determine whether a car is moving or not."

Embedded systems provider Adesto Technologies doesn't necessarily differentiate between edge and endpoint, given that the company provides devices for IoT edge servers as well as IoT edge devices. "While we don't tend to use the word 'endpoint' in our own communications, perhaps definitionally, 'endpoint' is aligned with the edge devices," said Adesto CTO Gideon Intrater. "AI in these devices would typically be some amount of local inference, with the algorithms running as a program on a processor, using a dedicated accelerator, through near-memory processing or in-memory computing."

Gideon Intrater, Adesto

He added that edge AI "is becoming a reality across just about every application. We see a great opportunity in industrial and building implementations where AI can provide benefits through predictive and preventive maintenance, quality control in manufacturing, and many other areas. The industry is just getting started, and every day that passes, we expect AI to do more for us. When our older devices without AI don't intuitively understand our needs, we often get frustrated because we have other devices that will provide intuitive capability. The end consumer doesn't know what goes into making an AI solution work; they just expect it to work."

IT'S STILL EARLY DAYS

So we are clear on the definition: You either sit in the camp that says the edge is everything that's not in the cloud or with those who clearly identify the endpoint as the meeting point of the physical world with the digital world, mostly the sensors. But the specific application will determine the point at which the intelligence might need to be added, with an increasingly blurred line between edge and endpoint and a somewhat heterogeneous compute infrastructure.

Markus Levy, NXP

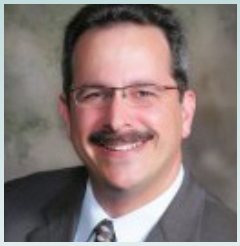
The next questions are: Who would want it, and what are the market expectations for edge AI? "This is something we're all still figuring out," said NXP's Levy. "The industry leaders are well engaged in implementing it; I can't name names, but we have a wide range of customers doing all kinds of machine learning at the edge. However, if you look at the technology adoption cycle, I still believe the majority of the industry is not even at the 'early adopter' stage, and this will really begin to unfold toward middle to late 2020.

"Customers are still comprehending the cool things that are possible with machine learning. But I typically give a few guidelines. [First,] can it save money — for example, by making a factory assembly line run faster or more efficiently by replacing headcount that was previously doing visual inspection? [Second,] can it make money — for example, by adding a cool feature to a product that makes it more useful? Maybe this is a barcode scanner that, by using machine learning, can remove wrinkles in a package that were previously making it impossible to scan accurately."

Infineon's Furtner asked essentially the same question in a different way: "What is the benefit of edge AI?"

He added, "The great thing about the edge is that we can turn its 'weaknesses' with respect to constraints into strengths. People do care about things like ease of use, functionality, privacy, security, cost, climate, or sustainable use of resources. These are all benefits we can make possible with edge AI. We are convinced that AI at the right places enhances our life and that there are many use cases for AI in endpoints. Edge AI is used for predictive maintenance and further automation or robotics, home automation, or smart farming, to name a few applications. With our work on low-power AI-enabled sensors, we make intuitive sensing more ubiquitous, spurring new applications in the home or city that can make lives easier, safer, and greener. Being non-dependent on the cloud enables fully new usage models in industry or home applications that cater for privacy and security."

Also, he said, edge AI provides the ability to drive value out of the exploding amount of IoT data in a more resource-efficient and thus sustainable way, which is paramount in times of climate change.



Jeff Bier, Edge AI and Vision Alliance

Bier of the Edge AI and Vision Alliance said application requirements would drive the need for edge AI in five key areas:

- **Bandwidth.** Even with 5G, there may not be sufficient bandwidth to send all of your raw data up to the cloud.
- **Latency.** Many applications require faster response time than you can get from the cloud.
- **Economics.** Even if a given application can technically use the cloud in terms of bandwidth and latency, it may be more economical to perform AI at the edge.
- **Reliability.** Even if a given application can technically use the cloud in terms of bandwidth and latency, the network connection to the cloud is not always reliable, and the application may need to run regardless of whether it has this connection. In such cases, edge AI is needed. An example is a face-recognition door lock; if the network connection is down, you still want your door lock to work.
- **Privacy.** Even if a given application can technically use the cloud in terms of bandwidth, latency, reliability, and economics, there may be many applications that demand local processing for privacy reasons. An example is a baby monitor or bedroom security camera.

HOW SMART SHOULD THE EDGE BE?

This might seem an obvious question, but the reality is that you need to be pragmatic, and each application is unique.

"Smart is generally not the limiting factor; the limit is memory capacity," said NXP's Levy. "In practice, memory limits the size of the machine-learning model that can be deployed, especially in the MCU domain. And to go one level deeper here, a machine-learning model for, say, a vision-based application will require more processing power and more memory. Again, processing power is more of a factor when real-time response is required.

"An example I give is a microwave oven with an internal camera to determine what kind of food was put in: A 1- or 2-second response time could be sufficient, thereby enabling the use of something like an NXP i.MX RT1050. The amount of memory will dictate the size of the model, which in turn dictates the number of food classes the machine can recognize. But what if food is inserted that isn't recognized? Now go to the gateway or the cloud to figure out what it is, then use that information to allow the smart edge device to retrain.

"To directly answer the question about how much 'smart' to include, it all boils down to tradeoffs of performance, accuracy, cost, energy. To add to this, we are also working on an application that uses auto-encoders for another form of ML: anomaly detection. In short, auto-encoders are quite efficient, and one example we implemented took only 3 Kbytes and did inferencing in 45 to 50 µs — easily the job of an MCU."

Infineon's Furtner echoed the pragmatic approach. "Edge AI is heavily constrained concerning energy consumption, space, and cost," he said. "In this case, the question is not how much smartness we should put into the edge but how much smartness we can afford in the edge. And the follow-up question would be, which of the known AI techniques can be slimmed down in a way that they are sufficiently 'tiny' to be applied in the edge?

"So certainly, power consumption limits the amount of endpoint intelligence. These endpoints are often powered by small batteries or even depend on energy harvesting. Data transmission costs a lot of energy, too."

Consider a smart sensor, Furtner said: "For local AI to function properly under these circumstances, it has to be optimized for its specific properties and behaviors. In addition, some new sensors will only become possible through the embedded AI — for instance, environmental sensors for liquids and gases. There are many reasons for endpoint AI. Intelligent data usage and reduction or fast real-time local reactions are obvious ones. Data privacy and security are others. Massive sensor raw data can be processed where it is generated, whereas intensive compute tasks remain in the cloud. Recent advances in lowest-power neural computing — for example, edge TPUs, neuromorphic technologies — shift this boundary in favor of the edge and endpoint nodes."

Imagination Technologies' Grant said, "To our mind, it's obvious to put as much intelligence as possible in the edge, and then software optimization can be used during the lifetime of the device." He likened the approach to the games console industry model: A vendor releases a new console that later is optimized with software updates over the life of the hardware.

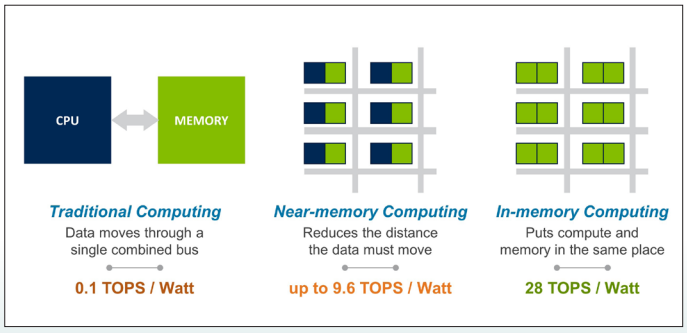
Adding a neural network accelerator to a system-on-chip (SoC) is not significant from a cost or size viewpoint, said Grant, "so the opportunities to speed up at the edge are really dramatic."

Arm's Bergey said, "As heterogeneous compute becomes ubiquitous throughout infrastructure, it is critical that we are able to identify where it makes most sense to process data, and this will vary from application to application and may even change based on time of the day. The market requires solutions that will enable the handing off of different roles to different layers of AI in order to gain the kind of overall insight that drives real business transformation. At the edge, AI is set to play a dual role. At a network level, it could be used to analyze the flow of data for network prediction and network function management, intelligently distributing that data to wherever it makes most sense at the time, whether that's the cloud or elsewhere."

Andrew Grant, Imagination Technologies

Adesto's Intrater said that "the decision of how much 'smarts' should be put at the edge is dependent on the specific application and how much latency it can handle (not much for real-time mission-critical applications), what the power envelope is (very small for battery-operated devices), and security and privacy concerns, as well as whether there is an internet connection. Even with an internet connection, you wouldn't want to send everything to the cloud for analytics, because of the bandwidth expense. The division of the smarts across the edge and the cloud is about balancing all these concerns."

He continued, "You could also do AI on a local edge server, and of course, training and analytics are often done in the cloud. Often, it is not a straightforward decision of where AI happens; the smarts are often distributed, with some happening in the cloud and some in the edge device. A typical AI system has such a split between which AI is done locally and



Adesto Technologies is exploring in-memory AI computing with RRAM technology, in which individual memory cells serve as both storage elements and compute resources. (Image: Adesto Technologies)

which is done remotely. Alexa and Siri are a good example, where there are algorithms in the device for voice/keyword recognition, and then the interactions from there take place in the cloud."

ENABLING TECHNOLOGIES

The Edge AI and Vision Alliance's Bier sees "many key enabling technologies" for edge AI. "Perhaps the most obvious is high-performance, energy-efficient, inexpensive processors that are good at running AI algorithms, but there are many others," he said. "Some of the most important are software tools, to enable efficient use of these processors, and cloud platforms, to aggregate metadata from edge devices and to manage the provisioning and maintenance of edge devices."

As you'd expect, most of the companies we spoke to provide a range of devices and intellectual property for edge AI. Infineon provides sensors, actuators, and microcontrollers, including neural network accelerators and hardware-security modules for the IoT, said Furtner. "Power efficiency, safety, and security are part of our key competencies. With our portfolio, we help to link the real with the digital world, offering secure, robust, and energy-efficient AI solutions for the edge."

Ni at Xilinx said that taking AI edge products to market is non-trivial: Engineers need to blend machine-learning technologies with conventional algorithms like sensor fusion, computer vision, and signal transformation. "Optimizing all the workloads to meet end-to-end responsiveness requires an adaptable domain-specific architecture that allows programmability in both hardware and software," he said. "Xilinx SoCs, FPGAs, and ACAPs provide such adaptable platforms to allow continuous innovation while meeting the end-to-end product requirements."

NXP's enabling technologies include hardware and software. "There are customers that use our low-end Kinetis or LPC MCUs for some smart functionality," said Levy. "It starts to get more interesting at our i.MX RT crossover processor level, where we provide integrated MCUs with Cortex-M7s running at 600 to 1,000 MHz. Our new RT600 includes an M33 and HiFi4 DSP, whereby we enable medium-performance machine learning by running in a heterogeneous mode, using the DSP to accelerate various components of a neural network.

"Moving way up the spectrum, our latest i.MX 8M Plus combines four A53s with a dedicated neural processing unit [NPU] that delivers 2.25 TOPS and two orders of magnitude more inference performance (and runs less than 3 W). This high-end NPU is critical for applications such as real-time speech recognition (i.e., NLP), gesture recognition, and live video face and object recognition."

From the software perspective, Levy said that NXP provides its eIQ machine-learning software development environment to enable open-source ML technologies across the NXP portfolio, from i.MX RT to i.MX 8 apps processors and beyond. "With eIQ, we give customers the option to deploy ML on the compute unit of their choice: CPU, GPU, DSP, or NPU. You'll even see heterogeneous implementations that run a voice application like keyword detection on the DSP, a face recognition on the GPU or CPU, and a high-performance video application on the NPU, or any combination thereof."

Arm's Bergey said, "As we move toward a world of 1 trillion IoT devices, we're facing an infrastructural and architectural challenge that's greater than ever before — and, as such, the technology we need to answer this great opportunity is constantly evolving. At Arm, our focus is on providing highly configurable, scalable solutions that meet the performance and power requirements to enable AI everywhere."

For AI at the edge, Adesto provides enabling technologies including ASICs with AI accelerators, NOR flash memory for storing the weights in an AI chip used for voice and image recognition, and smart edge servers that connect legacy and new data to cloud analytics such as IBM Watson and Microsoft Azure.

"We are also exploring in-memory AI computing with our RRAM [resistive RAM] technology, where individual memory cells serve as both storage elements and compute resources," said Intrater. "In this paradigm, the matrices of the deep neural networks become arrays of NVM [non-volatile memory] cells, and the weights of the matrices become the conductance of the NVM cells. The dot product operations are done by summing up the current that results from applying the input voltages on to the RRAM cells.

"Since there is no need to move the weights between the compute resource and the memory, this model can achieve an unrivaled combination of power efficiency and scalability."

OUR TAKE

We see a clear distinction between edge AI and endpoint AI: The endpoint is the point at which the physical world interfaces with the digital world. But the way the edge is defined is very elastic. Some vendors say that everything that is not in the data center is the edge; some define endpoints as a subset of the edge.

Ultimately, the definition is irrelevant. It comes down to the application and how much intelligence one can practically place at the endpoint or at the edge. And that involves tradeoffs among memory availability, performance needs, cost, and energy consumption. Those decisions will determine how much inferencing and analysis can be done at the edge, how many neural network accelerators are needed, whether this is part of a system-on-chip or whether it sits with a CPU, GPU, or DSP. This is not to forget new, innovative ways of looking at solving the challenge, using techniques like in-memory computing and AI.

