Unlocking Efficiency: Exploring the Advantages of Heat Pumps

ALSO IN THIS ISSUE

Special Report: Autonomous Vehicles
PAGE 10

Special Report: Supercomputers
PAGE 44

Embedded World MCU Innovations
PAGE 56

COVER STORY:
Unlocking Efficiency: Exploring the Advantages of Heat Pumps
PAGE 6
Ready to automate and accelerate productivity?
Stay one step ahead in your planning and business automation. Connect your software systems directly using our REST API without the need of a third party provider. Access current and future stock availability, technical product parameters, data sheets and more – all in real time in a common standard format.

Discover the capabilities of our API!
www.we-online.com/API

**Highlights**
- High data quality directly from manufacturer
- Information in realtime
- Future availability of stock
- Individual configuration
- Reliable and safe connection

WE meet @ PCIM Europe
Hall 6 - 342
Microsoft, Amazon, IBM and IQM Choose France. Will You?

The chandeliers in the Hall of Mirrors at the Château de Versailles shone brightly in May when French President Emmanuel Macron welcomed 180 French and foreign business leaders to the seventh edition of the Choose France Summit. Annum mirabilis: France set a record by securing €15 billion in foreign investments, with the potential to create 10,000 jobs. Notably, the red carpet was ceremoniously rolled out for AI and quantum players looking to set up in France.

"Nous choisissons la France," proclaimed Brad Smith, vice president and chairman of Microsoft, as he shook hands with Macron at the Choose France Summit. Building on its 41-year presence in the country, Microsoft pledged to invest €4 billion in cloud and AI infrastructure and bring "up to 25,000 of the most advanced GPUs to the country by the end of 2025." The Redmond, Washington-based company aims to train 1 million people and support 2,500 French AI startups via the newly launched Microsoft GenAI Studio by 2027.

Amazon, for its part, announced a plan to invest more than €1.2 billion in its Amazon Web Services cloud infrastructure "to support France’s flourishing generative AI opportunity" and logistics infrastructure to increase delivery speeds. The new investment is expected to create more than 3,000 additional permanent jobs in the country.

U.S. tech giant IBM committed to extending the reach of IBM France Lab Paris-Saclay, which was unveiled at Choose France 2019 with a mission to target AI research. The lab’s focus will now include quantum computing, with a new investment of €45 million and the recruitment of 50 researchers and engineers over the next year. The objective is to contribute to the French quantum ecosystem through education, the development of technological use cases and collaborative research.

Turning south, Espoo, Finland-based IQM Quantum Computers confirmed a collaboration with CEA-Leti in Grenoble to set up a pilot line dedicated to the high-throughput production of large-scale quantum processors. IQM also committed to building an industrial-scale manufacturing facility in France. Production is scheduled to start in 2027, and the overall project represents an investment of €100 million.

The list doesn’t end there: In total, 56 investment projects were served up at this year’s Choose France. Is it appropriate to proudly shout cocorico? As a Frenchwoman, I don’t deny my enthusiasm.

The Élysée claimed 10,451 foreign investment projects in France and 507,940 jobs maintained or created over the 2017–2023 period. Summit organizers also referred to the annual EY European Attractiveness Survey 2024, which shows that France tops the league table for foreign investment. The U.K. has moved up to second place with a 6% increase, while Germany is relegated to third place on a sharp, 12% drop from the previous survey.

Unsurprisingly, the Élysée didn’t specify that the number of foreign direct investment projects had fallen by 5% in 2023, according to the 23rd EY European Attractiveness Survey. Moreover, in R&D, an attractiveness driver for France in recent years, projects have been slackening (down 15% in 2023 from 2022), with investment intentions less marked than in the U.K. (46% versus 54%).

It’s easy to be dazzled by the royal splendor of the Château de Versailles, but the Choose France Summit is nothing more than an international communications campaign, and the French government makes no secret of it. Nevertheless, would these investments have taken place without the lavish reception? Most certainly.

Money doesn’t fall from heaven. Large corporations are not philanthropists but financiers with established KPIs to measure success. France has some unique strengths, but it’s worth highlighting the advantageous measures to reduce corporate taxes and social charges. Besides, France’s research tax credit covers 30% of all R&D expenses up to €100 million, and 5% above that threshold.
COVER STORY

6 Unlocking Efficiency: Exploring the Advantages of Heat Pumps
Heat pumps are rapidly gaining popularity as a sustainable and energy-efficient solution for heating and cooling homes. This article explores the technology behind heat pumps, their types, their benefits and why they are a smart choice for environmentally conscious homeowners.

EDITOR’S LETTER
3 Microsoft, Amazon, IBM and IQM Choose France. Will You?

SPECIAL REPORT: AUTOMONOUS VEHICLES
10 Pondering the Trajectory of AV Adoption

AV TRENDS

AV SENSORS
14 Automated Driving Is Transforming the Sensor and Computing Market
16 Avalanche Photodiode Innovations Target LiDAR Cost Challenges
17 How Versatile Technologies Can Boost AV Adoption
20 New Developments in MIPI’s High-Speed Automotive Sensor Connectivity Framework

SOFTWARE-DEFINED VEHICLES
23 Ethernet Connectivity Drives Zonal Architectures

AV SECURITY
25 Opinion: Are We Prepared for Cyberthreats in the New Era of Transportation?
28 Opinion: Certifying RISC-V: Industry Moves to Achieve RISC-V Core Quality

Subscribe to EE Times Europe: www.eetimes.eu/subscribe-now
Unsubscribe from EE Times Europe: www.eetimes.eu/unsubscribe
AI-POWERED AVS

30  Overcoming Unbalanced Training Data for Safer Autonomous Driving
31  Racing on Algorithms
34  What is Nvidia Doing in Automotive?
38  Nvidia GTC 2024: Why Nvidia Dominates AI

PARTNER CONTENT

42  AMD-Powered Advantech AIMB-723 Industrial Motherboard Future-Proofs AOI Deployments

SPECIAL REPORT: EUROPE’S SUPERCOMPUTERS

44  Lumi, Leonardo, MareNostrum: Conversations with Execs Running Three of the World’s Top 10 Supercomputers
45  Lumi: CSC’s Manninen on Managing Europe’s Biggest Supercomputer—and AI’s Expectations
47  Cineca’s Bassini on the Secrets of Leonardo’s Supercomputing Power
49  BSC’s Girona on MareNostrum’s Evolution and Delivering World-Class Compute from Barcelona

THE INDUSTRY

52  Is C/C++ Memory-Safe?
56  Three MCU Innovations at Embedded World 2024
57  A Blueprint for the Integrated Photonics Industry?
58  Delft Quantum Ecosystem Addresses Connectivity Challenges

ONLINE MUST READS

▶ A Conversation with GlobalFoundries Europe’s Manfred Horstmann tinyurl.com/2nnbs9c8
▶ Imec.xpand Launches a €300M Semiconductor-Focused Fund tinyurl.com/fbbc87j4
▶ Insurmountable Problems Only Quantum Computers Can Address tinyurl.com/bdhzjv2j
Unlocking Efficiency: Exploring the Advantages of Heat Pumps

By Omara Aziz, Power Technology Segment Leader EMEA, Arrow Electronics

Heat pumps are rapidly gaining popularity as a sustainable and energy-efficient solution for heating and cooling homes. Unlike traditional heating systems that generate heat, heat pumps work by transferring heat from one place to another, providing an efficient way to manage indoor temperatures year-round. This article explores the technology behind heat pumps, their types, their benefits and why they are a smart choice for environmentally conscious homeowners.

The European Union (EU) has established ambitious goals to reduce carbon dioxide emissions by 2030 as part of its broader climate strategy, committing to install 60 million heat pumps by 2030. This approach aims to tackle climate change and drive a transition toward a low-carbon economy. The International Energy Agency has estimated that heat pumps have the potential to reduce global CO₂ emissions by at least 500 million tons in 2030.

**TYPES OF HEAT PUMPS**

Several types of heat pumps use air, water, ground (geothermal) or waste as a heat source. However, the most common type in Europe is the air-source heat pump (ASHP), which extracts heat from the outside air. It is cost-effective and relatively easy to install, making it a popular choice for residential use. The ASHP is classified as air-to-air and air-to-water:

- Air-to-air heat pumps extract heat from the outside air and transfer it to the indoor air. They can also work in reverse to cool the indoor space. This is the most common type of heat pump for residential heating and cooling.
- Air-to-water heat pumps extract heat from the air but transfer it to water, which can then be used to heat homes.

![Heat-pump refrigeration cycle](Figure 1)

Figure 1: Heat-pump refrigeration cycle
to water, which can be used for heating spaces through radiators and underfloor heating or even for hot water supply.

Heat pumps operate on the simple principle of moving the heat from one location to another. In winter, a heat pump extracts heat from the outside air and transfers it indoors to heat the desired space. For cooling purposes in summer, the process reverses; the heat pump removes heat from the inside and releases it outside.

Inverters based on wide-bandgap devices like silicon carbide or gallium nitride offer higher efficiency due to lower switching losses and can operate at higher frequencies.

This transfer of heat is made possible through the refrigeration cycle. A heat pump uses a refrigerant to absorb heat at a low temperature and release it at a higher temperature. The system consists of an indoor unit, an outdoor unit and a network of pipes that carry the refrigerant between the two. The compressor, part of the outdoor unit, plays a key role in pressurizing the refrigerant to facilitate heat transfer. All components—circulation pump, fan and compressor—are operated by a motor drive to increase energy efficiency and reduce losses.