There is broad consensus on this: How much intelligence you put in depends on your application and requires a pragmatic approach based on the available resources. ■

SPECIAL REPORT: AI AT THE EDGE

Chip Startups for AI in Edge and Endpoint Applications

By Sally Ward-Foxton

As the industry grapples with the best way to accelerate artificial-intelligence performance to keep up with requirements from cutting-edge neural networks, there are many startup companies springing up around the world with new ideas about how this is best achieved. They are attracting a lot of venture capital funding, and the result is a sector rich in not just cash but in novel ideas for computing architectures.

A data center can count as the edge, depending on where it is. The key concept of edge computing is that the data is processed in or near the same geographical location as the data is generated or gathered.

Here at EE Times Europe, we are currently tracking about 60 AI chip startups in the United States, Europe, and Asia, ranging from companies reinventing programmable logic and multicore designs to developers of entirely new architectures and others using futuristic technologies such as neuromorphic (brain-inspired) architectures and optical computing.

Here is a snapshot of 10 that we think show promise or at least have some interesting ideas. We've categorized them by where in the network their products are targeted: data centers, endpoints, or devices for the artificial intelligence of things (AIoT).

AI IN THE DATA CENTER

Yes, a data center can count as the edge — depending on where it is. The key concept of edge computing is that the data is processed in (or near) the same geographical location as the data is generated or gathered. This includes gateway or hub devices but also on-premises servers that accelerate companies' individual AI applications. Think servers that accelerate image classification for X-rays or CT scans in a hospital or medical research facility, or gateways that receive status data from the factory floor and process it on-site.

Graphcore

Based in Bristol, U.K., Graphcore hit the news when an early funding round valued the company at more than US$1 billion, making it the first Western semiconductor startup to be designated a unicorn.

The company's Intelligence Processing Unit (IPU) chip has a massively parallel architecture with more than 1,200 specialized cores, which can each run six program threads. There is also substantial on-chip memory — hundreds of megabytes of RAM — plus, importantly, 45 terabytes of memory bandwidth. That allows entire machine-learning models to be stored on the chip.

Graphcore's IPU chip is available in a Dell server for edge compute applications.

Groq

Founded in Silicon Valley by a team from Google, Groq employs 70 people and has raised US$67 million in funding to date. It officially unveiled its enormous chip, which is capable of 1,000 TOPS (1 peta-OPS), late last year at SC19.

The company's software-first approach means its compiler handles many of the control functions that would normally happen in hardware, such as execution planning. Software orchestrates all the dataflow and timing required to make sure calculations happen without stalls, making latency, performance, and power consumption entirely predictable at compile time.

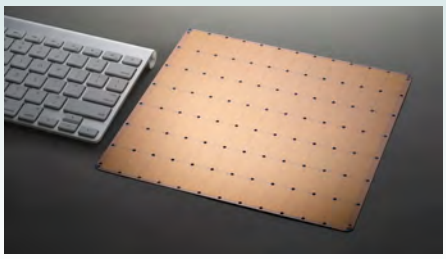
Groq is targeting data-center applications and autonomous vehicles with its tensor streaming processor (TSP) chip. The part is sampling now on a PCIe board.

Cerebras Systems

California-based Cerebras is famous for resurrecting the wafer-scale chip idea, which was abandoned in the 1980s.

The company's mammoth, 46,225-mm^2 die takes up an entire wafer. It consumes 15 kW

Graphcore's IPU chip has more than 1,200 cores. (Image: Graphcore)



The Cerebras chip takes up an entire wafer. (Image: Cerebras)

and packs 400,000 cores and 18 GB of memory onto 84 processor tiles. If the stats seem staggering, remember that one Cerebras chip is designed to replace thousands of GPUs.

Cerebras says it has solved problems that previously plagued wafer-scale designs, such as yield (it routes around defects), and that it has invented packaging that counters thermal effects. The company has raised more than US$200 million and says its rack system is running in a handful of customer data centers.

Cambricon Technologies

One of China's first AI chip companies, but by no means its last, Cambricon was founded in 2016 by two researchers from the Chinese Academy of Sciences who are brothers.

Citing the lack of agility in CPU and general-purpose GPU (GPGPU) instruction sets for the acceleration of neural networks, they developed their own instruction set architecture (ISA), a load-store architecture that integrates scalar, vector, matrix, logical, data transfer, and control instructions.

Cambricon's first product, Cambricon-1A, is used in tens of millions of smartphones and other endpoint devices such as drones and wearables. Today, second-generation chips include two parts for the cloud plus an edge compute chip, the Siyuan 220, that was designed to fill a gap in the company's portfolio. The chip offers 8-TOPS performance and consumes 10 W.

Cambricon (along with Horizon Robotics; see below) is currently one of the world's most valuable chip startups: The company has raised US$200 million thus far, giving it a market valuation of about US$2.5 billion.

AI IN THE ENDPOINT

"Endpoint" refers to devices at the end of the network, wherein the data is processed inside the same device that collected the data. This

might include everything from security cameras to consumer electronics and appliances. Of course, there is some gray area, given that some devices can be used as either gateways or endpoints (consider autonomous vehicles and smartphones).

Hailo Technologies

Founded in 2017 in Tel-Aviv, Israel, by former members of the Israel Defense Forces' elite intelligence unit, Hailo has about 60 employees and has raised US$21 million to date.

Hailo's AI co-processor, the Hailo-8, can handle 26 TOPS with a power efficiency of 2.8 TOPS/W. (Image: Hailo)

Hailo's AI co-processor, the Hailo-8, can handle 26 TOPS with a power efficiency of 2.8 TOPS/W. It targets advanced driver-assistance systems (ADAS) and autonomous driving applications. Its architecture mixes memory, control, and compute blocks, and adjacent blocks are allocated to compute each layer of a neural network by software. Minimizing data transmission on and off the chip helps save power.

Mass production of the Hailo-8 is due to begin in the first half.

Kneron

With 150 employees in San Diego and Taiwan and US$73 million in funding, Kneron was one of the first startups to get silicon on the market, achieving the goal in May 2019. The company has several customers already announced for its first-generation KL520 chip and says it made "millions of dollars" in revenue in 2019.

The KL520 is optimized for convolutional neural networks (CNNs) and can run 0.3 TOPS at 0.5 W (equivalent to 0.6 TOPS/W). That is enough for facial recognition in IP security cameras, for example, but the chip can also be found in smart door locks and doorbells.

The company started out making neural networks for facial recognition and now offers those alongside intellectual property for its neural processing unit (NPU). A second-generation chip, due this summer, will be able to accelerate both CNNs and recurrent neural networks (RNNs), the company said.

Mythic

Mythic was founded in 2012 at the University of Michigan. The company, now based in Austin, Texas, has raised US$86 million to develop its analog compute chip, which uses a processor-in-memory technique based on flash transistors for power, performance, and cost advantages over CPUs and GPUs.

Processor-in-memory is not new, but Mythic says it has figured out the tricky compensation and calibration techniques that cancel out noise and allow reliable 8-bit computation. The company plans to sell standalone chips as well as multichip processing cards.

Because the device can handle image processing on HD video at 30 frames/second, one of Mythic's key target markets is security cameras and on-premises aggregators for security camera systems.

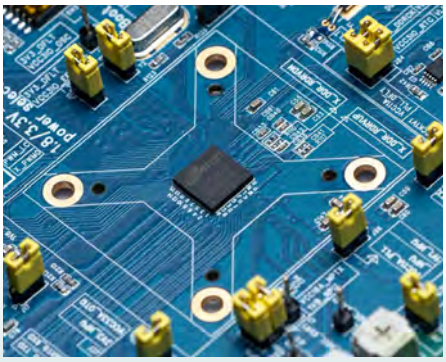
Mythic CEO Mike Henry said in an interview at CES that its chip would be sampling "shortly."

Horizon Robotics

Founded in 2015 in Beijing, Horizon Robotics had raised about US$600 million by the end of 2019, for a valuation of US$3 billion. Today, the company has more than 500 employees and holds more than 600 patents.

Horizon Robotics' Brain Processing Unit (BPU) was originally designed for computer-vision applications. The second generation of the BPU is a 64-bit multi-instruction, multi-data (MIMD) core that can handle all types of neural networks (not just convolutional networks). It uses the company's sparse neural networks to predict the movement of objects and for scene parsing. A third generation will add acceleration for decision-making algorithms and for other parts of AI outside of deep learning.

Horizon Robotics has two chip product lines: Journey, for automotive, and Sunrise, for AIoT. The first generation of Journey and Sunrise chips launched in December 2017, with a second generation based on BPU 2.0 arriving in autumn 2019.



Kneron's KL520 will be succeeded by the KL720 this summer. (Image: Kneron)

Journey 2 offers 4 TOPS at 2 W for L3/L4 autonomous driving and, combined with Horizon's own algorithms for perception, achieves 90% core utilization.

AI IN THE IOT (TINYML)

In this category, we consider chips with microcontroller levels of compute that operate in resource-constrained environments at ultra-low power. AI inference on the endpoint device in these circumstances is very attractive because it increases latency, saves bandwidth, helps privacy, and saves power associated with RF transmission of data to the cloud.

GreenWaves

A spin-out from the University of Bologna, GreenWaves (Grenoble, France) uses multiple RISC-V cores in an ultra-low-power ML application processor for battery-powered sensing devices. The company relies on its custom instruction set extensions to facilitate DSP operations and AI acceleration at minimal power consumption.

GreenWaves' second-generation product, GAP9, uses 10 cores. One is used as a fabric controller, and nine make up the compute cluster, with the controller and cluster in separate voltage and frequency domains, so they each consume power only when necessary. The GAP9 also takes advantage of a state-of-the-art fully depleted silicon-on-insulator (FD-SOI) process technology to minimize power consumption.

GreenWaves' figures have GAP9 running MobileNet V1 on 160 × 160 images with a channel scaling of 0.25 in just 12 ms, with power consumption of 806 µW/frame/second.

GAP9 samples are set to arrive in this year's first half.

Eta Compute

Eta Compute's design for AI processing in ultra-low-power IoT devices uses two cores: an Arm Cortex-M3 microcontroller core plus a DSP. Both leverage the company's clever dynamic voltage and frequency-scaling techniques to run at the lowest possible power levels, which is achieved without phase-locked loops (PLLs). The AI workload, allocated via software, can run on either core or both. Using this technique, always-on image processing and sensor fusion can be achieved with a power budget of 100 µW.

The company also optimizes neural networks for ultra-low-power applications that will run on its ECM3532 device.

Founded in 2015, Eta Compute employs 35 people in the U.S. and India and has raised US$19 million in funding to date. ■

Sally Ward-Foxton *is a staff correspondent at AspenCore.*

SPECIAL REPORT: AI AT THE EDGE

Seeing the AI Inference Market with 2020 Vision: Five Predictions

By Geoff Tate

A new class of artificial-intelligence chips is coming to market in 2020, and the entries will be optimized for inference — not for graphics, training, or digital signal processing. In fact, according to Nvidia in its recent quarterly earnings call, the AI inference market exceeded AI training spending in the data center in 2019.

What is surprising here is that very few startups introduced new products this past year, considering the number of startups funded and the number of companies that have been in business for five or more years. Flex Logix expects that to change this year. Below are our top five predictions based on what we see in the market and the many conversations we are having with customers.

1. THROUGHPUT OVER MEANINGLESS BENCHMARKS

A year ago, customers were asking about tera-operations per second (TOPS) and ResNet-50 at various batch sizes. Today, leading customers have developed models that work for their applications, and how a solution runs their model is what's important to them. The thing that matters most is throughput on megapixel images (not meaningless benchmarks), and more companies are going to figure this out in 2020.

2. INFERENCE THROUGHPUT AT LOWEST COST

In the server market, some customers will want to get more inference per PCIe slot, but the path to expanding the market will be to deliver inference throughput at lower price points. Having more TOPS doesn't necessarily correlate with higher throughput; what customers really want is high throughput per dollar. As an example, if you compare the Flex Logix InferX X1 device with a market-leading graphics processor, the GPU may offer 3× to 4× the throughput, with 10× the TOPS, but it also uses 8× the number of DRAMs. The InferX X1 architecture is a lot more resource-efficient.

3. PREDICTION ACCURACY REQUIRES HIGHER THROUGHPUT PER DOLLAR AND PER WATT

Applications are being starved for inference throughput and have been getting by with down-sampling their native megapixel images and processing only a fraction of the frames per second. In the market for systems that need 5- to 30-W chips/modules with heat sinks and no fans (now primarily served by Nvidia Xavier AGX and now NX), customers will want to get more throughput for the same power and price as the solutions they are using today. The only way to get higher prediction accuracy will be to get higher throughput per dollar and higher throughput per watt.

4. SELECTING THE RIGHT SOLUTION FOR AUTOMOTIVE

The automotive market for inference will be in the millions of dollars, but deployment takes time. Car companies and their suppliers have already selected solutions for the 2024–2025 model years and will evaluate alternatives in 2020 for the 2026–2027 model year. A year ago, all of the car companies/suppliers planned their own chips, but almost all have dropped those plans and will use merchant market solutions instead.

5. ADVANTAGE FOR SOLUTIONS WITH BF16

While 8-bit integer (INT8) offers the highest throughput per dollar and throughput per watt, winning solutions will need to have a BFloat16 (BF16) option because it allows customers to quickly ramp production. For many customers, the cost and complexity of quantizing may never be economical.