COEFFICIENT OF PERFORMANCE

The attractiveness of the heat pump is that it creates more heating energy than it takes. The coefficient of performance (CoP) for a heat pump is a measure of its efficiency in transferring heat. It is defined as the ratio of the heat energy transferred by the heat pump to the energy input required to achieve that transfer, typically in the form of electrical energy. Figure 2 shows a heat pump that uses only 1 kW of electricity to achieve a CoP of 4 kW. The heat pump is very efficient, as 3 kW of energy has been sourced from the air.

Heat pumps have a high CoP. At a high level, you get more heat energy out of the system than the electrical energy you put into it. The higher the CoP, the more energy-efficient the device is. Another way to look at it, the higher the CoP, the more money you can save on your heating or air conditioning bill.

DRIVING A HEAT-PUMP INVERTER

Heat pumps use a motor drive that operates a compressor to convert electricity into heat in the vapor-compression cycle. Driving a heat-pump compressor efficiently involves a combination of proper compressor selection, efficient motor control, advanced inverter technology and optimized system design. The goal is to achieve high energy efficiency, reliability and performance that contributes to the overall CoP of the heat pump.

Whether the heat-pump system is a monoblock or split between outdoor and indoor units, the objective is to drive the compressor, fan and circulation pump efficiently. Selecting the right topology, such as totem-pole and Vienna rectifier for the power-factor-correction and inverter stages, especially for three-phase systems, results in a high-efficiency inverter design.

WIDE-BANDGAP DEVICES IN HEAT PUMPS

Inverters based on wide-bandgap devices like silicon carbide (SiC) or gallium nitride (GaN) offer higher efficiency due to lower switching losses and can operate at higher frequencies. This leads to smaller and more efficient inverters.

Figure 2: Heat pump that uses only 1 kW of electricity to achieve a CoP of 4 kW
Advantages of WBG semiconductors over IGBTs in heat pumps include:

- **Higher efficiency**: WBG devices have lower switching losses and can operate at higher frequencies than traditional silicon-based devices. This leads to greater overall efficiency, reducing energy consumption and operational costs.

- **Higher-temperature operation**: SiC and GaN devices can operate at much higher temperatures than silicon devices. This allows for more compact heat management systems and can reduce cooling requirements for power electronics in heat pumps.

- **Smaller size and weight**: Higher switching frequencies allow for smaller inductors and capacitors, which can reduce the overall size and weight of the heat-pump system. This feature is particularly beneficial for compact installations or portable heat-pump applications.

- **Greater reliability and durability**: WBG devices generally exhibit greater reliability because of their high thermal conductivity and durability at elevated temperatures. This leads to longer lifespans and reduced maintenance costs.

**HEAT PUMPS IN ACTION: COMPANY SOLUTIONS**

Heat pumps can be customized and tailored to specific needs, from residential units designed for single-family homes to large-scale systems for commercial buildings and industrial complexes. This flexibility ensures that heat pumps can be deployed in various settings. They also work well with renewable energy sources like solar and wind power. The integrated systems that combine heat pumps with solar panels or wind turbines can further reduce energy costs and environmental impact.

Modern heat pumps are equipped with smart technology and connected to smart homes that can be easily controlled via a touchscreen and allow for remote monitoring and automation. Connectivity to the smart home and the internet will play a crucial role in enabling the energy transition. This enhances efficiency by enabling users to control their heating and cooling systems remotely and make adjustments based on weather conditions or occupancy.

**Heat pumps represent a significant step forward in the quest for energy-efficient heating and cooling solutions.**

**CONCLUSION**

Heat pumps represent a significant step forward in the quest for energy-efficient heating and cooling solutions. Their versatility, energy efficiency and environmental benefits make them an attractive choice for a wide range of applications. As technology advances and companies continue to innovate, heat pumps are prepared to play a central role in reducing energy consumption and combating climate change. Whether in residential, commercial or industrial settings, heat pumps offer a compelling solution for a sustainable future.

Arrow Electronics has always been centered on promoting energy efficiency, and we are eager to contribute to this discussion by offering a comprehensive portfolio for single- and three-phase solutions across all power classes, making it easy to select the most suitable solution for heat-pump systems. A wide product portfolio of SiC power devices, IGBTs, gate driver ICs, intelligent power modules, motor controllers and reference design boards can be found at Arrow.com.
The PowerUP and Fortronic Conference & Exhibition is where the vanguard of power electronics engineering converges with sustainability and innovation. This event is bringing the latest developments in power electronics engineering and renewable energy, debated by global industry leaders across two days of keynotes, panels, and technical presentations, with an exhibition area for networking.

The conference will dig into several key topics:

- Power semiconductors (GaN, SiC, IGBTs)
- Power management, converters, sensors, passives
- Tools/test and measurement
- Renewable energy
- E-mobility and infrastructures/smart grid
- Motor control
- And more!

For more information about sponsorship opportunities, please contact: sales@aspencore.com

powerup.fortronic.it
Pondering the Trajectory of AV Adoption

By Anne-Françoise Pelé

Over the past decade, autonomous driving technologies have advanced significantly, moving beyond the Gartner hype cycle with real-world deployments. Robotaxis are now on the streets of San Francisco, Beijing and Shanghai, and the first L3 autonomous systems have been launched. By 2035, McKinsey expects that autonomous driving will create between US$300 billion and US$400 billion in revenue.

However, full autonomy remains a challenge, and it might be some time before we can read a book and enjoy the scenery while our AV drives us safely to our destination. Safety—for us, our passengers and those around our vehicle—relies on the sensors operating properly. Today’s vehicles are equipped with dozens of camera, radar and LiDAR units, which work together with sophisticated algorithms and powerful processors to deliver intelligence and autonomous functions. The promise of fully autonomous, safe and reliable AVs will be possible only with a collaborative, multi-disciplinary approach.

An AV is a safety-critical system that leaves no room for error, malfunction or compromise. Any failure can cause accidents that result in serious injuries or even death. Who is responsible for an accident with a self-driving car? What sensing solutions are needed to make AVs smarter? What are the latest developments in sensor fusion technology? How do software-defined vehicles reset the safety bar? How is AI making AVs more reliable? Can AI-powered self-driving cars race on F1 tracks? What is Nvidia’s impact on AI-defined AVs?

This Special Report will try to answer these fundamental questions and open up new conversations. It relies on contributions by IDTechEx, Yole Intelligence, MIPI Alliance, Secure-IC, Breker Verification Systems and VSI Labs, as well as discussions with Phlux Technology, Innoviz Technologies, BlueWhite, Analog Devices Inc., the Computer Vision Laboratory at Sweden’s Linköping University and the Technical University of Munich Autonomous Motorsport Team.
The automotive industry is going through the most transformative journey in its history. Only 15 years ago, autonomous cars were a pipe dream; electric cars were the butt of jokes; and safety was reactive, aiming to minimize the consequences of collisions rather than prevent them. Now, in 2024, autonomous mobility services are available to the public in multiple locations across the U.S. and China; electric cars are a desirable mainstream option; and the future of safety is preventative, trying to ensure crashes never happen. This article considers three critical aspects of future vehicles: sensors, software and safety.

SENSORS

Radar, cameras and LiDAR form the core trio of sensors synonymous with autonomous driving. All three have been evolving in the automotive industry and continue to progress. LiDAR is coming down to a price point more palatable for OEMs, radar’s performance and utility continues to evolve, and high-resolution camera technologies pioneered for smartphone cameras are being leveraged in autonomous driving systems.

Radar is perhaps the most interesting of the three categories. Automotive radar has come a long way since it was introduced to vehicles around the turn of the century. Back then, radar was good for detecting the distance to the vehicle ahead, and that was about it. Its major advantage was that it worked under almost any weather conditions the car might encounter. Snow, rain, fog, sandstorm—the radar did not care. For fully autonomous driving, however, its resolution simply was not sufficient.

In recent years, advances in semiconductor technology and the emerging demands of the autonomous driving industry have forced radar to evolve. It has sprouted more transmitter and receiver channels, growing from 1 × 2 to 3 × 4 and then 12 × 16Tx/Rx arrays from leading Tier 1 companies and 48 × 48 arrays from leading startups. Large antenna arrays with many transmitting and receiving channels have boosted the resolution of modern radar.

However, transmitting and receiving channels are like pixels on a camera: Simply adding more pixels does not necessarily make a better camera. As far back as 2013, the Nokia Lumia 1020 had 41-megapixel cameras, yet the iPhone 14 from 2022 uses 12-megapixel sensors. Which would take a better photo? There is more at play than just the pixel count. Sensitivity, dynamic range and post-processing software make a significant difference to the end image. Likewise, radar performance is not just a measure of transmitters and receivers.

An emerging priority for radar is dynamic range, or the ability of radar to detect both...
high-reflectivity objects and low-reflectivity objects simultaneously. Consider, for example, a small child standing near a car. The reflection from the car is orders of magnitude brighter than that from the child. For the radar, detecting the child is like trying photograph a grain of rice in front of the sun. It’s very easy for the child’s reflection to be lost to noise.

Advances in semiconductor technology and radar design are improving radar’s dynamic range. For example, at CES 2024, Mobileye showed an example of a radar that could detect a wooden pallet next to a metal railing at nearly 240 meters.

**It is an exciting time as the automotive industry goes through the most transformative journey in its history.**

Processing is another area in which radar utility can be improved. The industry has started using the nomenclature “4D imaging radar,” referring to azimuth (horizontal angle), elevation (vertical angle), distance and relative velocity. But radar has another dimension that could further improve its performance: intensity, or reflection strength. By considering the strength of a reflection, radar could determine whether something is metallic (high reflectivity) or organic (low reflectivity). With this distinction, radar would become an even more complete sensor, offering low cost, operation under all weather conditions, accurate ranging and velocity and, finally, accurate object classification. Researchers are working on intensity detection capabilities now, and the technology could soon make its way onto radar or, alternatively, onto the car’s central autonomous driving and sensor fusion computer.

Previously, radar data was processed on the radar itself. Local processing makes sense for a task like automatic emergency braking, which requires only a flag that something is stopped in front of the vehicle. But radar data is so rich that startups like Zendar, Zadar Labs and others want to handle the processing on a much more powerful, central computer. Such machines not only would handle radar data but would accept high-resolution video feeds from multiple cameras, LiDAR data and more.

With so much data to process, the industry is moving to more advanced semiconductors and more powerful systems-on-chip. Nvidia’s Thor, expected in 2025, is the prime example. With 2,000 TOPS of processing power, and utilizing TSMC’s 4-nm node technology, this SoC has more in common with a high-end gaming PC than with the chips running systems in most consumer cars today. With the advent of Chat GPT, the demand for AI within cars and the emergence of software-defined vehicles, a modern high-end electric car will have more in common with a computer than with an average car from 20 years ago.

**SOFTWARE**

Horsepower is becoming democratized. Scaling an electric motor to double or triple the horsepower is much cheaper than doubling the output of an internal combustion engine. Cars that would put the supercars of yesterday to shame are now entering the market at less than US$50,000. In combination with this, electric cars do not possess the differentiating characteristics of engines. They do not have turbo lag or differing exhaust notes; most do not even have more than one gear. In short, future cars will be differentiable not by their performance but by their software. Software-defined vehicles, connectivity and autonomy also provide a route to perfect safety. The problem with sensors, even the best and most advanced sensors, is that they cannot see around corners or through buildings. No matter how many radar, LiDAR or camera units are fitted to a car, they cannot see that a speeding car on an adjacent city street is about to jump a red light. But connected vehicles could.

Connectivity and software offer a pathway to collective perception, in which cars and infrastructure would share their environmental knowledge with each other. Armed with...
such information, a connected car could react appropriately to the location and movement of every pedestrian, cyclist and vehicle on the road around it, including those beyond its line of sight. Combine collective perception with a responsible autonomous driver, and a country full of such cars could end road traffic accidents altogether.

This scenario is, of course, a long way off; autonomy on the roads is in the early days, and collective perception is a university research project. However, autonomous vehicles are already starting to show their potential safety benefits.

SAFETY

In October 2023, California ordered General Motors to remove its driverless Cruise cars from state roads, stating that GM had “misrepresented” the safety of the technology. It is clear from this incident and others that AV safety claims will always and should always be subject to public scrutiny. This is not to disparage AVs or their safety, however. Rather, public oversight should be undertaken not with the motivation to eliminate autonomous cars but with the intent to improve them.

Unlike accidents involving human drivers, the intelligence gathered on a single AV collision can generate improvements that can be rolled out via over-the-air updates to the entire fleet. In this case, a collision provides an opportunity to prevent that type of collision from ever happening again, something far beyond the capabilities of human drivers.

By all metrics, the safety of AVs is improving. The data from California shows as much. Going back to 2015, when records were first kept, autonomous cars would barely make it 1,000 miles without the safety driver behind the wheel intervening. Fast forward to 2023, and Cruise submitted 576,000 miles worth of driving without a single disengagement. By comparison, IDTechEx estimates that human drivers in the U.S. are involved in collisions approximately once per 200,000 miles, and in cities like San Francisco, this rate nearly doubles. So perhaps the argument can be made that autonomous cars are already safer than human drivers.

On the other hand, a look through the data on driverless autonomous cars (without a safety driver on board) reveals that, on average, a driverless Waymo is involved in collisions once every 52,000 miles. Cruise fairs slightly better, with one collision per 65,000 miles. So, they are colliding at roughly twice the rate of average San Franciscoan drivers. But again, this is not the whole story. Out of 33 recorded collisions involving a driverless Cruise vehicle, blame can only be attributed to the autonomous driving system in six cases, which equates to once per 344,000 miles. Likewise, Waymo’s autonomous driving system could be held responsible roughly for five collisions, equating to one per 238,000 miles. So by this metric, autonomous drivers are once again safer than the average human driver.

Another metric that should be considered is the rate at which fatalities and serious injuries occur with AVs. Once again, this is lower for AVs than for human drivers.

However, there is a big issue with how this discussion is framed, and it concerns the “average” human driver. Clearly, human driving proficiency is a spectrum, encompassing those who are incapable of staying below the speed limit and off their phones, the super-vigilant types who have never had an accident, and everyone in between. The question, then, is with which human drivers an autonomous car should be compared. Are autonomous drivers safe enough when they are safer than 90% of human drivers? Should they be better than every human driver? But how would that be defined? A 17-year-old with a fresh license may never have crashed, i.e., one collision per infinity miles.

Complicating matters is that the true safety record for human driving is not fully known. Not all collisions are reported, and not all dangerous situations or near-misses end in collisions. The industry just about has a grasp of how safe the average person is by measuring the rate at which people crash, cause injuries and cause fatalities. However, insufficient data exists to understand human driving safety across the full spectrum of abilities to the same depth as autonomous driving safety.

It was important to show with data and numbers that AVs could exceed the driving standards of the average human driver. However, to go beyond this and show that autonomous cars are safer than any human driver, the benchmark will need to change or evolve. Even beyond near misses, there are many metrics that could be used to better understand human driving safety, such as how often drivers go through red lights, how often they tailgate, how often they disobey traffic rules, how often they exceed speed limits and many more. With these metrics, the industry could better understand human driving performance and make solid comparisons to build a case that AVs are safer.

Between sensors, software and safety, it is an exciting time for the automotive industry. The advent of AVs in recent years is an era-defining moment. In 100 years’ time, examples of today’s autonomous cars and robotaxis will be exhibited in museums as part of the story of how the automotive industry followed a pathway to infallible road safety.

James Jeffs is senior technology analyst at IDTechEx.
Eyes-off driving has been around for years but appears much more complex than OEMs anticipated. One challenge is the high number of sensors required not just around the car but also inside it, to monitor the driver’s attention. Apart from sensors, computing requirements are increasing so that all the data from the various sensors can be gathered and fused to deliver an accurate view of the car’s surroundings. That alone requires a considerable investment from the entire supply chain, not only in terms of money but also in terms of resources, knowledge, time and regulation. Recent difficulties encountered by robotaxi companies have not helped gain customers’ trust in autonomous driving.

Another concern for OEMs is the extent of their responsibility if an accident occurs while a vehicle is in automated driving mode. In 2017, Audi integrated a LiDAR, supplied by Valeo, in its Audi A8 sedan, but in 2020, the German carmaker gave up Level 3, eyes-off automated driving developments because of the lack of a legal framework. This lack of regulation and the resulting liability issues prompted OEMs to stick with hands-off applications. Of course, hands-off driving has become quite significant, and some OEMs are more advanced than others. Tesla has led for many years, though some others are now catching up. However, the main competitors in this domain are not only global European or U.S. OEMs but also Chinese automotive players.

Chinese OEMs, especially new ones like XPeng, Li Auto and Xiaomi, are pushing very hard to implement new ADAS features. All are focusing on the Navigate on Autopilot (NOA) feature. Their approach is based on a similar set of sensors, including cameras, radar and LiDAR used by global OEMs, but those players are going a step further. A detailed analysis reveals they are embedding more than 20 sensors, many more than necessary to enable hands-off driving features.

Computing is also part of the game and follows the same evolution. Computing power, usually supplied by Nvidia, ranges from 500 TOPS to 1,000 TOPS, while the Mobileye Eye Q5 performs at about 25 TOPS. Some argue that Nvidia is using brute force to process all the data generated by the sensors, and this solution is not as optimized as Mobileye’s SoC. In any case, Nvidia’s platform is much more open for OEMs, while Mobileye’s is still considered a black box, even if it is improving.

Facing this market evolution, Chinese OEMs are putting a lot of effort into these new developments, waiting for new regulations that could come shortly, as testing permits for automated driving are
Automated Driving Is Transforming the Sensor and Computing Market

is expected to grow from ~US$12 billion in 2022 to ~US$27 billion in 2028, at a CAGR of 14%. Regarding computing, the growth is even more dynamic, with the computing market for ADAS expected to grow from ~US$3 billion in 2022 to ~US$8 billion in 2028, at a CAGR of 18%.

We’re in a whole new automotive era, with software and software-defined vehicles leading the charge. These trends aren’t just hype; they’re paving the way for OEMs to explore fresh applications and revenue streams. This shift will revolutionize E/E architecture, with domain controllers set to replace single ECUs.

**Eyes-off driving applications are not as simple as just adding more sensors and computing power to the car.**

Finally, LiDAR is emerging quite rapidly, and its performance is increasing, especially its range and resolution. The latest LiDAR technology will be able to generate up to 12 million points per second to enrich the point cloud.