It's going to be an exciting year in the AI/inference market. There is no doubt that the availability of high-throughput inference capabilities will change the way people live, work, and play. ■

Geoff Tate *is the CEO of Flex Logix.*

IMAGE: SHUTTERSTOCK

SPECIAL REPORT: AI AT THE EDGE

Edge Intelligence Ticks Many Boxes for AI

By Dennis Goldenson

As adoption rates rise for artificial intelligence and machine learning (ML), the ability to process large amounts of data in the form of algorithms for computational purposes becomes increasingly important.

To help make the expanding use of data applications across billions of connected devices more efficient and valuable, there is growing momentum to migrate the processing from centralized third-party cloud servers to decentralized and localized processing on-device, commonly referred to as edge computing. According to SAR Insight & Consulting's latest AI/ML embedded chips database, the global number of AI-enabled devices with edge computing will grow at a compound annual growth rate of 64.2% during the 2019–2024 period.

DATA COMPUTATION AT THE EDGE, NO NETWORK NEEDED

Edge AI takes the algorithms and processes the data as close as possible to the physical system — in this case, locally on the hardware device. The advantage is that the processing of data does not require a connection. The computation of data happens near the network edge, where the data is developed, instead of in a centralized data-processing center. Determining the right balance between how much processing can and should be done on the edge will become one of the most important decisions for device, technology, and component providers.

Given the training and inferencing engines that produce deep-learning predictive models, edge processing usually requires an x86 or Arm processor from suppliers such as Intel, Qualcomm, Nvidia, and Google; an AI accelerator; and the ability to handle speeds of up to 2.5 GHz with 10 to 14 cores.

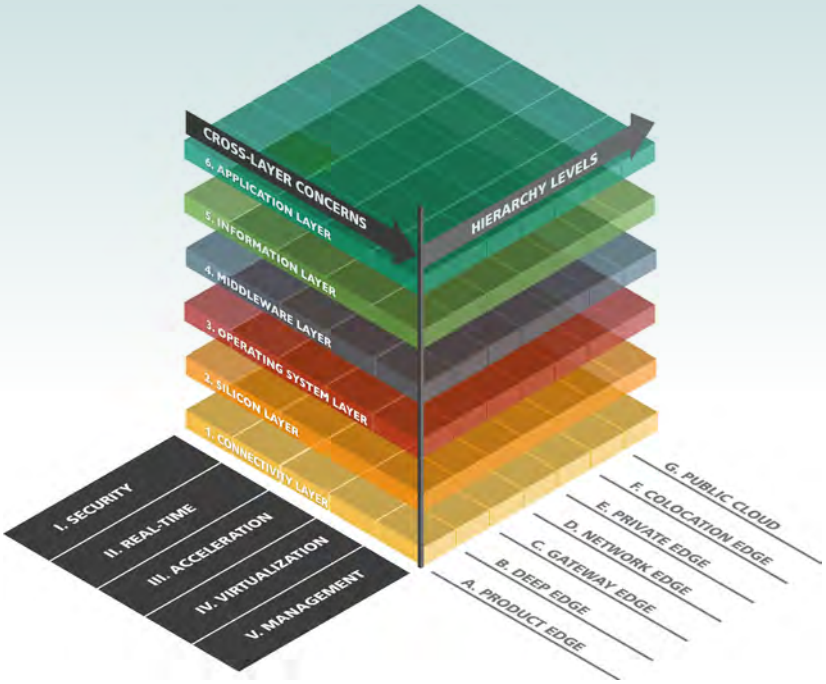
The ability of edge computing to provide immediate and reliable data for time-sensitive needs builds confidence, increases customer engagement, and, in many cases, saves lives.

REAL-TIME RESULTS FOR TIME-SENSITIVE APPS

Given the expanding markets and expanding service and application demands placed on computational data and power, there are several factors and benefits driving the growth of edge computing. Because of the shifting needs of reliable, adaptable, and contextual information, a majority of the data is migrating locally to on-device processing, resulting in improved performance and response time (in less than a few milliseconds), lower latency, higher power efficiency, improved security because data is retained on the device, and cost savings because data-center transports are minimized.

One of the biggest benefits of edge computing is the ability to secure real-time results for time-sensitive needs. In many cases, sensor data can be collected, analyzed, and communicated straightaway, without having to send the data to a time-sensitive cloud center.

Scalability across various edge devices to help speed local decision-making is fundamental. The ability to provide immediate and reliable data builds confidence, increases customer engagement, and, in many cases, saves lives. Just think of all of the industries — home security, aerospace, automotive, smart cities, health care — in which the immediate interpretation of diagnostics and equipment performance is critical.



Proposed reference architecture model for edge computing
(Image: European Edge Computing Consortium)

AI EDGE DEVELOPMENTS

Innovative organizations such as Amazon, Google, Apple, BMW, Volkswagen, Tesla, Airbus, Fraunhofer, Vodafone, Deutsche Telekom, Ericsson, and Harting are now embracing and hedging their bets for AI at the edge. A number of these companies are forming trade associations, such as the European Edge Computing Consortium (EECC), to help educate and motivate small, medium-sized, and large enterprises to drive the adoption of edge computing within manufacturing and other industrial markets.

The goals of the EECC initiative include specification of a reference architecture model for edge computing, development of reference technology stacks (EECC edge nodes), identification of gaps and recommendation of best practices by evaluating approaches within multiple scenarios, and synchronization with related initiatives/standardization organizations.

LOOKING OVER THE EDGE

The advancement of AI and machine learning is providing numerous opportunities to create smart devices that are contextually aware of their environment. The demands placed on smart machines will benefit from the growth in multi-sensory data that can compute with greater precision and performance. Edge computing provides an opportunity to turn AI data into real-time value across almost every industry. The intelligent edge is the next stage in the evolution and success of AI technology. ■

Dennis Goldenson *is director of artificial intelligence and machine learning at SAR Insight & Consulting.*

SPECIAL REPORT: AI AT THE EDGE

Putting AI into the Edge Is a No-Brainer; Here's Why

By Duncan Stewart and Jeff Loucks

In 2020, Deloitte predicts that more than 750 million edge AI chips — full chips or parts of chips that perform or accelerate machine-learning tasks on-device, rather than in a remote data center — will be sold, representing US$2.6 billion in revenue. Furthermore, the edge AI chip market will grow much more quickly than the overall chip market. By 2024, we expect unit sales of edge AI chips to exceed 1.5 billion, possibly by a great deal. That represents compound annual unit sales growth of at least 20%, more than double the longer-term forecast of 9% CAGR for the overall semiconductor industry.

These edge AI chips will likely find their way into an increasing number of consumer devices, such as high-end smartphones, tablets, smart speakers, and wearables. They will also be used in multiple enterprise markets: robots, cameras, sensors, and other devices for the internet of things.

The consumer market for edge AI chips is much larger than the enterprise market, but it is likely to grow more slowly, with a CAGR of 18% expected between 2020 and 2024. The enterprise edge AI chip market is growing much faster, with a predicted CAGR of 50% over the same time frame.

Nevertheless, this year, the consumer device market will likely represent more than 90% of the edge AI chip market, both in terms of the numbers sold and their dollar value. The vast majority of these edge AI chips will go into high-end smartphones, which account for more than 70% of all consumer edge AI chips currently in use. Indeed, not just in 2020 but for the next few years, AI chip growth will be driven principally by smartphones. We believe that more than a third of the 1.56 billion-unit smartphone market this year may contain edge AI chips.

Because of the extremely processor-intensive requirements, AI computations have almost all been performed remotely in data centers, on enterprise core appliances, or on telecom edge processors — not locally on devices. Edge AI chips are changing all that. They are physically smaller, relatively inexpensive, use much less power, and generate much less heat, making it possible to integrate them into handheld devices as well as non-consumer devices such as robots. By enabling these devices to perform processor-intensive AI computations locally, edge AI chips reduce or eliminate the need to send large amounts of data to a remote location, thereby delivering benefits in usability, speed, and data security and privacy.

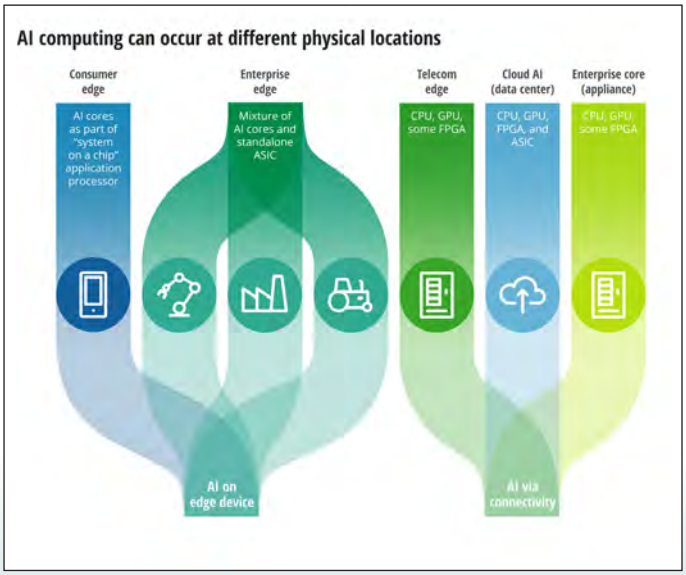


Figure 1: *Locations in which intelligence can be embedded* (Image: Deloitte Insights)

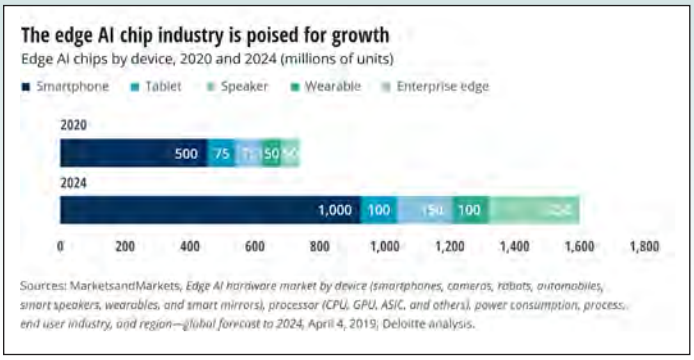


Figure 2: *The edge AI chip market* (Image: Deloitte Insights)

Keeping the processing on the device is better in terms of privacy and security; personal information that never leaves a phone cannot be intercepted or misused. And when the edge AI chip is on the phone, it can do all these things even when not connected to a network.

Of course, not all AI computations have to take place locally. For some applications — for instance, when there is simply too much data for a device's edge AI chip to handle — sending data to be processed by a remote AI array may be adequate or even preferred. In fact, most of the time, AI will be done in a hybrid fashion: some portion on the device and some in the cloud. The preferred mix in any given situation will vary depending on exactly what kind of AI processing needs to be done.

THE ECONOMICS OF EDGE AI IN SMARTPHONES

Smartphones aren't the only devices that use edge AI chips; other device categories — tablets, wearables, smart speakers — contain them as well. In the short term, these non-smartphone devices will likely have much less of an impact on edge AI chip sales than smartphones, either because the market is not growing (as for tablets) or because it is too small to make a material difference (for instance, smart speakers and wearables combined are expected to sell a mere 125 million units in 2020). Many wearables and smart speakers depend on edge AI chips, however, so penetration is already high.

Currently, only the most expensive smartphones — those in the top third of the price distribution — are likely to use edge AI chips. But putting an AI chip in a smartphone doesn't have to be price-prohibitive for the consumer.

It's possible to arrive at a fairly sound estimate of a smartphone's edge AI chip content. To date, images of phone processors in Samsung, Apple, and Huawei show the naked silicon die with all its features visible, allowing identification of which portions of the chips are used for which functions. A die shot of the chip for Samsung's Exynos 9820 shows that about 5% of the total chip area is dedicated to AI processors. Samsung's cost for the entire SoC application processor is estimated at US$70.50, which is the phone's second-most expensive component (after the display), representing about 17% of the device's total bill of materials. Assuming that the AI portion costs the same as the rest of

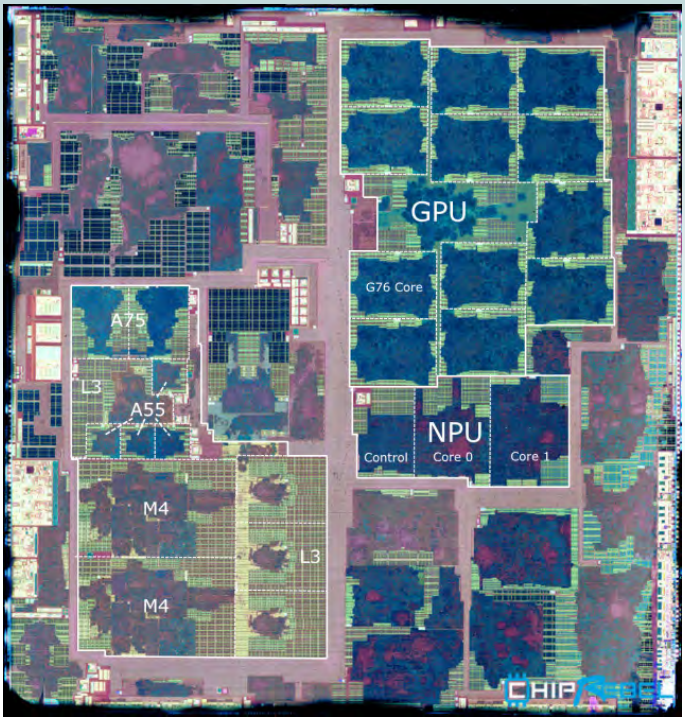


Figure 3: *A die shot of the chip for Samsung's Exynos 9820 shows that about 5% of the total chip area is dedicated to AI processors.* (Image: ChipRebel; Annotation: AnandTech)

the components on a die-area basis, the Exynos's edge AI neural processing unit (NPU) represents roughly 5% of the chip's total cost. That translates to about US$3.50 each.