Despite this new generation of sensors, eyes-off driving remains exceedingly challenging for the handful of OEMs engaged in this direction. As already mentioned, they will be responsible during automated driving. Therefore, they are restricting the operational domain design to ensure safety. The current use case is highway driving up to 60 km/hour, though it will soon be updated to 150 km/hour with the oncoming generation of sensors.

On the issue of responsibility, at Yole Group, we clearly see two development paths. While global OEMs are targeting the highway use case, Chinese OEMs are focused on urban applications. This can also explain the difference in the number of sensors embedded in cars. Indeed, the Mercedes S-Class, which was the first car with eyes-off capabilities, uses 13 sensors, while Chinese OEMs like XPeng, BYD, Lotus and JM Motors use between 21 and 24 sensors in their cars.

The computing needs for eyes-off applications are exploding because there is a need to fuse data from all the sensors. To do that, the E/E architecture is changing from a distributed architecture, which could still exist for hands-off applications, to a more centralized architecture in which single ECUs are replaced by domain controllers initially and possibly by zone controllers later. These controllers possess the necessary computing power to process raw data coming from sensors. The goal is to have a single unit to enable more complex algorithms and reduce processing latency.

As for market trends, the sensor market for cameras, radar and LiDAR is expected to grow from ~US$12 billion in 2022 to ~US$27 billion in 2028, at a CAGR of 14%. Regarding computing, the growth is even more dynamic, with the computing market for ADAS expected to grow from ~US$3 billion in 2022 to ~US$8 billion in 2028, at a CAGR of 18%.

We’re in a whole new automotive era, with software and software-defined vehicles leading the charge. These trends aren’t just hype; they’re paving the way for OEMs to explore fresh applications and revenue streams. This shift will revolutionize E/E architecture, with domain controllers set to replace single ECUs.

**REFERENCE**

Yole Group, which examines the latest innovations through conversations with leaders in the automotive sector, recently released a market study, “Semiconductor Sensors for Automotive 2024,” which may be read at tinyurl.com/Sz7ehts.

**Pierrick Boulay** is a senior technology and market analyst for automotive semiconductors at Yole Group.
Avalanche Photodiode Innovations Target LiDAR Cost Challenges

By Robert Huntley

Without sensors, an autonomous vehicle cannot perceive its surrounding environment and potential hazards. The same goes for the plethora of advanced driver-assistance systems (ADAS), which are increasingly becoming the norm in today’s vehicles. Some ADAS functions, such as blind-spot detection and adaptive cruise control, are required to detect over different distances, leading automotive engineers to select suitable sensor technology, such as ultrasonic, radar and LiDAR. Increasingly, through sensor fusion techniques, data from multiple detection methods are “fused” together to create a more accurate, multi-dimensional image of the environment surrounding the vehicle.

LiDAR is a popular method of detecting objects, typically up to 200 meters away, by emitting short pulses of infrared light from a laser and detecting reflections. The time of flight yields the distance between the vehicle and the object. Initially, a LiDAR subsystem was expensive, particularly those using spinning mechanical and optical methods to control the field of view (FOV). However, with its growing popularity and the development of MEMS-based FOV control, costs have been reduced significantly.

As with any sensing application, the sensor’s specifications dictate the performance characteristics of the complete subsystem. The receiving sensor in a LiDAR camera is typically a photodiode, converting the reflecting light into electrical energy. Avalanche photodiodes (APDs) are highly sensitive, and their operation uses an avalanche process that yields a multiplier factor, creating more electrical output from a given number of reflected photons.

COST STILL A MAJOR CONSIDERATION

To date, the fabrication of APDs has focused on using an indium gallium arsenide (InGaAs) process operating with a 905-nm laser. However, Phlux Technology (Sheffield, U.K.) claimed it has developed an InGaAs technology capable of operating at 1,550 nm that is 12× more sensitive than traditional APDs and overcomes many of the challenges associated with 905-nm lasers. EE Times Europe spoke with CEO and co-founder Ben White to learn why the 1,550-nm wavelength offers better performance for automotive applications.

“Phlux Technology started as a research project at the University of Sheffield,” White told EE Times Europe. “My co-founder and professor [and I] had an early breakthrough when we discovered how a particular alloy composition of antimony exhibited some unique electron properties of very low-noise behavior you could exploit for a detector.

“There is a growing adoption of LiDAR for automotive applications,” he said. “However, the cost of the laser is still a limiting challenge for mass LiDAR rollout. With our highly sensitive APD, subsystem designers can use a lower-cost laser to achieve the same performance figures as an expensive laser with a traditional InGaAs APD. That reduces the overall bill of materials down to a point where it can be mass-produced.”

EE Times Europe asked White if there are other limiting factors besides cost. “Reliability is a massive concern for automotive, since the cost of a recall, even if it’s only due to the failure of a few units, can create really bad press,” he said. “So it is essential to have extremely low failure rates, both naturally and randomly, throughout the product’s lifetime, and as we understand, it is one of the issues with today’s LiDAR systems.”

905 NM VS. 1,550 NM

White told EE Times Europe that the LiDAR industry currently appears divided into two camps regarding the best wavelength to use.

“At 905 nm, you can use cheap gallium arsenide lasers and low-cost silicon detectors.”

However, he cautioned, “At that wavelength, you are quite close to the wavelengths the eye absorbs since your cornea can still focus light onto your retina, so you have to use low-power lasers. But in so doing, you don’t have the photon budget required to see out the required 250 meters over a reasonable field of view using a single channel. You’ll see companies using 256 lines of laser detectors instead to get around that. Yes, you can use low-cost components, but you end up using a horrendous number of them to make up. So you have sacrificed cheap components for system complexity.”

White said that Phlux is focused on the 1,550-nm wavelength with its high-sensitivity Aura InGaAs photodiodes range. “At this wavelength, you can use large amounts of laser power without damaging the human eye. The cornea doesn’t focus on the light, but also, the goo in your eye absorbs that wavelength.”

IMPROVED APD PERFORMANCE AT 1,550 NM

While using 1,550 nm offers an eye-safe wavelength, many APDs are not particularly sensitive at that shortwave infrared wavelength. “With our Aura Noiseless APDs, we can deliver a much more competitive position against the cheaper, silicon-based APDs, rebalancing the design tradeoffs using low-cost lasers,” White said. “The detectors in use today haven’t evolved much in the last 25 years. The new applications, such as LiDAR, bring much higher performance demands. What differentiates us from others making detectors for 1,550 nm is that we add an antimony alloy to the compound semiconductor during manufacturing. With a 12× better LiDAR image resolution for a given laser power and an APD gain of up to 120, we can discern the smallest signals above the noise floor of a connected transimpedance amplifier and offer size, weight and bill-of-material cost reductions.”

Robert Huntley is a contributing writer for EE Times Europe.
While advancements in software and hardware components are crucial, ensuring the reliability of hardware components takes precedence. It is therefore imperative that they undergo rigorous testing to guarantee the accuracy of the data collected by these components.

Innoviz, a Tier 1 supplier of automotive-grade LiDAR sensors, launched its InnovizTwo LiDAR sensors and demonstrated the sensors’ resilience in a successful winter test.

**LiDAR SENSORS PAIRED WITH AI-POWERED PERCEPTION SOFTWARE**

LiDAR sensors, which use laser beams to measure precise distances and movement in an environment in real time, give AVs a precise perception of their surroundings. At some point, AVs will be subjected to harsh environmental conditions like snow and rain, which can alter the data collected by the sensors, leading to accidents. As part of a test, Innoviz subjected its sensors to over 120 hours of driving in 10 countries. But as indispensable as testing may be for the sensor itself, “InnovizTwo is designed as a platform rather than a single configurable product,” Elad Hofstetter, CBO of Innoviz Technologies, said in an interview with EE Times Europe.

The data collected by InnovizTwo sensors is used to improve Innoviz’s proprietary AI-powered perception software. “The hardware development team works hand in hand with the software team to ensure that the LiDAR design aligns with the requirements of the perception software,” Hofstetter said. “For instance, the hardware team may modify LiDAR components to improve data quality or increase scanning capabilities based on feedback from the software team regarding perception algorithm requirements. Conversely, the software team provides valuable insights to the hardware team by identifying potential issues or limitations in LiDAR’s performance that may impact perception algorithms.”

This collaborative approach enables iterative improvements to hardware and software components, resulting in a more robust and reliable solution. Furthermore, the perception software has calibration and semantic point-cloud (Figure 1) features to enhance the functionality of AVs.

**Figure 1:** InnovizTwo point cloud
This approach again means an easier integration and customization of LiDAR systems for vehicles used across applications like agriculture, construction and mining operations, public safety and defense as well as healthcare and medical services.

**RETROFITTING FARMING EQUIPMENT FOR AUTONOMOUS OPERATIONS**

The key to the widespread adoption and eventual commercialization of AVs may lie in the versatility of the technologies predominant in automated driving systems. The ability of autonomous driving technologies to be employed over a wide range of vehicles could pave the way for AV ecosystems. These technologies, which could be used not only for a tractor but also a miller or drone, could enable a single platform where the user could monitor its use.

BlueWhite, an Israeli-American technology company, demonstrates that versatility by providing AV solutions for agriculture. It develops retrofit kits that add autonomous operation to existing tractors and other farm equipment, enabling the creation of a network of AVs and the automation of their tasks. Like Innoviz’s InnovizTwo, BlueWhite’s technology intricately fuses advanced hardware and software components.

The BlueWhite Pathfinder retrofit kit, installed on any tractor, serves as the hardware backbone, imbuing vehicles with autonomous potential to bridge the reliability gap in AVs.

Integrating the perception software into the solution also contributes to the versatility of the sensor. The company configures scan patterns, resolution and other features of the sensor data via the software to meet the needs of different autonomy levels and use cases. "For instance, it can prioritize front-facing long-range detection for L3 highway driving or expand coverage to include sides and junctions for L4 applications," Hofstetter said.

---

**Figure 2: Levels of driving automation in SAE J3016 standards**

- **SAE LEVEL 0**: You are driving whenever these driver support features are engaged — even if your feet are off the pedals and you are not steering.
  - You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety.
- **SAE LEVEL 1**: You are driving when these automated driving features are engaged — even if you are seated in the driver’s seat.
  - When the feature requests, you must drive.
  - These automated driving features will not require you to take over driving.
- **SAE LEVEL 2**: These features are limited to providing warnings and momentary assistance.
  - Example Features: automatic emergency braking, blind spot warning, lane departure warning.
- **SAE LEVEL 3**: These features provide steering support to the driver.
  - Example Features: lane centering control.
- **SAE LEVEL 4**: These features provide steering and brake/acceleration support to the driver.
  - Example Features: traffic jam chauffeur, local driverless taxi.
- **SAE LEVEL 5**: These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met.
  - Example Features: same as level 4, but feature can drive everywhere in all conditions.

---

How Versatile Technologies Can Boost AV Adoption

The key to the widespread adoption and eventual commercialization of AVs may lie in the versatility of the technologies predominant in automated driving systems.
culture—from labor shortages to operational safety—while being compatible with any tractor brand or model. However, its solution can be employed only on small or family farms. For large-scale agricultural activities, it is far from being a replaceable solution. BlueWhite can retrofit autonomous driving only into simple vehicles and equipment that do not operate day and night. This means that BlueWhite can’t automate the operation of the huge farming equipment typically used on large farms.

VERSATILITY TO DRIVE AV ADOPTION

Although its technology can be extended to other applications, BlueWhite plans to address the challenges of evolving machine architecture to facilitate the adoption of autonomous operation, creating open APIs so that different tractors can work autonomously with different implements and ensuring full farm connectivity so that operational reports and relevant data can be shared.

“In our roadmap, BlueWhite plans to expand from perennial crops to other specialty crops, such as vegetables; livestock-feeding applications; and wherever advanced autonomy can provide a significant impact to farmers,” Ascher said.

Concerns about the reliability of AVs, especially L4 AVs, have prompted technology players in the AV industry to subject their solutions to rigorous testing. More versatile technologies can enable more data to be collected, perhaps speeding adoption of Level 4 AVs.

MULTI-LAYERED SAFETY MECHANISMS

Retrofitting autonomous driving systems has faced more stigma over safety than building an AV from scratch. This is particularly true for BlueWhite, as the company promises a fully automated network of farm equipment.

“At BlueWhite, we implement redundancies and a multi-safety-layer approach directly into our system architecture,” Alon Ascher, CBO of BlueWhite, told EE Times Europe. “For instance, to enhance the perception of our autonomous tractor, we employ sensor fusion. This approach combines data from multiple sensors, such as LiDAR, cameras, GNSS [global navigation satellite system] and INS [inertial navigation system], to comprehensively understand and visualize the surrounding environment.

“We also deploy advanced algorithms and policies to control vehicle behavior, ensuring safe driving even in GPS- and communication-denied environments,” Ascher said. “Additionally, we incorporate hardware-layer redundancy and a physical bumper in front of the machinery, along with ‘Safety Stop’ buttons on the machine for field operators.”

BlueWhite said it seeks to solve the problems of modern agriculture—from labor shortages to operational safety—while being compatible with any tractor brand or model. However, its solution can be employed only on small or family farms. For large-scale agricultural activities, it is far from being a replaceable solution. BlueWhite can retrofit autonomous driving only into simple vehicles and equipment that do not operate day and night. This means that BlueWhite can’t automate the operation of the huge farming equipment typically used on large farms.

VERSATILITY TO DRIVE AV ADOPTION

Although its technology can be extended to other applications, BlueWhite plans to address the challenges of evolving machine architecture to facilitate the adoption of autonomous operation, creating open APIs so that different tractors can work autonomously with different implements and ensuring full farm connectivity so that operational reports and relevant data can be shared.

“In our roadmap, BlueWhite plans to expand from perennial crops to other specialty crops, such as vegetables; livestock-feeding applications; and wherever advanced autonomy can provide a significant impact to farmers,” Ascher said.

Concerns about the reliability of AVs, especially L4 AVs, have prompted technology players in the AV industry to subject their solutions to rigorous testing. More versatile technologies can enable more data to be collected, perhaps speeding adoption of Level 4 AVs.

Saumitra Jagdale is a contributing writer for EE Times Europe.
The need to integrate these greater numbers of image sensors, with ever-higher resolutions, creates multiple design challenges, including connecting multi-gigabit image data streams to application processors and processing the huge volume of raw image data generated by the sensors. This number is expected to go up significantly to enable higher levels of automation, potentially requiring as many as 45 image and other sensors to deliver the most advanced SAE Level 3 (and beyond) features. As the number of image sensors in vehicles has increased, the MIPI Camera Serial Interface 2 (CSI-2) standardized image sensor protocol has become the core imaging protocol used within automotive systems.

Automotive image sensors, which include cameras, LiDAR and radar, play a critical role in advanced driver-assistance systems (ADAS) and autonomous driving. Today, the systems that deliver advanced SAE Level 2 features typically utilize a combination of up to 10 camera and radar sensors. This number is expected to go up significantly to enable higher levels of automation, potentially requiring as many as 45 image and other sensors to deliver the most advanced SAE Level 3 (and beyond) features. As the number of image sensors in vehicles has increased, the MIPI Camera Serial Interface 2 (CSI-2) standardized image sensor protocol has become the core imaging protocol used within automotive systems.

To support these evolving electrical and electronic (E/E) architectures that leverage the latest high-speed automotive networking protocols.

To enable high-speed, software-oriented architectures, the automotive industry is evolving away from the domain-specific architectures of today and toward cross-domain centralized architectures. These centralized architectures use just a few powerful computers compared with the high number of individual electronic control units (ECUs) used in today’s domain-specific architectures. Several powerful vehicle computers will connect to all the vehicle image sensors via a small number of zonal ECUs, not only enabling software-defined vehicles but also allowing wiring harnesses to be simplified for weight and cost savings.

To support these evolving E/E architectures, MIPI developed MIPI Automotive SerDes Solutions (MASS), a framework for connecting sensors and displays over long cable lengths with functional safety (FuSa) and security built in. Since its initial release in 2020, several components within the MASS framework have evolved, and in 2024, major advancements include the addition of image sensor security functionality, an even faster MIPI A-PHY v2.0 serializer/deserializer (SerDes) physical layer and the further development of an A-PHY compliance program.

**COMPONENTS OF THE MASS FRAMEWORK**

Representing a collaborative cross-industry effort to meet the needs of the automotive industry, MASS consists of four principal components that, when combined, create an end-to-end high-speed connectivity framework for sensors and displays:

- The reuse of widely adopted higher-layer protocols: MASS leverages a suite of higher-layer application protocols used in billions of devices and already widely adopted in automotive. These
protocols include CSI-2 for image sensors; MIPI DSI-2 and VESA eDP/DP for displays; and lower-speed protocols for command and control, such as I2C, GPIO, Ethernet, SPI and the emerging MIPI I3C protocol. Using these proven protocols drives economies of scale, reduces NRE/development costs and provides backward and forward compatibility.

• **SerDes physical layer:** MASS is built on MIPI A-PHY, the first industry-standard, long-reach, asymmetric, SerDes physical layer interface with high noise immunity. A-PHY is designed to meet the needs of the automotive industry and eliminates the need for proprietary asymmetric PHYs and bridges, simplifying in-vehicle communication networks, and reducing costs, cable harness weight and development time.

• **Functional safety:** MASS standardizes features to help applications meet the FuSa requirements of ISO 26262:2018 “Road vehicles — Functional safety” and enable designers to build systems that meet common Automotive Safety Integrity Level (ASIL) specifications, from ASIL-B through ASIL-D.

• **Security:** MASS leverages MIPI’s camera security framework for protecting CSI-2 image sensor data. The framework enables key security functionality, including the authentication of imaging system components, data integrity protection and data encryption. MASS also supports High-Bandwidth Digital Content Protection for display applications.