Similarly, Apple's A12 Bionic chip dedicates about 7% of the die area to machine learning. At an estimated US$72 for the whole processor, that percentage suggests a cost of US$5.10 for the edge AI portion. The Huawei Kirin 970 chip, estimated to cost the manufacturer US$52.50, dedicates 2.1% of the die to the NPU, suggesting a cost of US$1.10. (Die area is not the only way to measure what percentage of a chip's total cost goes toward AI, however. According to Huawei, the Kirin 970's NPU has 150 million transistors, representing 2.7% of the chip's total of 5.5 billion transistors. That would suggest a slightly higher NPU cost of US$1.42).

Although the cited cost range is wide, it's reasonable to assume that NPUs cost an average of US$3.50 per chip. Multiplied by half a billion smartphones (not to mention tablets, speakers, and wearables), that makes for a large market, despite the low price per chip. At an average cost of US$3.50 to the manufacturer, and a probable minimum of US$1, adding a dedicated edge AI NPU to smartphone processing chips starts looking like a no-brainer. Assuming normal markup, adding US$1 to the manufacturing cost translates into only US$2 more for the end customer. That means that NPUs and their attendant benefits — a better camera, offline voice assistance, and so on — can be put into even a US$250 smartphone for less than a 1% price increase.

SOURCING AI CHIPS: IN-HOUSE OR THIRD PARTY?

Companies that manufacture smartphones and other devices vary in their approaches to obtaining edge AI chips, with the decision driven by factors such as phone model and, in some cases, geography. Some buy application processor/modem chips from third-party providers, such as Qualcomm and MediaTek, which together captured roughly 60% of the smartphone SoC market in 2018.

Both Qualcomm and MediaTek offer a range of SoCs at various prices; while not all of them include an edge AI chip, the higher-end offerings (including Qualcomm's Snapdragon 845 and 855 and MediaTek's Helio P60) usually do.

At the other end of the scale, Apple does not use external AP chips at all: It designs and uses its own SoC processors, such as the A11, A12, and A13 Bionic chips, all of which have edge AI.

Other device makers, such as Samsung and Huawei, use a hybrid strategy, buying some SoCs from merchant market silicon suppliers and using their own chips (such as Samsung's Exynos 9820 and Huawei's Kirin 970/980) for the rest.

OVER 50 AI ACCELERATOR COMPANIES VYING FOR EDGE AI IN ENTERPRISE AND INDUSTRIAL

If edge AI processors used in smartphones and other devices are so great, why not use them for enterprise applications, too? This has, in fact, already happened for some use cases, such as for some autonomous drones. Equipped with a smartphone SoC application processor, a drone is capable of performing navigation and obstacle avoidance in real time and completely on-device, with no network connection at all.

However, a chip optimized for a smartphone or tablet is not the right choice for many enterprise or industrial applications. As discussed earlier, the edge AI portion of a smartphone SoC accounts for only about 5% of the total area and about US$3.50 of the total cost and would

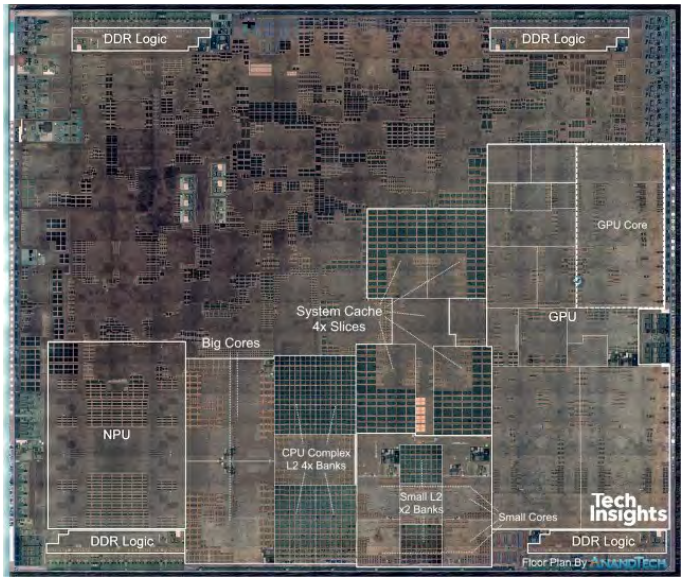


Figure 4: *Apple's A12 Bionic chip dedicates about 7% of the die area to machine learning.* (Image: TechInsights/AnandTech)

use about 95% less power than the whole SoC does. What if someone built a chip that had only the edge AI portion (along with a few other required functions, such as memory) and that cost less, used less electricity, and was smaller?

Well, they have. In all, as many as 50 different companies are said to be working on AI accelerators of various kinds. The standalone edge AI chips available in 2019 were targeted at developers, who would buy them one at a time for about US$80 each. In volumes of thousands or millions, these chips will likely cost device manufacturers much less to buy: some as little as US$1 (or possibly even less), some in the tens of dollars. We are, for now, assuming an average cost of about US$3.50, using the smartphone edge AI chip as a proxy.

Besides being relatively inexpensive, standalone edge AI processors have the advantage of being small. They are also relatively low-power, drawing between 1 and 10 W. For comparison, a data-center cluster (albeit a very powerful one) of 16 GPUs and two CPUs costs US$400,000, weighs 350 pounds, and consumes 10,000 W.

With chips such as these in the works, edge AI can open many new possibilities for enterprises, particularly with regard to IoT applications. Using edge AI chips, companies can greatly increase their ability to analyze — not just collect — data from connected devices and convert the analysis into action while avoiding the cost, complexity, and security challenges of sending huge amounts of data into the cloud. Issues that AI chips can help address include the following:

Data security and privacy

Collecting, storing, and moving data to the cloud inevitably exposes an organization to cybersecurity and privacy threats, even when companies are vigilant about data protection. This immensely important risk is becoming even more critical to address as time goes on. Regulations about personally identifiable information are emerging across jurisdictions, and consumers are becoming more cognizant of the data that enterprises collect, with 80% of them saying that they don't feel that companies are doing all they can to protect consumer privacy. Some devices, such as smart speakers, are starting to be used in settings such as hospitals, where patient privacy is regulated even more stringently.

By allowing large amounts of data to be processed locally, edge AI chips can reduce the likelihood that personal or enterprise data will be intercepted or misused. Security cameras with machine-learning processing, for instance, can reduce privacy risks by analyzing the video to determine which segments of the video are relevant and sending only those to the cloud. Machine-learning chips can also recognize a broader range of voice commands so that less audio needs to be analyzed in the cloud. More accurate speech recognition can deliver the additional bonus of helping smart speakers detect the "wake word" more accurately, thus preventing it from listening to unrelated conversation.

Low connectivity

A device must be connected for data to be processed in the cloud. In some cases, however, connecting the device is impractical. Drones are an example. Maintaining connectivity with a drone can be difficult depending on where they operate, and both the connection itself and uploading data to the cloud can reduce battery life. In New South Wales, Australia, drones with embedded machine learning patrol beaches to keep swimmers safe. They can identify swimmers who have been taken by riptides or warn swimmers of sharks and crocodiles before an attack, all without an internet connection.

(Too) big data

IoT devices can generate huge amounts of data. For example, an Airbus A-350 jet has more than 6,000 sensors and generates 2.5 terabytes of data each day it flies. Globally, security cameras create about 2,500 petabytes of data per day. Sending all this data to the cloud for storage and analysis is costly and complex. Putting machine-learning processors on the endpoints, whether sensors or cameras, can solve this problem. Cameras, for example, could be equipped with vision-processing units (VPUs) — low-power SoC processors specialized for analyzing or pre-processing digital images. With edge AI chips embedded, a device can analyze data in real time, transmit only what is relevant for further analysis in the cloud, and "forget" the rest, reducing the cost of storage and bandwidth.

Power constraints

Low-power machine-learning chips can allow even devices with small batteries to perform AI computations without undue power drain. For instance, Arm chips are being embedded in respiratory inhalers to analyze data, such as inhalation lung capacity and the flow of medicine into the lungs. The AI analysis is performed on the inhaler, and the results are then sent to a smartphone app, helping health-care professionals to develop personalized care for asthma patients. In addition to the low-power edge AI NPUs currently available, companies are working to develop "tiny machine learning": deep learning on devices as small as microcontroller units. Google, for instance, is developing a version of TensorFlow Lite that can enable microcontrollers to analyze data, condensing what needs to be sent off-chip into a few bytes.

Low-latency requirements

Whether over a wired or wireless network, performing AI computations at a remote data center means a round-trip latency of at least 1 to 2 ms in the best case and tens or even hundreds of milliseconds in the worst case. Performing AI on-device using an edge AI chip would reduce that to nanoseconds — critical for applications in which the device must collect, process, and act upon data virtually instantaneously. Autonomous vehicles, for instance, must collect and process huge amounts of data from computer-vision systems to identify objects, as well as from the sensors that control the vehicle's functions. They must then convert this data into decisions immediately — when to turn, brake, or accelerate — in order to operate safely. To do this, autonomous vehicles must process much of the data they collect in the vehicle itself. Low latency is also important for robots, and it will become more so as robots emerge from factory settings to work alongside people.

THE BOTTOM LINE: EDGE AI WILL BE VITAL FOR DATA-HEAVY APPS

The spread of edge AI chips will likely drive significant changes for consumers and enterprises alike. For consumers, edge AI chips can make possible a plethora of features — from unlocking their phone to having a conversation with its voice assistant or taking mind-blowing photos under extremely difficult conditions — and without the need for an internet connection.

But in the long term, edge AI chips' greater impact may come from their use in the enterprise, where they can enable companies to take their IoT applications to a whole new level. Smart machines powered by AI chips could help expand existing markets, threaten incumbents, and shift how profits are divided in industries such as manufacturing, construction, logistics, agriculture, and energy. The ability to collect, interpret, and immediately act on vast amounts of data is critical for many of the data-heavy applications that futurists predict will become widespread: video monitoring, virtual reality, autonomous drones and vehicles, and more.

That future, in large part, depends on what edge AI chips make possible: bringing the intelligence to the device. ■

Duncan Stewart *and* **Jeff Loucks** *are with Deloitte's Center for Technology, Media and Telecommunications. This article is based on an article originally published by Deloitte for its "TMT Predictions 2020" report.*

OPINION

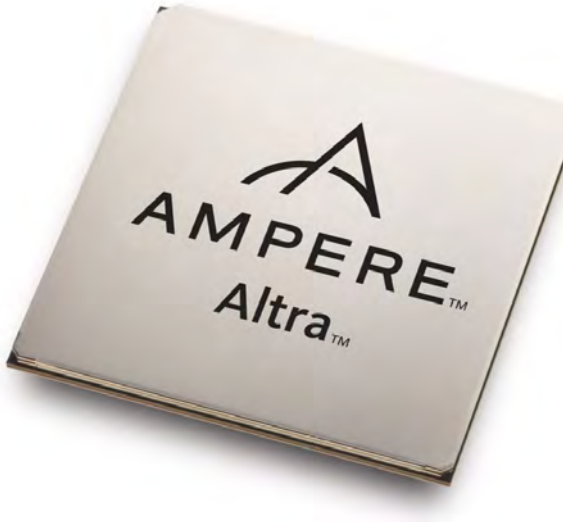
Arm in the Data Center: Is This Finally the Year?

By Sally Ward-Foxton

A recent high-profile product launch has pushed the concept of using Arm-based CPUs in the data center back into the limelight. Ampere, a Silicon Valley startup that focuses exclusively on Arm-based CPUs for the data center, launched its Altra CPU. Built on up to 80 Arm Neoverse N1 cores, Altra consumes 210 W at peak workload.

Ampere designed the Altra specifically for hyperscalers that need CPUs for their cloud offerings (an earlier Ampere CPU, eMAG, is optimized for enterprise data-center applications, which have slightly different requirements). Ampere says the Altra delivers on power efficiency (performance per watt) as well as watts per core, allowing for denser racks and maximizing performance per rack.

"Ampere's Altra server processor news is more than just a milestone for Arm in the data center; it is a turning point for the industry in terms of what is possible in data-center computing," said Chris Bergey, senior vice president and general manager for Arm's infrastructure business, in a blog post. "Today's announcement demonstrates the power of building the right compute for the right applications. Ampere took the N1 platform and integrated its own innovations to design an SoC uniquely built for applications across hyperscale cloud and edge markets while bringing it to market quickly to address evolving compute requirements."



Ampere's Altra chip, based on 80 Arm 64-bit CPU cores, is designed for performance, scalability, and power efficiency in cloud applications.
(Image: Ampere)

BUMPY JOURNEY

Let's not forget that the road to Arm in the data center has been peppered with potholes thus far.

An early startup in the space, Calxeda, was working on a 120 quad-core Cortex-A9 design (480 cores total) but closed its doors at the end of 2013 amid criticism of its 32-bit architecture, which some said was unsuitable for servers. A partnership with Hewlett-Packard for enterprise servers failed to save the company.

Fast-forward to 2016. Broadcom had its own project in the works: a 64-bit multicore Arm architecture, codenamed Vulcan, for networking, storage, communications, big data, and security applications. But the effort was a casualty of Broadcom's acquisition that year by Avago.

Arm-based CPUs are finding favor at Amazon, but the rest of the industry has yet to follow suit. Will the need for heterogeneous computing and power efficiency prove undeniable in 2020?