### MIPI CSI-2 FEATURES FOR AUTOMOTIVE IMAGING SYSTEMS

The CSI-2 protocol connects image sensors to application processors and includes features that benefit automotive imaging systems, enabling sensor aggregation optimization, superior objective image quality and energy-consumption reduction. Provisions within CSI-2 also alleviate RF emissions and support “region of interest” extraction, always-on inferencing, wire reduction, longer-reach connectivity and the reduction of current leakage. Features that particularly

---

**MIPI CSI-2 Stack**

**MIPI A-PHY Performance**

<table>
<thead>
<tr>
<th>Downlink Gear Data Rate</th>
<th>Modulation</th>
<th>Modulation Bandwidth (GHz)</th>
<th>Max Net App Data Rate (Gbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 3 Gbps</td>
<td>NRZ-8B/10B</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>PAM4 (Optional)</td>
<td>0.5</td>
<td>1.8</td>
</tr>
<tr>
<td>G2 4 Gbps</td>
<td>NRZ-8B/10B</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>PAM4 (Optional)</td>
<td>3.6</td>
<td>2</td>
</tr>
<tr>
<td>G3 6 Gbps</td>
<td>PAM4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>G4 12 Gbps</td>
<td>PAM8</td>
<td>2</td>
<td>10.8</td>
</tr>
<tr>
<td>G5 16 Gbps</td>
<td>PAM16</td>
<td>2</td>
<td>14.4</td>
</tr>
<tr>
<td>G6 24 Gbps</td>
<td>PAM8</td>
<td>4</td>
<td>21.6</td>
</tr>
<tr>
<td>G7 32 Gbps</td>
<td>PAM16</td>
<td>4</td>
<td>28.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Uplink Gear Data Rate</th>
<th>Modulation</th>
<th>Modulation Bandwidths (MHz)</th>
<th>Max Net App Data Rate (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GB 100 Mbps</td>
<td>NRZ-8B/10B</td>
<td>50</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>PAM4 8B/10B</td>
<td>125</td>
<td>100</td>
</tr>
<tr>
<td>200 Mbps</td>
<td>PAM4 8B/10B</td>
<td>400</td>
<td>1100</td>
</tr>
</tbody>
</table>

**MIPI A-PHY SerDes performance table**
New Developments in MIPI’s High-Speed Automotive Sensor Connectivity Framework

benefit vehicle imaging systems include:

- **Always-on sentinel conduit**: Enables always-on vision systems in which combinations of ultra-low-power image sensors and video digital-signal processors continuously monitor their surrounding environments and wake their higher-power host ECU only when significant events happen.

- **Multi-pixel compression**: Provides optimized pixel compression for the latest generation of ultra-high-resolution image sensors with multi-pixel color filter arrays.

- **RAW-28 color-depth pixel encoding**: Supports both higher image quality and signal-to-noise ratio in high-dynamic-range automotive image sensors.

- **Smart region of interest**: Enables more efficient analysis of images using AI-based vision inferencing algorithms.

The use of CSI-2 within automotive applications is augmented with MIPI Camera Service Extensions (CSE), which provides functional safety and security, and the MIPI Camera Command Set (CCS), which provides command and control.

**HIGHER UPLINK AND DOWNLINK SPEEDS IN THE NEXT A-PHY SERDES UPDATE**
The A-PHY physical layer interface addresses the need for a standardized long-reach, highly reliable, asymmetric SerDes solution to simplify the integration of image sensors and displays into automotive E/E architectures. It addresses key features of A-PHY:

- Downlink data rates of up to 32 Gbps in A-PHY v1.1.1, with up to 64 Gbps supported in A-PHY v2.0 (expected Q2 2024)
- Uplink data rates of up to 200 Mbps in A-PHY v1.1.1, with up to 1.6 Gbps supported in A-PHY v2.0
- High reliability, with an ultra-low packet error rate of <10⁻¹⁸ over the lifetime of a vehicle
- High resiliency, with ultra-high immunity to automotive EMI effects
- Bounded low latency (maximum 6 µs)
- Support for multiple cable types (coaxial, shielded differential pair and star quad)
- Long reach (up to 15 meters in length with four inline connectors)
- Multiple power-over-cable options, including support for 48-V operation with the Power over A-PHY v1.0 specification

IN THE NEXT A-PHY SERDES UPDATE
- Higher uplink and downlink speeds
- Support for multiple higher-layer protocols through its generic data-link layer and a set of protocol adaptation layers (PALs) that map these protocols to A-PHY’s A-Packet format.
- In addition to PALs that support native coupling to CSI-2 and DSII-2, PALs have been developed for several lower-bandwidth control interfaces, including I²C, GPIO, Ethernet, SPI and the emerging I3C interface.

END-TO-END, APPLICATION LAYER, FUNCTIONAL SAFETY AND SECURITY FOR AUTOMOTIVE SYSTEMS

The MIPI camera security is effective in both the command and control (CSE) layer and the payload layer. CSE v2.0 (May 2024) includes MIPI Camera Security Extensions (CSE) that add security functionality to the MIPI Camera Interface (CSI). CSE includes functional safety and security for imaging applications. FuSa functions were included in CSE v1.0, and security functions were added in the recently introduced CSE v2.0 (May 2024). Services are implemented using the Service Extension Packet (SEP) or Frame-Based Service Extension Data (FSED) protocol, providing flexibility, particularly when implementing CSE services within legacy solutions.

CSE FuSa services include message counters, cyclic redundancy checks, built-in self-tests with error-injection capabilities and a baseline timeout-monitoring service to detect loss of communication.

CSE security services enable authentication, integrity protection and (optional) encryption of image sensor data. They provide a high degree of flexibility to balance the required security levels against processing efficiency, thermal regulation and power consumption. Other features include:

- **Choice of cipher suites**: Options include an "efficiency" cipher suite providing AES-CMAC data integrity only (no encryption) targeted toward sensors with limited hardware resources, as well as a "performance" cipher suite that provides AES-GMAC data integrity and optional AES-CTR encryption aimed at sensors with dedicated hardware support.

- **Choice of tag modes**: These multiple tag mode options allow the implementer to choose how often the tag is computed and transmitted.

- **Granular security controls**: These controls provide highly granular "source selective" control over the different segments of the CSI-2 image frame to enable a "sliding scale" of security levels.

MIPI camera security is effective in both its security extent and implementation flexibility. When both security and FuSa service extensions are enabled, security is layered on top of functional safety; from a source (or transmitter) perspective, security is applied to the image data first, followed by functional safety.

Ariel Lasry is vice chair of the MIPI A-PHY working group at the MIPI Alliance.
Ethernet Connectivity Drives Zonal Architectures

By Robert Huntley

Here’s nothing as constant as change, so the saying goes. That is undoubtedly the case for the automotive industry as vehicle manufacturers experience significant pressure to transform their operations. The drivers of change are many. Adding hybrid and battery-electric platforms to the lineup stresses engineers with weight, range and cost challenges. At the same time, they continue to deliver innovations and personalization enhancements to current models. The electrical/electronic (E/E) architecture is the most intense of all the areas undergoing significant change.

E/E ARCHITECTURES ADAPTING TO CHANGE

Vehicle electrification, semi and full autonomy, and enhanced connectivity services have established and benefited from two interlinked megatrends: the zonal E/E architecture and the rise of the software-defined vehicle (SDV).

The zonal E/E architecture has been in the cards for some time. It represents an evolution from a legacy domain approach to cabling that connects similar functions, such as chassis and body/comfort, to those based on physical location. By using zone-based electronic control units (ECUs), there is an opportunity to significantly lower the weight of the cable harness.

Complementing the architectural changes, an SDV employs a service-oriented approach, decoupling hardware devices from the software. Implementing a robust, resilient and fit-for-purpose network throughout the car is paramount for these megatrends to succeed.

Ethernet has made significant advances and already provides a gigabit backbone connecting ECUs. Ethernet is a trusted and reliable network protocol; however, routing a thick, four-twisted-pair cable to every sensor node goes against the initiatives to reduce vehicle weight and simplify E/E architectures. One of Ethernet’s recent advances is the development of single-pair Ethernet, and for automotive applications, 10BASE-T1S is rapidly gaining adoption.

10BASE-T1S ETHERNET CONNECTIVITY TO SENSORS AND ACTUATORS

The 10BASE-T1S specification is part of the IEEE 802.3cg single-pair Ethernet standard and offers 10-Mbps networking. What makes 10BASE-T1S particularly relevant for automotive applications is its multidrop topology, a feature that allows multiple nodes to be connected to the host over a single unshielded twisted pair. It supports at least eight nodes and bus lengths up to 25 meters. The protocol provides shared network usage by implementing a physical-layer collision-avoidance technique, and the maximum latency is dictated by the number of nodes and the amount of data to transfer.

EE Times Europe spoke with Fionn Hurley, marketing director of the automotive networking group at Analog Devices Inc., to find out why 10BASE-T1S looks set to become the missing link for automotive Ethernet-based connectivity.

RAPID INCREASE IN ZONAL ARCHITECTURE ADOPTION

EE Times Europe asked Hurley if implementing zonal architectures had been a long time coming or if there was a strong push for adoption. “The first time we started hearing about zonal architectures was around 2018–2019,” he said. “It’s been a key topic of conversation at ADI for the last few years, and we’re already seeing a pretty rapid rise in the push toward zonal architectures, with the first steps already taken by some of the leading automotive OEMs. We’re seeing some domains move toward zone-based before other domains—sort of a hybrid mix of architecture. Based on what we’re seeing today, body/chassis and powertrain domains will move first, while leaving other domains, such as infotainment and safety systems, to follow.”
Ethnic Connectivity Drives Zonal Architectures

10BASE-T1S multidrop, shared network capability (Source: Analog Devices Inc.)

as ADAS and infotainment, where there is a need for high-bandwidth connections, to use another architecture.

REDUCING CABLES, REDUCING WEIGHT
Hurley said that one of the driving forces toward adopting a zonal architecture is to reduce the vehicle's weight. "The cable harness, typically representing several kilometers, is a significant proportion of a vehicle's weight. The location of the zones is another consideration. The doors, for example, are a good location to gather together the sensors and actuators. Another broader aspect of zonal architectures is to reduce the number of ECUs. In the past, the number of ECUs in a vehicle rose exponentially as new functions were added. Now, there is a desire to reduce them, providing cost savings for the OEM and maintenance costs for the customer."