Then, at the end of 2018, Qualcomm abruptly shut down its Centriq division, which produced its line of Arm-based server chips.

Among those false starts, however, there is one notable success story: Amazon. The hyperscaler's 2015 acquisition of system-on-chip maker Annapurna Labs resulted in the Graviton chip — the most famous success story thus far for Arm in the data center.

Given that Amazon is its own customer for this chip, it may not be the best indicator of the market for Arm in the data center. But the company has said that developing its own chips optimized its precarious price/performance balance. Specifically, according to Amazon, Graviton A1 instances reduced customer costs by 45% for scale-out workloads such as micro-services and web servers compared with general-purpose Intel Xeon EC2 instances.

Amazon renewed its commitment to Arm architectures with the launch of the Graviton 2 in December 2019. This chip boosted performance to 7× that of its predecessor, with 4× the number of cores. Significantly, Amazon put the Graviton 2 into its Elastic Compute Cloud (EC2) instances, which are intended for high-intensity workloads, such as high-performance computing and application servers; previously, those instances were based on Intel Xeon. The shift signaled Amazon's belief in Arm's vision of diverse CPU architectures (heterogeneous computing) for even the most intense workloads in the cloud.

GAINING MOMENTUM

Amazon's vote of confidence with the launch of Graviton 2 in December, plus the Ampere launch this month, has given Arm's cause a bit of momentum. So will 2020 be the year that Arm finally makes serious inroads into the data center?

Certainly, the timing is right from a technology perspective; power efficiency demands have never been stronger, for example, and may dictate a different approach than scaling up existing CPUs. The need for heterogeneous compute to process workloads such as AI also supports Arm's case; we are seeing specialized accelerators such as GPUs, FPGAs, and even some ASICs getting more traction.

Arm's success in this market, however, depends on more than its customers' ability to meet hyperscalers' stringent demands for performance, power efficiency, and price. It also depends on whether those customers can grab any market share from Intel's dominant Xeon platform. And Arm isn't the only other option here: AMD and Nvidia are also in the fray.

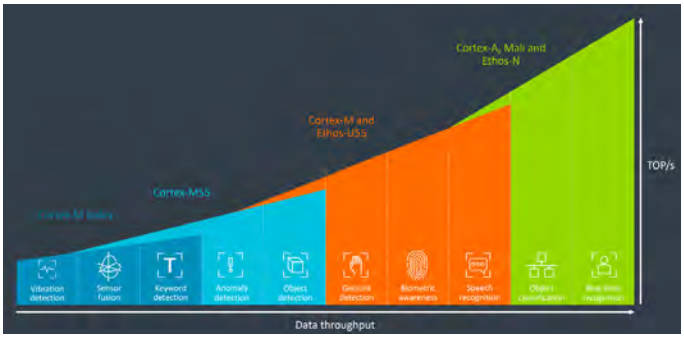
So while we may be seeing a few rays of hope, it's not quite the light at the end of the tunnel for Arm's infrastructure business — at least not yet. ■

Sally Ward-Foxton *is a staff correspondent at AspenCore.*

MICROCONTROLLERS

Adapting the Microcontroller for AI in the Endpoint

By Sally Ward-Foxton



Used in tandem, Arm's Cortex-M55 and Ethos-U55 have enough processing power for applications such as gesture recognition, biometrics, and speech recognition. (Image: Arm)

What do you get when you cross AI with the IoT? The artificial intelligence of things (AIoT) is the simple answer, but you also get a huge new application area for microcontrollers, enabled by advances in neural network techniques that mean machine learning is no longer limited to the world of supercomputers. These days, smartphone application processors can (and do) perform AI inference for image processing, recommendation engines, and other complex features.

Bringing this kind of capability to the humble microcontroller represents a huge opportunity. Imagine a hearing aid that can use AI to filter background noise from conversations, smart-home appliances that can recognize the user's face and switch to their personalized settings, and AI-enabled sensor nodes that can run for years on the tiniest of batteries. Processing the data at the endpoint offers latency, security, and privacy advantages that can't be ignored.

However, achieving meaningful machine learning with microcontroller-level devices is not an easy task. Memory, a key criterion for AI calculations, is often severely limited, for example. But data science is advancing quickly to reduce model size, and device and IP vendors are responding by developing tools and incorporating features tailored for the demands of modern machine learning.

TINYML TAKES OFF

As a sign of this sector's rapid growth, the TinyML Summit, a new industry event held in February in Silicon Valley, is going from strength to strength. The first summit, held last year, had 11 sponsoring companies; this year's event had 27, and slots sold out much earlier, according to the organizers. Attendance at TinyML's global monthly meet-ups for designers has grown dramatically, organizers said.

"We see a new world with trillions of intelligent devices enabled by TinyML technologies that sense, analyze, and autonomously act together to create a healthier and more sustainable environment for all," said Qualcomm Senior Director Evgeni Gousev, co-chair of the TinyML Committee, in his opening remarks at last month's conference.

Gousev attributed this growth to the development of more energy-efficient hardware and algorithms, combined with more mature software tools. Corporate and venture-capital investment is increasing, as are startup and M&A activity, he noted.

Today, the TinyML Committee believes that the tech has been validated and that initial products using machine learning in microcontrollers should hit the market in two to three years. "Killer apps" are thought to be three to five years away.

A big part of the tech validation came last spring when Google demonstrated a version of its TensorFlow framework for microcontrollers for the first time. TensorFlow Lite for Microcontrollers is designed to run on devices with only kilobytes of memory (the core runtime fits in 16 KB on an Arm Cortex-M3; with enough operators to run a speech keyword-detection model, it takes up a total of 22 KB). It supports inference but not training.

BIG PLAYERS

The big microcontroller makers, of course, are watching developments in the TinyML community with interest. As research enables neural network models to get smaller, the opportunities get bigger.

Most have some kind of support for machine-learning applications. For example, STMicroelectronics has an extension pack, STM32Cube.AI, that enables mapping and running neural networks on its STM32 family of Arm Cortex-M–based microcontrollers.

Renesas Electronics' e-AI development environment allows AI inference to be implemented on microcontrollers. It effectively translates the model into a form that is usable in the company's e2 studio, compatible with C/C++ projects.

NXP Semiconductors said it has customers using its lower-end Kinetis and LPC MCUs for machine-learning applications. The company is embracing AI with hardware and software solutions, albeit primarily oriented around its bigger application processors and crossover processors (between application processors and microcontrollers).

STRONG ARM-ED

Most of the established companies in the microcontroller space have one thing in common: Arm. The embedded-processor–core giant dominates the microcontroller market with its Cortex-M series. The company recently announced the brand new Cortex-M55 core, which is designed specifically for machine-learning applications, especially when used in combination with Arm's Ethos-U55 AI accelerator. Both are designed for resource-constrained environments.

But how can startups and smaller companies seek to compete with the big players in this market?

"Not by building Arm-based SoCs, because [the dominant players] do that really well," laughed XMOS CEO Mark Lippett. "The only way to compete against those guys is by having an architectural edge … [that means] the intrinsic capabilities of the Xcore in terms of performance, but also the flexibility."

XMOS's Xcore.ai, its newly released crossover processor for voice interfaces, will not compete directly with microcontrollers, but the sentiment still holds true. Any company making an Arm-based SoC to compete with the big guys better have something pretty special in its secret sauce.

SCALING VOLTAGE AND FREQUENCY

Startup Eta Compute released its much-anticipated ultra-low-power device during the TinyML show. The ECM3532 can be used for machine learning in always-on image-processing and sensor-fusion applications with a power budget of 100 µW. The chip uses an Arm Cortex-M3 core plus an NXP DSP core — either or both of which can be used for ML workloads. The company's secret sauce has several ingredients, but the way it scales both clock frequency and voltage on a continuous basis, for both cores, is key. The approach saves a lot of power, particularly because it's achieved without a phase-locked loop (PLL).

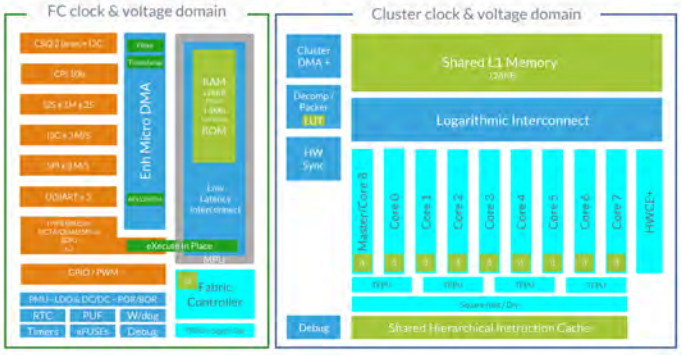
With viable competitors to Arm now out there, including the up-and-coming instruction-set architecture offered by the RISC-V foundation, why did Eta Compute choose to use an Arm core for ultra-low-power machine-learning acceleration?

"The simple answer is that the ecosystem for Arm is just so well-developed," Eta Compute CEO Ted Tewksbury told EE Times Europe. "It's just much easier to go to production [with Arm] than it is with RISC-V right now. That situation could change in the future … RISC-V has its own set of advantages; certainly, it's good for the Chinese market. But we're looking primarily at domestic and European markets right now with the ecosystem for [our device]."

Tewksbury noted that the major challenge facing the AIoT is the breadth and diversity of the applications. The market is rather fragmented, with many relatively niche applications commanding only low volumes. Altogether, however, this sector potentially extends to billions of devices.

"The challenge for developers is that they cannot afford to invest the time and the money in developing customized solutions for each one of those use cases," Tewksbury said. "That's where flexibility and ease of use become absolutely paramount. And that's another reason why we chose Arm — because the ecosystem is there, the tools are there, and it's easy for customers to develop products quickly and get them to market quickly without a lot of customization."

After keeping its ISA under lock and key for decades, Arm finally announced in October that it would allow customers to build their own custom instructions for handling specialist workloads such as machine learning.



The architecture of GreenWaves' GAP9 ultra-low-power AI chip now uses 10 RISC-V cores. (Image: GreenWaves)

That capability, in the right hands, may also offer the opportunity to reduce power consumption.

Eta Compute can't take advantage of it just yet, because the new policy does not apply retroactively to existing Arm cores, so it is not applicable to the M3 core that Eta is using. But could Tewksbury see Eta Compute using Arm custom instructions in future product generations to cut power consumption further? "Absolutely, yes," he said.

ALTERNATIVE ISAs

RISC-V has been getting a lot of attention this year. The open-source ISA allows the design of processors without a license fee, whereas designs based on the RISC-V ISA can be protected as with any other type of IP. Designers can pick and choose which extensions to add, including their own customized extensions.

French startup GreenWaves is one of several companies using RISC-V cores to target the ultra-low–power machine-learning space. Its devices, GAP8 and GAP9, use eight- and nine-core compute clusters, respectively. Each device also has an additional core that handles control functions.

Martin Croome, vice president of business development at GreenWaves, explained to EE Times Europe why the company uses RISC-V cores.

"The first reason is RISC-V gives us the ability to customize the cores at the instruction-set level, which we use heavily," said Croome, adding that the custom extensions are used to reduce the power of both machine-learning and signal-processing workloads. "When the company was formed, if you wanted to do that with any other processor architecture, it was either impossible or it was going to cost you a fortune. And the fortune it was going to cost you was essentially your investor's money going to a different company, and that is very difficult to justify."

GreenWaves' custom extensions alone give its cores a 3.6× improvement in energy consumption over unmodified RISC-V cores. But Croome also said that RISC-V has fundamental technical benefits that are simply due to its being new.

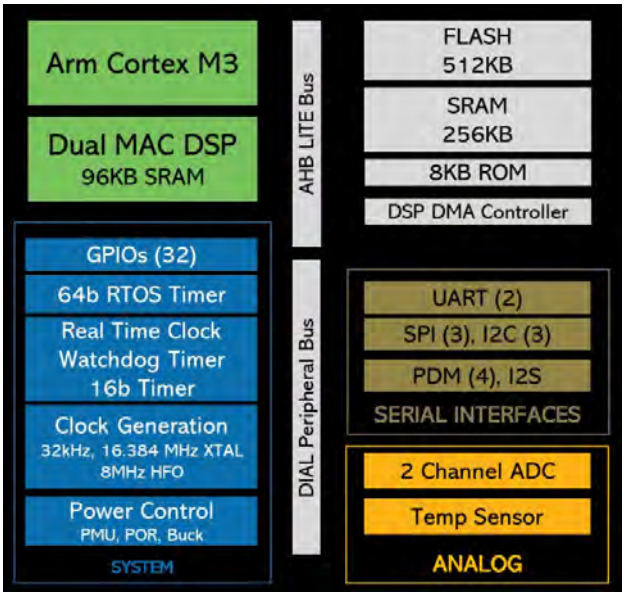
"It's a very clean, modern instruction set; it doesn't have any baggage," he said. "So from an implementation perspective, the RISC-V core is actually a simpler structure, and simple means less power."

Croome also cited control as an important factor. The GAP8 device has eight cores in its compute cluster, and GreenWaves needs very fine, detailed control over the core execution to allow maximum power efficiency. RISC-V enables that, he said.

"In the end, if we could have done all of that with Arm, we would have done all of that with Arm; it would have been a much more logical choice … because no one ever got fired for buying Arm," he joked. "The software tools are there to a level of maturity which is far higher than RISC-V … but, that said, there's now so much focus on RISC-V that those tools are increasing in maturity very fast."

In summary, while some see Arm's hold on the microprocessor market as weakening, in part because of increased competition from RISC-V, the company is responding by allowing some customized extensions and developing new cores designed for machine learning from the outset.

In fact, there are both Arm and non-Arm devices coming to the market for ultra-low-power machine-learning applications. As the TinyML community continues to work on reducing neural network model size and developing dedicated frameworks and tools, this sector will blossom into a healthy application area that will support a variety of device types. ■



Eta Compute's ECM3532 uses an Arm Cortex-M3 core plus an NXP CoolFlux DSP core. The machine-learning workload can be handled by either or both cores. (Image: Eta Compute)

Sally Ward-Foxton *is a staff correspondent at AspenCore.*

MICROCONTROLLERS

Added Arm Core Yields a More Flexible iMOTION Controller

By Maurizio Di Paolo Emilio

Manufacturers increasingly tend to consider the actual motor control unit as a basic functional block rather than a diversifying feature of the final product; this is particularly apparent for auxiliary systems such as water drainage pumps or fans normally found in domestic appliances. Infineon Technologies' new IMC300 combines the iMOTION Motion Control Engine (MCE) motor with an additional microcontroller based on the Arm Cortex-M0 core. The IMC300 portfolio targets variable-speed drives that require high application flexibility.

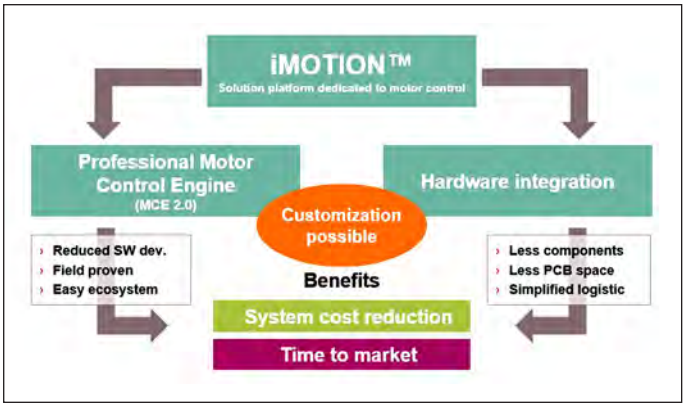
Controlling a motor's speed and direction presupposes the mode of operation of the motor in use and requires different techniques and circuits depending on the type of motor and the different application requirements. The purpose of a motor controller is to be able to act manually or automatically on the electric motor (start-stop, advance-inversion, speed, torsion, and protection against voltage overloads). Electric-motor control requires electronic circuits, which, until a few years ago, were made with discrete components because of the voltages and currents involved. Engine control today is at the forefront of R&D activities to achieve efficient microelectronic solutions on two levels: the computational software and power electronics.

iMOTION TECHNOLOGY

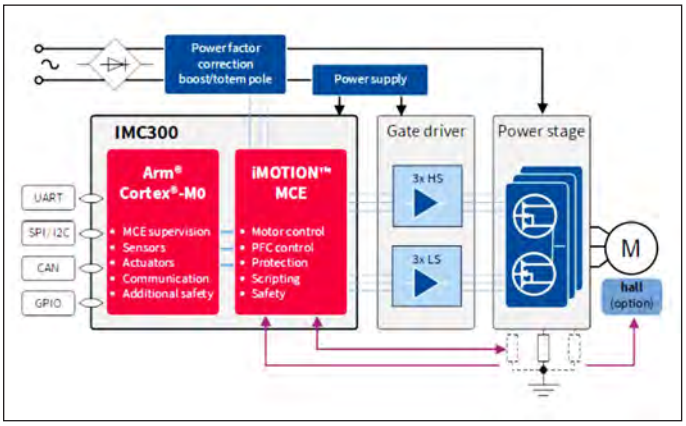
iMOTION identifies a family of products that integrate a ready and certified motor control algorithm with the hardware for productive use, with power factor correction (PFC) available as an option. In power electronics solutions for motor control, it is necessary to ensure flexibility and adaptability to the customer's target application, which often goes well beyond the configuration of the motor under consideration.

iMOTION ICs integrate all the control and analog interface functions required for sensor-field–oriented control (FOC) motor control using DC shunt current measurements or DC connection. In addition, they feature Infineon's patented algorithm, which eliminates software coding from the motor control algorithm development process. iMOTION solutions fit into home appliances and small drives such as fans. The solutions combine controller, SmartDrive, and SmartIPM devices.

The SmartDrive device is integrated with the gate driver (MOSFETs or IGBTs can be used). The highest integration level is achieved in the SmartIPM family, which combines the MCE with the gate driver and a three-phase full bridge, resulting in a complete inverter system in one small PQFN package.



iMOTION drives integration in motor control. (Image: Infineon Technologies)



IMC300 circuit application (Image: Infineon Technologies)

Targeting the regulation of variable-speed motor control systems, the ICs in Infineon's iMOTION IMC300 family integrate an additional, user-programmable microcontroller. The dual-core controller integrates the required hardware, software, and user program to regulate a permanent-magnet synchronous motor (PMSM). It enables a flexible motor control system at the lowest system and development cost.

Protection features of the IMC300 include under/overvoltage, overcurrent, overtemperature, motor gate kill, rotor lock, fault reporting, minimum dead time, and shoot-through. Control features include sinusoidal FOC, sensored/sensorless, three- or two-phase modulation, field weakening, zero vector braking, integrated PFC, and integrated bootstrap. The IMC300 and its predecessor, the IMC100, share the same implementation of MCE 2.0, providing a ready-to-use solution with PFC control, multiple protection features, and a scripting engine.

By applying MCE for motor control, customers can focus on the system application, which works completely independently on the built-in Arm microcontroller. That MCU offers a flexible set of peripherals and can serve a multitude of purposes, such as system functions, specific communication, or drive monitoring.

"The main idea behind this to make it easy for the customer to turn the motor," said Ingo Skuras, product marketing manager for iMOTION at Infineon Technologies. "We've seen a trend for some years now toward electronic control of most motors — not just on or off, but variable-speed drives, as they are called. And we do have solutions for that, which basically the customer can use without any programming: Just configure them to the respective motor uses, and then it's running. Typically, it takes half an hour to an hour to [accomplish this]."

The MCE integrates all the necessary hardware and software components, as well as all the necessary protection functions, resulting in a reduced BOM. It undergoes continuous improvements; typically, two versions are released per year.

A motor control algorithm that transfers more intelligence to peripherals such as A/D converters or timers reduces the load on the CPU core. With solutions that include small controllers integrating modern 32-bit architectures, this freed-up CPU capacity is made available to the application designer.

The IMC300 devices are pre-certified for applications requiring functional safety according to UL/IEC 60730 (Class B). ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

INTERNET OF THINGS

Thermal and Vibration Energy Powers IoT Devices

By Maurizio Di Paolo Emilio

Achieving so-called zero-power sensors will require harvesting energy from sources in the environment. After narrowing down one's options to available sources, the next criteria will be the amount of energy available and the amount of energy needed. Solar and wind harvesting can provide a solid foundation for high-power solutions. Heat is often readily available as waste by-product from engines, machines, and other sources. Thermal-gradient harvesting is the process of capturing environmental heat and putting it to use. And among the many ways to tap environmental phenomena for energy, the use of piezoelectric devices to convert vibrations into electrical energy seems particularly effective, with the ability to produce hundreds of microwatts (μW/cm^2), depending on size and construction.

THE THERMAL GRADIENT

Energy harvesting through temperature gradients is done using pyroelectric and thermoelectric solutions. The use of pyroelectrics is limited because it requires a variable temperature input, whereas other approaches can provide nonstop operation for hundreds of thousands of hours but at low efficiency. Thermoelectric solutions are enabled by Peltier cells.

"Examples of thermoelectric materials are bismuth telluride, lead telluride, cobalt triantimonide, and silicon germanium, [all of] which can provide good performance," said Alfred Piggott, founder and CTO of Applied Thermoelectric Solutions. "Using these materials, a thermoelectric generator can achieve up to 9% to 11% efficiency in an ideal application with a properly designed thermoelectric generator. Which material is best depends on many considerations, but mainly, the decision is based on the application, the budget, and the design of the thermoelectric generator."

Ideal thermoelectric materials should have low thermal conductivity, high electrical conductivity, and a high Seebeck coefficient. The thermoelectric effect leveraged for energy harvesting is attributed to Thomas Johann Seebeck. In a thermoelectric device, voltage is produced when the different temperatures are combined. Likewise, a temperature difference occurs when voltage is applied. The ability of a material or device to generate voltage per unit temperature is called the Seebeck effect.

The material usually used to create the p and n regions (bismuth telluride, or Bi_2Te_3) enables output voltages of 0.2 mV/K per cell, while higher values are obtained if the thermoelectric converter uses multiple p and n pairs (20 mV using 10 cells at $\Delta T = 10K$). The equivalent model of the source is represented by a Thévenin generator with an RT output resistor, and the maximum power that can be supplied to the load is obtained by resistive impedance adaptation $R_{load} = RT$.

A temperature difference between two points results in a flow of thermal energy from the highest temperature point to the lowest temperature point. Heat will flow until thermal equilibrium is reached and can be used to collect reusable energy. The process of extracting energy from the heat exchange is governed by the laws of thermodynamics.

Jean Charles Athanase Peltier discovered that by passing an electric current through the intersection of two conductors, heating or cooling would occur. The direction of the flow determines the direction of the temperature change, either upward or downward. The heat produced or absorbed is relative to the electric current, and the proportionality constant is called the Peltier coefficient.

THE PIEZO EFFECT

Mechanical vibration is another method to provide a sufficient energy solution for electronic systems. Oscillations of the piezoelectric transducer through the use of special masses and special systems that allow movement have been widely used in energy-harvesting applications.

Piezoelectric converters exploit the direct piezoelectric effect, i.e., the property of some crystals to generate a potential difference when subjected to mechanical strain. This effect occurs at the nanoscale and is reversible. Recently, polymeric plastic matrix piezoelectric materials (such as polyvinylidene difluoride, or PVDF) have been developed, and efforts are under way to find new materials and develop more advanced manufacturing processes.

The piezoelectric effect converts kinetic energy in the form of vibration or shock into electrical energy. Piezoelectric generators (energy harvesters) offer a robust and reliable solution by converting the vibrational energy normally wasted in the environment into usable electricity. They are ideal for applications that need to charge a battery, power a supercapacitor, or directly power remote sensor systems (**Figure 1**).

Figure 1: *The S234-H5FR-1803XB piezo crystal converts vibrations into electrical energy.* (Image: Piezo.com)

The total performance of the system depends on many factors such as the input vibrations, the geometry and material of the transducer, the mass that causes the vibrations, and the electronic interface. For this reason, even during the early design phases, a rapid and reliable quantitative estimate of the transducer and circuit junction behavior is strongly desired to optimize the system as a whole.

The analysis of the piezoelectric effect can be schematized with the circuit shown in **Figure 2**. The inductor L_M represents the equivalent inertial mass, the capacity C_M refers to the elasticity of the transducer, and the resistor R_M represents the mechanical losses. The mechanical part is powered by the force generator F_{IN}, opposite to the feedback force generator α-V_p, which is controlled by the voltage that develops on the output of the device on the capacity C_p (inverse effect of the piezoelectric). At the same time, the mechanical speed ż produces a current βż that supplies both capacitive outputs (direct effect of the piezoelectric) and other possible electrical loads connected to the transducer. Therefore, model identification involves the independent parameters L_M, C_M, R_M, C_p, α, and β (**Figure 1**). α and β are thermal coefficients related to the system.

POWER MANAGEMENT IC

Temperature differences can be used to create electricity. Waste heat from solar thermal and geothermal systems and even discharge flows from household appliances can be harvested.

Suppose we use battery-powered wireless IoT devices that operate in an environment that has thermal gradients generated by a human body, an oven, and a motor. Without energy harvesting, the batteries of such devices need to be replaced because they discharge energy. This generates operating costs. Depending on the available temperature gradients, thermoelectric generators (TEGs) can generate from 20 μW/ cm² to 10 mW/cm².

TEGs and piezo transducers combined with PMICs can charge batteries in IoT gear.