SHORTENING CYCLE TIMES
EE Times Europe asked whether the cycle times for bringing new vehicle platforms changed as these megatrends took hold.
"We are seeing car platforms refreshed more frequently," Hurley said. "Fifteen years ago, it was probably a six-year cycle; now, it's significantly shortened and continuing to reduce. New entrants into the market are also forcing some of the traditional OEMs to accelerate the refresh of their vehicles. There is a desire across the industry to update platforms, and in the interim, some platform functions are updatable with over-the-air updates, bringing the latest features to customers as early as possible."

10BASE-T1S ADOPTION ESCALATING
"Over the past two years, we've seen a step up in engagement and rollout of 10BASE-T1S in automotive," Hurley said. "If you look at the topics discussed at this year's Automotive Ethernet Congress, I'd say that between 50% and 75% of the papers had a focus on 10BASE-T1S. Three years ago, there was probably only one paper about it, so it's really taken off. 10BASE-T1S fits very well with the transition to zonal-based architectures, and Ethernet to the edge also makes a lot of sense. Over the years, we've all worked on technologies that were either too early or too late—10BASE-T1S really seems to intersect the transition to zone-based architecture at just the right time. 10BASE-T1S is a growing market, and we're working with OEMs across the globe."

MULTIDROP: A KEY FEATURE OF 10BASE-T1S
EE Times Europe asked Hurley whether multidrop is paramount to the success of 10BASE-T1S automotive adoption. "Nearly every customer we talk to and every application we work on uses multidrop," he said. "It really helps its adoption and with reducing the cable harness weight challenge. Ethernet has been used in automotive since about 2013 and has a suite of tools, has time-sensitive networking to give you determinism and comes with security and safety capabilities. Although the physical layer is new, and it's multidrop, you can layer it on top of the existing suite of tools to yield an effective solution."

CONNECTOR FLEXIBILITY, POWER DELIVERY CAPABILITY
Hurley noted that the IEEE specifies no connector for 10BASE-T1S. "The IEEE typically shies away from defining the cable and connector types, which allows OEMs to continue using the popular connector methods already in use. Unshielded and unjacketed cables are becoming the norm for 10BASE-T1S to make it an economically viable solution. Using an unjacketed pair opens options for building your cable harness and simplifies implementation. Customers can use their existing multifunction connectors rather than having to design in specialist connectors and through-bulkhead fixtures, significantly aiding 10BASE-T1S's adoption."

EE Times Europe asked Hurley whether automotive customers will implement 10BASE-T1S's power-over-data-line (PoDL) power delivery capability. "Even though power delivery has been in the 802.3bu specification since 2016, Ethernet's automotive deployment has initially been from ECU to ECU, so PoDL doesn't fit very well there due to the high power consumption requirements. However, connecting to edge sensors and actuators tends to be lower-power, so now we see that PoDL does bring value, so it's desirable and beneficial. It's something we will definitely see the rollout of in the not-too-distant future."

The zonal E/E architecture represents an evolution from a legacy domain approach to cabling that connects similar functions, such as chassis and body/comfort, to those based on physical location.

TAKING 10BASE-T1S ONE STEP FURTHER
Hurley told EE Times Europe that ADI had taken 10BASE-T1S one step forward with its Ethernet to the Edge Bus (E2B) technology. "E2B aims to simplify the design process by removing the need for microcontrollers in edge nodes for 10BASE-T1S applications and fits well into a fully Ethernet automotive architecture. Earlier this year, we announced our collaboration with BMW Group to implement our E2B 10BASE-T1S technology for the ambient lighting system in future vehicles. E2B removes microcontrollers from edge sensors and actuators, moving the software to an ECU, greatly simplifying software development and qualification tasks."
The automotive user experience has evolved dramatically, and the trend is accelerating, driven by advances in vehicle connectivity and autonomy. Connectivity makes it possible for the car to be safer and more aware, thanks to vehicle-to-everything technologies. Autonomy enables the car to behave more intelligently, thanks to advanced driver-assistance systems (ADAS).

The downside of these technology enhancements is the increased exposure to attacks. Connectivity can be leveraged as an attack vector, and autonomy can drastically augment an attack’s impact. In practice, the scary scenario is that of a remote attacker infecting a car through its wireless interface and taking control of the vehicle by abusing its ADAS capabilities.

Trust in hardware and software is key in the realm of connected vehicles. Security-by-design principles must be ingrained into the very foundation of electronic control units (ECUs). Safety and security considerations are pivotal, with tailored security levels adapted to each ECU. This approach ensures that every component of the vehicle’s hardware contributes to its overall security posture, laying a solid foundation for trustworthy automotive systems.

Imagine a scenario in which a hostile state or a rogue hacker group gains access to the control systems of connected vehicles. The consequences could be catastrophic: widespread accidents, traffic gridlock and even deliberate targeting of individuals or critical infrastructure.

**THE CHALLENGES OF CONNECTED-VEHICLE CYBERSECURITY**

Ensuring vehicle cybersecurity is a multifaceted challenge that encompasses aspects of software development, cryptography, operating systems and life-cycle management. One significant hurdle lies in implementing safe and secure coding practices tailored for automotive environments. Adherence to stringent standards helps to ensure that the software running in connected vehicles is robust and resilient against potential attacks and exploits.

Integrating artificial intelligence into automotive systems complicates cybersecurity efforts. AI-driven functionalities, such as autonomous driving and ADAS capabilities, introduce new attack vectors and complexities in securing connected vehicles. Essential protections, including secure boot mechanisms, communication protocols, fault and intrusion-detection mechanisms and memory protection, are paramount to safeguarding AI-driven functions from exploitation by malicious actors.

The adoption of post-quantum cryptography will likewise be crucial for safeguarding communications and data exchange within vehicles in the near future. With the advent of quantum computing, some of the traditional cryptographic algorithms will become vulnerable to brute-force attacks, necessitating the deployment of quantum-resistant cryptographic algorithms to maintain vehicle security.

Secure operating systems are required that enable secure applications to run in an isolated manner within software-defined vehicles. By compartmentalizing critical functions and services, these operating systems mitigate the risk of unauthorized access and tampering. Managing the security life cycle of connected vehicles is also fundamental to ensure that security measures are maintained from IC manufacturing to decommissioning.

The commitment to cybersecurity doesn’t end with the development and deployment of vehicles; it extends throughout their life cycle. ISO/SAE 21434 and the European Cyber Resilience Act underscore the importance of life-cycle cybersecurity management. Such approaches to cybersecurity ensure that
Ensuring vehicle cybersecurity is a multifaceted challenge that encompasses aspects of software development, cryptography, operating systems and life-cycle management.

Security certification is mandatory in building trust at every level of the autonomous-vehicle ecosystem. Each component must undergo rigorous certification processes to validate its security measures and ensure its trustworthiness. From individual ECUs to complex vehicle systems, robust security certification protocols are necessary for instilling confidence in the hardware components that power self-driving capabilities. This includes obtaining ISO 26262 ASIL-D certification, implementing robust security measures and adhering to industry standards to fortify the resilience of automotive systems against cyberthreats throughout their operational lifespan.

COLLABORATIVE EFFORTS TO SECURE THE FUTURE OF MOBILITY

In addressing the challenges of securing autonomous vehicles, embedded cybersecurity emerges as an indispensable cornerstone of automotive innovation. Through a concerted effort to prioritize cybersecurity measures, adopt best practices and deploy advanced solutions for cyberthreat detection and monitoring, we can fully embrace the transformative potential of connected vehicles while effectively mitigating risk.

However, the responsibility for ensuring robustness in autonomous vehicles extends beyond manufacturers and developers. Collaboration among governments, academia and industry stakeholders is essential to anticipate emerging threats and mitigate risk effectively. Moreover, public awareness and education campaigns are needed to foster a culture of cybersecurity consciousness among users of connected vehicles.

Collaborative security management across the value chain is crucial in securing connected vehicles. It requires collaboration and accountability on the part of manufacturers, suppliers, service providers and regulators. Each player must attest to its commitment to security management and implement robust measures accordingly. By working together to uphold cybersecurity standards and practices, the automotive industry can build a foundation of trust and reliability in vehicle technology.

By addressing these challenges head-on and fostering collaboration across borders, we can ensure that the promise of autonomous vehicles is not overshadowed by the dark cloud of cyber insecurity. As we navigate the road ahead, let us ensure that embedded cybersecurity serves as the bedrock upon which a safer, more secure automotive future is built.

Moreover, this new paradigm paves the way for new services and business models to monetize security. While there are indeed challenges, there is also a space of opportunities awaiting exploration and innovation.

Hassan Triqui is co-founder, CEO and president of Secure-IC.
CHIPLETS: BUILDING THE FUTURE OF SoCs

Chiplets: Building the Future of SoCs is a virtual event exploring the innovative world of chiplet technology. Discover how chiplets, the building blocks of heterogeneous multi-die systems, are shaping the future of high-performance computing across industries, driven by AI advancements. Dive deep into discussions covering the entire chiplet ecosystem, from design to testing, and explore initiatives like chiplet marketplaces and standards. Key topics include:

- Chiplet definitions, standards, and technologies
- Integrating multiple dies
- Ensuring quality and security
- Real-world implementations
- Assessing economic viability
- The chiplet ecosystem
- And more!

DON’T MISS THIS OPPORTUNITY TO EXPLORE THE POTENTIAL AND CHALLENGES OF CHIPLET-BASED SYSTEMS!