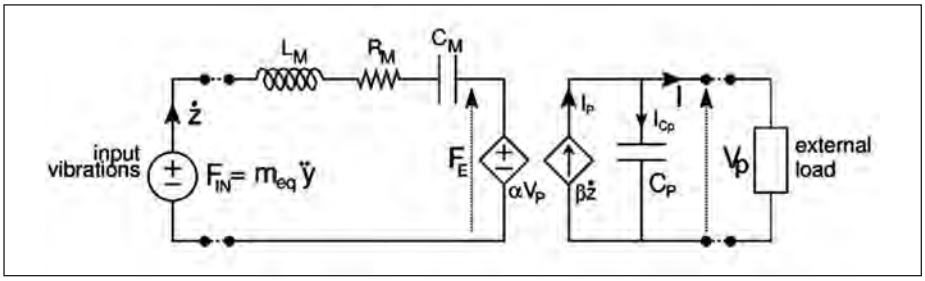


Figure 2: *Equivalent circuit of piezoelectric effect* (Image: "A piezoelectric vibration based generator for wireless electronics." Smart Materials and Structures 13 (2004) 1131–1142)

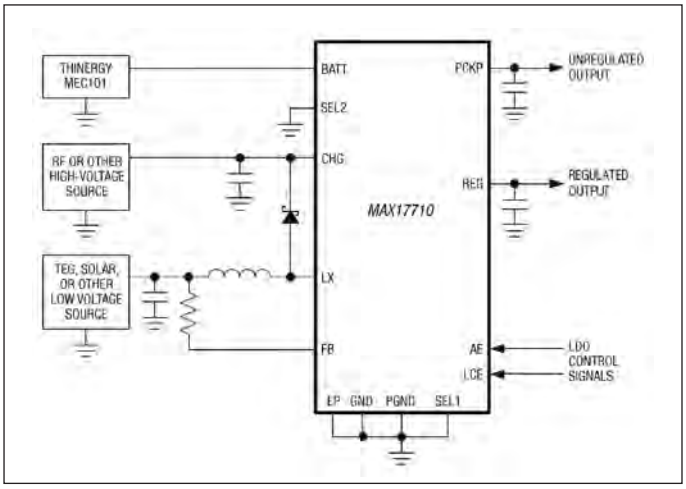


Figure 3: *Simplified operating circuit of the MAX17710* (Image: Maxim Integrated)

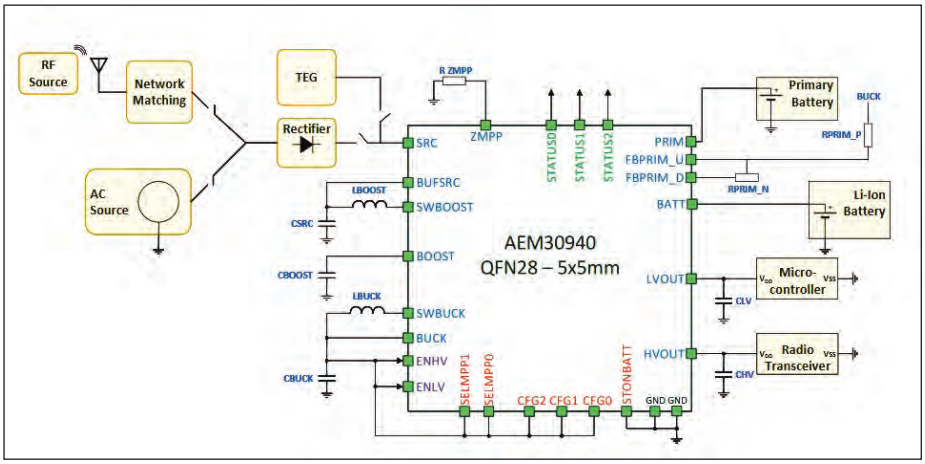


Figure 4: *Typical application circuit for the AEM30940* (Image: e-peas)

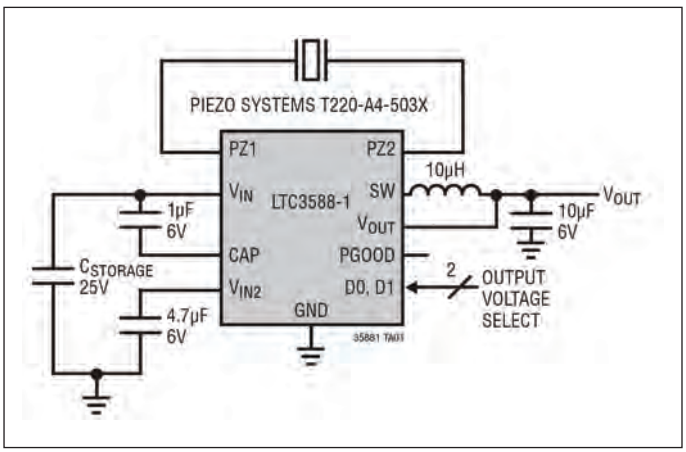


Figure 5: *Complete energy storage solution, optimized for high-impedance generators such as piezoelectric transducers* (Image: Analog Devices)

Design considerations for thermoelectric-energy–harvesting systems "include electrical and thermal requirements, thermoelectric materials, application-specific considerations, durability targets, selling price, and engineering budget," said Piggott.

Vibrations are a ubiquitous energy source. Every car on the road creates vibrations on the asphalt and in the cab. Given the number of cars on the world's motorways, the appeal of deriving energy from vibrations is clear.

Maxim Integrated's MAX17710 PMIC is a complete system for charging and protecting micropower-storage cells and can manage poorly regulated sources such as energy-harvesting devices with output levels ranging from 1 µW to 100 mW (**Figure 3**).

The AEM30940 integrated energy management subsystem from e-peas extracts DC power from a TEG, piezo generator, microturbine generator, or high-frequency RF inputs to simultaneously store energy in a rechargeable element and supply the system with two independent regulated voltages. The PMIC integrates an ultra-low-power boost converter to charge a storage element, such as a lithium-ion battery, a thin-film battery, or a super- or conventional capacitor. It can start operating with empty storage elements at an input voltage as low as 380 mV and an input power of just 3 µW (**Figure 4**).

The LTC3588-1 integrated circuit shown in **Figure 5** offers a complete energy storage solution, optimized for high-impedance generators such as piezoelectric transducers. The Analog Devices circuit features a low-loss full-wave rectifier and a high-efficiency synchronous buck converter that transfers energy from a storage device at the input to a regulated voltage output capable of supplying loads up to 100 mA. It is available in a 3 × 3-mm DFN or 10-conductor MSE package.

To design a fully autonomous wireless sensor system effectively, you need low-power-consumption microcontrollers and transducers that consume a minimum amount of electricity using low-energy environments. A power solution for such systems may consist of storing mechanical, thermal, or electromagnetic energy available in the local environment of the sensors themselves.

Supercapacitors are the technological prerequisite to energy harvesting. They have the functional characteristics of electrolytic capacitors and rechargeable batteries but can store 10× to 100× more energy per unit volume or mass than an electrolytic capacitor. They can accumulate charge at far higher speeds and withstand far more charge-discharge cycles than typical rechargeable batteries. ■

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

EMBEDDED COMPUTING

SODIMM-Format SoM Targets Embedded Designers

By Maurizio Di Paolo Emilio

Direct Insight presented the TRITON-TX8M system-on-module (SoM) in a SODIMM format at embedded world. The module is based on NXP's affordable i.MX8M Mini Quad Arm Cortex-A53 processor, which has four 64-bit Arm Cortex-A53 cores running up to 1.6 GHz (**Figure 1**).

A system-on-module is one step above an SoC. It incorporates connectivity, multimedia and display, GPIO, an operating system, and others in a single module. On the other hand, SoM-based designs are usually scalable to achieve a fully customized electronics assembly in terms of interfaces and form factors. SoMs can be replaced or upgraded within a carrier board. Some advantages of the SoM approach over SoC development include cost savings, reduced market risk, reduced customer design requirements, and reduced footprint.

"SODIMM modules are very good because they use a standard format; you can plug, unplug, and reprogram them into your carrier board," said David Pashley, Direct Insight's managing director.

Unlike an SoM, an SoC can contain various digital and analog functions on a single substrate. One of the most significant advantages of an SoC is that it is usually more energy-efficient. At the hardware level, the SoC still has one significant disadvantage: You are stuck in that configuration for life, and this limits the manufacturer-related applications.

Integrating everything on a single board also creates some problems. Once the SoC is built, it can no longer be modified. That's why it's vital to know what its destination is before you design it. Despite this precaution, the programmer will have to do somersaults to create software that does not exceed the limits imposed by the hardware. To overcome this limitation, SoM comes to our aid and offers us much more flexibility.

"There are problems with the adoption of system-on-modules," said Pashley. "It's not just as a component, because you have APIs and complex integrations to do, and to take advantage of the time-to-market opportunity, it all needs to go smoothly.

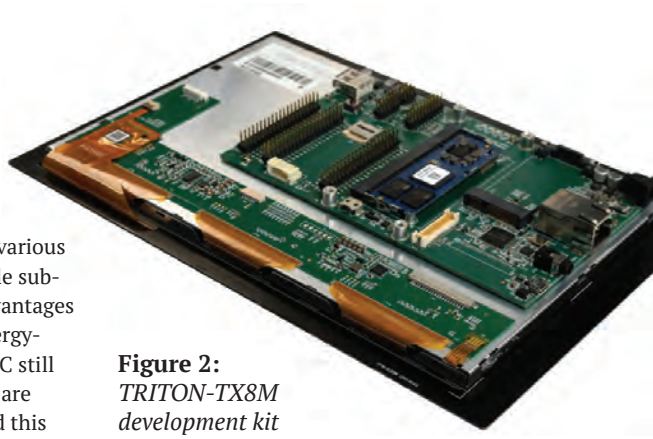


Figure 2: *TRITON-TX8M development kit* (Image: Direct Insight)

"The only limitations you have from a system-on-module compared with designing from the ground up with a system-on-chip is that, obviously from a hardware standpoint, there are fewer pins. The SODIMM module has 200 pins. The actual system-on-chip has over 400 pins, so you'll appreciate that there is some limitation in the number of interfaces."

The tiny TRITON module, 68 × 26 mm (LVDS version: 28 mm), comes with 1,024-MB or 2,048-MB DDR3L and a 4-GB eMMC. A second complex core is equipped with an Arm Cortex-M4F microcontroller, and full graphics capabilities are provided by powerful 3D and 2D GPUs.

The module provides a wide range of connectivity, including an Ethernet port, two USB 2.0 ports, a MIPI-DSI display or an alternative construction with an LVDS display, MIPI-CSI camera input, and many other interfaces. The i.MX8M Mini's VPU provides 1080p encoding and decoding capabilities. The TRITON-TX8M comes with Linux or Windows 10 IoT Core BSP, providing an ideal platform for mission-critical projects. Direct Insight is also developing a BSP for QNX 7.0. The modules in the TRITON-TX family based on i.MX6- and i.MX8 based on i.MX8 all take a single 3.3- to 5.5-V power supply and provide up to 300-mA, 3.3-V output power for use on the baseboard.

Direct Insight also offers a TRITON-TX8M development kit with a 10-inch touchscreen, 1,280 × 800 resolution, and other interfaces such as a MiniPCI expansion port, RJ45 Ethernet connector, 3.5-mm audio jack, 120-pin expansion head, MIPI-CSI camera port, and micro-SD card (**Figure 2**). ■

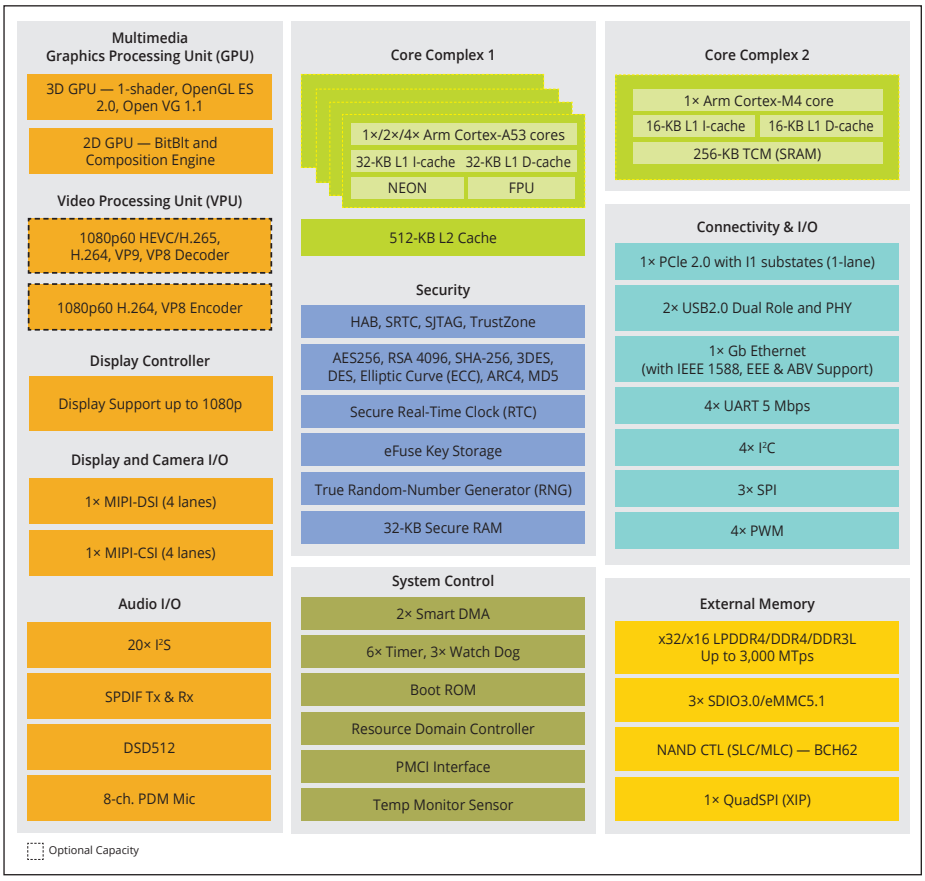


Figure 1: *i.MX8M Mini block diagram* (Image: Direct Insight)

Maurizio Di Paolo Emilio *is a staff correspondent at AspenCore, editor of Power Electronics News, and editor-in-chief of EEWeb.*

OPINION

Huawei Chooses France, But Will France Choose Huawei?

By Anne-Françoise Pelé

France secured a total of €8 billion in investments from business leaders in January at its third Choose France Summit, held at the Palace of Versailles. That doesn't include Huawei's plan to spend more than €200 million on a new wireless equipment plant in France. Should we light up Louis XIV's Hall of Mirrors and run the Grand Fountains in celebration? Not yet.

China's Huawei Technologies announced it would build its first European manufacturing plant — and the second outside China — on French soil. The "highly automated and intelligent facility" is expected to produce 4G and 5G equipment and to have a demo center that will showcase the wireless base station production, software loading, and testing process.

The €200 million investment, which covers the acquisition of the land, construction, and setup of the machine tools and equipment, is just the first phase. Huawei estimates that the project will create 500 jobs and that the direct annual economic activity will amount to €1 billion.

"This site will supply the entire European market, not just France," Huawei chairman Liang Hua told a press conference in late February.

So why center the operation in France? "As one of the world's most advanced manufacturing centers, France has mature industrial infrastructure and a highly educated labor pool, and its geographic positioning is ideal for Huawei," the company stated.

Indeed, Huawei has had a presence in the country for some time. In 2018, the telecom giant picked Grenoble for an R&D center dedicated to sensors and parallel processing software. At the time, Huawei already had three R&D centers in the Paris region and one in Provence-Alpes-Côte d'Azur. It currently employs 1,000 people in France and 12,000 in Europe.

A CHARM OFFENSIVE?

No matter how genuine its intentions, it is tempting to think that Huawei's aim is to court France — and Europe.

Huawei, caught up in a full-fledged trade war between Washington and Beijing, is accused of threatening U.S. security. The Trump administration has been pressuring European allies to ban the Chinese company from building their 5G infrastructure. Despite the arm twisting, the U.K. government recently announced that it has allowed Huawei to build non-critical parts of the country's 5G network, and the European Commission confirmed that it would not exclude Huawei on the continent.

IMAGE: SHUTTERSTOCK

Is the battle half-won for Huawei in Europe? Not really. Germany is struggling to reach consensus, and France is in the early stages of rolling out its 5G network and has not yet selected suppliers.

Wait. If Germany and France are involved in a decision-making process, assessing the pros and cons, isn't it the best time to offer reassurances and build confidence? Isn't it an opportune time to announce a plant likely to generate extra cash and extra jobs for the local economy?

No need to answer; Liang himself felt the need to justify Huawei's project when he said, "This is not a charm offensive."

According to Huawei, as one of the world's most advanced manufacturing centers, France has mature industrial infrastructure and a highly educated labor pool, and its geographic positioning is ideal for the company.

One detail caught my eye: The plant location has not been unveiled yet. Why would Huawei make such an announcement if it doesn't have a city nailed down? The only reason is to draw the attention of the media and let the news ricochet through the politicians as they fight for their regions.

The French Ministry of Economy diplomatically welcomed the announcement, regarding it as the obvious proof of France's attractiveness. Nonetheless, it is not rolling out the red carpet, as it usually does for foreign investors.

Words need to be chosen wisely before they are uttered, especially after the Chinese embassy in Paris urged the French government not to discriminate against Huawei. "If, due to security concerns, the French government truly does have to impose constraints on operators, it should establish transparent criteria around this and treat all companies equally," stated the embassy, specifying that security fears about Huawei were unfounded.

In a mid-February TV interview, French minister of economy and finance Bruno Le Maire said that France would not discriminate against Huawei, but "some decisions will be taken to protect our sovereignty." He further commented, "When you have a military camp or a nuclear plant with sensitive technologies, it is normal for the French state to protect them and apply restrictions."

Pro-European Le Maire added that it would be "perfectly understandable" to favor European 5G equipment suppliers, namely Nokia and Ericsson. "When we have two European 5G suppliers with quality equipment, it makes sense to see if they can provide us with solutions."

France's No. 1 mobile operator, state-controlled Orange, is committed to using 5G equipment from Nokia and Ericsson. Bouygues Telecom, however, has been collaborating with Huawei since 2012 on 4G and was one of the first operators worldwide to conduct field trials of 5G technology in real conditions in Bordeaux.

There has been rising discontent within the ranks of French telecom operators since French parliamentarians voted on a new law to give the government the power to control the rollout of 5G networks and infrastructure. Under the law, the telecom operators must seek formal permission from the French prime minister for their network projects. The process is burdensome, and operators are afraid it might delay rollouts.

Neither Huawei nor ZTE is mentioned, but there's no doubt that the law targets both Chinese equipment providers. It is commonly known as the Huawei Law.

5G will transform the way we live and work. Connecting everyone and everything, it promises higher economic and societal benefits than any previous generation of mobile technology.

There is a lot of money on the table, and much at stake, for mobile companies willing to compete. 5G is indeed expected to account for 15% of global mobile connections by 2025 and to contribute US$2.2 trillion (€1.98 trillion) to the global economy over the next 15 years, according to a report by telecom-industry lobby group GSMA, which represents the interests of 750 mobile operators.

GSMA, however, assessed that excluding Huawei and ZTE from Europe's rollout of 5G telecom networks would cost European mobile operators up to €55 billion. Nokia and Ericsson simply don't have the capacity to address migration from 3G and 4G networks to 5G, and Europe risks to lose ground in this race.

With the U.S. turning its back on Huawei, Europe has become a major battleground. Building a European 5G factory in France is well-played, but Huawei has not won the game yet.

We don't know exactly what the word "restrictions" encompasses and if France is not just being too politically correct to say no. ■

Anne-Françoise Pelé *is editor-in-chief of* eetimes.eu.

ADVERTISERS INDEX

EE Times EUROPE

For reader inquiries and address changes, please contact: *christiane.lockner@aspencore.com* or call +49 8092-247740

To unsubscribe, please go to: *www.eetimes.eu/unsubscribe/*

ASPENCORE

Aspencore Media GmbH
Rauwagnerstr. 5
85560 Ebersberg / Germany

EDITORIAL MANAGEMENT
Bolaji Ojo, Global Editor-in-Chief, *bolaji.ojo@aspencore.com*
Junko Yoshida, Global Co-Editor-in-Chief, *junko.yoshida@aspencore.com*
Jürgen Hübner, Managing Editor, *j.hubner@aspencore.com*
Anne-Françoise Pelé, Editor-in-Chief, *eetimes.eu; afpele@aspencore.com*
Brian Santo, Editor-in-Chief, *eetimes.com*
Echo Zhao, Chief Analyst, *echo.zhao@aspencore.com*
Judith Cheng, Managing Editor, *Judith.cheng@aspencore.com*

CONTRIBUTING EDITORS
George Leopold, *gleopold@gmail.com*
Ann Thryft, *athryft@earthlink.net*

STAFF CORRESPONDENTS
Matthew Burgess, *matthew.burgess@aspencore.com*
Anthea Chuang, *anthea.chuang@aspencore.com*
Nitin Dahad, *nitin.dahad@aspencore.com*
Maurizio Di Paolo Emilio, *maurizio.dipaolo@aspencore.com*
Yvonne Geng, *yvonne.geng@aspencore.com*
Amy Guan, *amy.guan@aspencore.com*
Susan Hong, *susan.hong@aspencore.com*
Illumi Huang, *illumi.huang@aspencore.com*
Barbara Jorgensen, *barb.jorgensen@aspencore.com*
Clover Lee, *clover.lee@aspencore.com*
Shao Lefeng, *lefeng.shao@aspencore.com*
Jenny Liao, *jenny.liao@aspencore.com*
Elaine Lin, *elaine.lin@aspencore.com*
Luffy Liu, *luffy.liu@aspencore.com*
Challey Peng, *challey.peng@aspencore.com*
Gina Roos, *gina.roos@aspencore.com*
Fendy Wang, *fendy.wang@aspencore.com*
Sally Ward-Foxton, *sally.wardfoxton@aspencore.com*
Demi Xia, *demi.xia@aspencore.com*
Franklin Zhao, *franklin.zhao@aspencore.com*
Momo Zhong, *momo.zhong@aspencore.com*

COPY AND DESIGN
Adeline Cannone, Design Director, *adeline@cannone.com*
Lori O'Toole, Chief Copy Editor, *lotoole@aspencore.com*
Diana Scheben, Production Chief, *diana.scheben@aspencore.com*

TECHNICAL AND ADMINISTRATIVE SERVICES
Tracey Bayer, *tracey.bayer@aspencore.com*

SALES & MARKETING
Philip Strange, Director of Sales, ASPENCORE EMEA, *philip.strange@aspencore.com*
Claudia Mellein, *claudia.mellein@aspencore.com*
Christiane Lockner, *christiane.lockner@aspencore.com*
The Hufmann Agency, *victoria@hufmann.info*, *norbert@hufmann.info*
Todd Bria, *todd.bria@aspencore.com*

Bolaji Ojo, Group Publisher, ASPENCORE

OPINION

Do We Need to Rethink Trade Shows?

By Nitin Dahad

The EE Times Europe team came out in force in Nuremberg, Germany, in late February to attend embedded world 2020, despite the avalanche of cancellations right up to the last minute, especially by the major chip companies. No Infineon, Microchip, NXP Semiconductors, Renesas, Silicon Labs, STMicroelectronics, Texas Instruments, Toshiba, Xilinx — the list goes on. More than 200 companies bowed out, according to the organizers.

But the feedback we gathered from visitors and exhibitors was unexpectedly positive. The comment we heard most often throughout the three days of the trade fair was that while it may have been quiet, those who attended had been able to spend more quality time with customers, partners, and suppliers.

Why was that so? Because they'd planned meetings in advance, and many individuals and smaller companies had made the trip anyway. And because there weren't hundreds of "tire kickers" milling about, exhibitors and visitors were able to spend more time with the people they had planned to see. I'm not saying tire kickers are bad — some casual booth visitors become leads, and some of those become new customers. But engaging with everyone is time-consuming and can be counterproductive, particularly when what you're really there for are the face-to-face meetings with the people you've been courting over many months of emails and telephone calls.

So while concern over the coronavirus kept some contingents home, there was still plenty of business and serious discussion going on at embedded world this year. Halls might have been half-empty and visitor numbers down to about 13,800, but much of the feedback was that the event had been worthwhile nonetheless.

Since then, we've seen other big events and meetings canceled. Nvidia moved its GPU technology conference, GTC 2020, from the San Jose Convention Center to an online event. In its statement, the company said that Nvidia founder and CEO Jensen Huang would still deliver a keynote address exclusively by live stream, and they'd work with the conference speakers to publish their talks online.

I also took a briefing from Intel's Barefoot Networks division, which changed the venue for the group's planned demo of the industry's first silicon photonics co-packaged optics switch. Rather than showcase the product at OFC 2020 in San Diego, the group decided to pull out and instead demo its technology on its own premises at specially arranged customer, media, and analyst days.

It's a common theme: Organizers and companies are finding creative ways to present their new products and technologies. In the case of this year's canceled or downsized events, the target audience is already primed; people who had planned to visit a company may be tempted to go online and see its live stream or webinar, view a talk, or download a paper. But we could see this year's contingency arrangements become a trend.

Many trade fairs and gatherings make their money by selling conference and visitor passes and selling exhibition space. The events of the past few months have certainly been an eye-opener for organizers and exhibitors alike. Some exhibitors might wonder whether there is a smarter way to use the hundreds of thousands of dollars they would normally spend on exhibition space, stand build, and travel costs — especially when you consider the cost of participating at an event like Mobile World Congress, which this year was canceled outright, with very little chance that refunds would be forthcoming.

There is probably still a need to participate at trade fairs, because they almost always generate business — sometimes, millions of dollars of it, people have told me anecdotally — if done the right way. A platform that offers face-to-face interaction can definitely be a worthwhile investment. But companies might pursue a better mix of conference booths and sessions, streaming and webinars, and face-to-face meetings.

IMAGE: NÜRNBERG MESSE / FRANK BOXLER

Maybe the lesson from embedded world is that the point of a trade fair is to meet serious people to do serious business. If you're in it just for lead generation, there may be a better way to generate leads. If you want to foster awareness of a new technology, maybe live-streamed events and webinars are a sound option.

We'll clearly have to wait a few months to understand how effective the alternative forms of communication and information dissemination have been in the wake of the many canceled events. But there is a chance that event organizers and participants will rethink the purpose of trade shows and consider making them part of a wider mix of marketing channels. ■

Nitin Dahad *is a staff correspondent at AspenCore.*

POWER ELECTRONICS NEWS

ASPENCORE

EE Times EUROPE

POWER UP EXPO
The Virtual Power Electronics
JUNE 16-19, 2020
www.powerup-expo.com

Event Cancelled?
We've Got the Solution.

ASPENCORE offers POWER-UP, the Online Virtual Expo for players in the power electronics industry to meet and explore business and design opportunities.

Showcase Your Company's Power Electronics Innovations & Products to a Global Audience. Engage Virtually With:

- Design Engineers
- Power and Power Train System Builders
- Software Developers
- Executives and other Decision-Makers
- Power Supply Chain Architects
- Component Suppliers & Distributors

For information on Virtual booths, keynote presentation opportunities, webinars, podcasts, technical papers, networking events and interviews with ASPENCORE editors go to

www.powerup-expo.com/engage

Principal Host
Maurizio Di Paolo Emilio
Physicist
EE Times Correspondent &
Power Electronics News Editor-in-Chief

www.powerup-expo.com

NEW: CoolSiC™ trench MOSFETs 650 V

Top efficiency / density (> 97%, > 35 W/inch^3) for HV SMPS

Utilizing the best of SiC

Combined with unique features enhancing performance, robustness and ease of use.
State-of-the-art trench semiconductor process, optimized to
minimize losses in applications and maximize reliability in operation.

What else is special?

- Gate-oxide reliability
- Robustness against parasitic turn-on
- Robust body diode
- Wide V_{GS} range for ease of use
- Short-circuit capability and thermal conductivity (suitability for harsh-condition / high-temperature operation, low dependency of $R_{DS(on)}$ with the temperature)

Interested in more details?

www.infineon.com/coolsic-mosfet-discretes