REGISTER NOW

chiplets.eetimes.com

For more information about sponsorship opportunities, please contact sales@aspencore.com

Embedded.com | EETimes
Certifying RISC-V: Industry Moves to Achieve RISC-V Core Quality

By Dave Kelf, Breker Verification Systems

At the RISC-V Summit North America in November 2023, Simon Davidmann, CEO of Imperas (now Synopsys), delivered a surprising keynote. His message: While RISC-V was an enormous leap forward, the limited focus on quality could ultimately kill the initiative. This was a view shared by many stakeholders.

Davidmann went on to explain that a quality expectation had been set across the semiconductor industry by companies like Arm and Intel. Bugs in their processors have been extremely rare, and users rely on that quality level, as a processor bug could kill an end product. Davidmann noted that Arm spends more than US$150 million on verification annually, runs a staggering $10^{15}$ (greater than the number of miles in a light year) verification clock cycles per core and has 30 years of experience under its belt. This is made easier by retaining complete control of the instruction set.

RISC-V users expect this “Arm quality” level, because without it, their designs will fail. But how can RISC-V core providers hope to meet this quality goal, and how can they prove it?

RISC-V: NEW PROCESSOR THINKING

Before the advent of RISC-V, most processor instruction set architectures (ISAs)—the fundamental code specifications used to drive the device—were the protected intellectual property of the company that developed the processor. This protection was critical to the company, as it allowed ecosystem control and blocked competition, enabling favorable business models.

Along came the RISC-V open ISA. Any company could use the instruction set for its processor implementation, software stack, tools and other ecosystem elements. Business models would no longer be dictated by the ISA copyright. In addition, the RISC-V ISA was more flexible than others, allowing the inclusion of differentiating custom instructions.

The semiconductor industry was enthusiastic. Coincident, unpopular business model changes from Arm drove companies to start RISC-V initiatives, and RISC-V gained considerable momentum. BCC Research estimated the market for RISC-V technology in 2021 at US$500 million, with an expected growth rate of 33%, to US$2.7 billion by 2027. Today, RISC-V International, the ISA governing body, reports 3,950 members in 70 countries, with 13 billion RISC-V IP cores on the market.

Arguably, one of the most important RISC-V market characteristics thus far is that most of those 13 billion processors have been relatively small, embedded devices. But we are now seeing the advent of larger, multicore application processors in complex systems, with an inevitable impact on verification solutions, given the processors’ far greater complexity.

In response, RISC-V International’s board has launched a certification committee to address the need for a provable quality benchmark for RISC-V.

VERIFYING A PROCESSOR

Processor verification is a complex business with too many facets to cover in this short article. However, even in the case of a simple, single-core embedded device, a large series of instructions in an almost-infinite range of combinations can be applied to make the device react in many different ways.

Intelligent, random instruction generators are often deployed for a more superficial check of the architecture. Checking that the processor does the right thing, based on the instruction stream applied, requires the use of either self-checking tests or a golden model of a RISC-V processor for comparison, both of which are hard to produce. The instruction

A typical RISC-V verification stack is a complex undertaking that needs a comprehensive test plan. (Source: Breker Verification Systems)
type and combinations are important. Checking branch prediction, verifying instruction privilege levels and ensuring exceptions are handled correctly are just a few examples.

Another aspect of instruction testing is to ensure compliance with the official ISA specification. Various tests are available, but they achieve only some of this architectural test requirement.

Instruction-level, architectural testing is just the start. Up to 80% of processor execution consists of load-store operations, and these must be efficient. The same applies to the interrupt mechanisms and many other interfacing requirements. Functional operation and performance-related issues must be checked together, as read-write hazards can cause unexpected problems that prove difficult to debug. Such a concerted effort requires in-depth, microarchitectural testing, which in turn requires test synthesis that combines different test sets to weed out complex, unpredictable corner cases.

**RISC-V International’s board recognized the need for a certification program to provide processor suppliers with an independent assessment of the performance of their devices.**

The core should also be tested in a system-on-chip. Depending on the core—or multicore—complexity, we start to see coherency issues, security vulnerabilities, complex interrupts and many more.

This is extreme verification, and it is what drives processor companies like Arm to invest so much. It is unreasonable to expect a small team to build such a verification environment. The only way to achieve an appropriate level of verification by smaller teams is dramatic test content reuse.

One advantage of a common ISA is that it enables multiple teams to combine forces by combining test suites. A vehicle for such a common reusable test set is a system-level verification intellectual property component (SystemVIP) targeting RISC-V. With this approach, a commercial or other organization would build a common test set by leveraging economies of scale, spreading the investment across many core developments. This is the purpose of commercial verification companies, as well as open-source projects by corporations that include Google.

Of course, such cooperation is difficult between commercial, competitive entities. However, a common organization can propel such an initiative, and RISC-V International is doing just that.

**RISC-V CERTIFICATION**

RISC-V International and other organizations have already implemented a RISC-V ISA compliance test suite. While the suite is a useful start, it is in no way comprehensive enough to be solely relied upon. RISC-V International’s board recognized the need to set up a rigorous certification program that would provide processor suppliers with an independent assessment of the performance of their devices, yielding information on which their customers could rely.

Certification programs exist for other industry standards, such as Wi-Fi, USB and PCIe. While these interface protocols require extensive test suites, their verification is simpler than verifying a full processor core. Arm and Intel undoubtedly have internal sign-off processes for their new cores, but these are proprietary. As such, no public program of this nature to certify a device as complex as a processor core that can be applied to a range of core architectures has been created before, and it is a daunting challenge.

The RISC-V board recently formed the RISC-V Certification Steering Committee (CSC) to tackle the task. Members include commercial RISC-V users, processor providers, test suite providers and other experts in this area. The committee is new but is moving aggressively to come up with a concrete plan that addresses what tests might be used, how the program would operate, to what degree the tests would check overall architecture and other definitional questions. By the end of the year, the CSC should have a comprehensive plan and will be well on its way to implementing it.

The RISC-V board fully expects the CSC to remove a critical potential barrier to the full deployment of RISC-V processors by providing a quality stamp of approval for the most complex devices that will be accepted across the semiconductor industry. This is no mean feat, and it will require a broad range of expertise to achieve verification levels not previously seen in the public domain. But it is absolutely necessary to achieve the same quality expectation for RISC-V cores as we observe for its commercial counterparts.

RISC-V is about to get a whole lot more reliable.

**REFERENCE**

1 BCC Research (December 2022). “RISC-V Technology: Global Market Outlook.” tinyurl.com/4teh82y3

Dave Kelf is CEO of Breker Verification Systems.
Overcoming Unbalanced Training Data for Safer Autonomous Driving

By Pat Brans

According to a Sweden-based expert on autonomous systems, self-driving cars won’t be available for at least another 10 years—and part of the holdup is data.

“Many of the remaining challenges standing in the way of fully autonomous vehicles have to do with the quality of data used to train the neural networks that control a vehicle,” said Michael Felsberg, full professor and head of the Computer Vision Laboratory at Sweden’s Linköping University. To ensure that AVs react appropriately to real-world road conditions and events, researchers are working on ways to fill the gaps in training data and correct for biases in the datasets, he told EE Times Europe.

Felsberg serves as a member of the Wallenberg AI, Autonomous Systems and Software Program (WASP) executive committee, representing Linköping University. He also collaborates with industrial players, including car and truck manufacturers and companies that produce support systems for vehicles.

Recording the data needed for training is expensive—and so is labeling it, largely because the task still requires human intervention. According to Felsberg, to bring down costs, the role of humans in the labeling process must be minimized through some form of weakly supervised learning whereby labels are assigned automatically or at least semiautomatically. But while many academic researchers and industrial players have been experimenting with weakly supervised learning, thus far, none of the methods are ready for widescale use.

The monetary cost of collecting and labeling data is not the only issue. An even bigger obstacle to reliable self-driving cars has to do with ensuring that the data used to train the neural networks will get the vehicle to do the right things. Not only is it impossible for AV manufacturers to collect enough data to cover all conceivable situations, but the data they do collect is likely to include biases that, if left uncorrected, can produce undesired behavior.

Compensating for Underrepresented Scenarios

Many of the most dangerous driving situations involve circumstances so rare that they are unlikely to be fully represented in real-world training data. No AV developer can expect to record enough real-world images and videos to cover all the different ways a pedestrian might run out in front of a vehicle, for example.

“One strategy for dealing with underrepresented data is to augment the recorded data with slight changes, such as geometric or chromatic variations, [made] directly in the training data,” Felsberg said. “But I would go so far as to say that it’s better to adjust the bias of the classifiers that result from the training than to apply fake data to the training process.”

Some people have floated the idea of using generative AI to produce supplemental training data for scenarios that are not close enough to the real data to be represented by slight manipulations. Felsberg thinks this approach would be catastrophic, however, given the propensity of generative AI to create absurd representations of the real world. When unrealistic data is used to train an autonomous system, the resulting network becomes unpredictable.

A better method, according to Felsberg, would be to move toward explainable AI—internal representations that reflect real scenarios by combining expert models (constructed from what human experts think are real-world scenarios) with machine-learned models (constructed from data collected in the field). Instead of altering the data used to train the model, a highly skilled technician could analyze and possibly alter the internal representation directly as part of a new step in the training process.

“If you understand the scenarios and have the right tools to manipulate the model’s internal representations of external objects, you have a more powerful way of influencing the outcome,” Felsberg said. “You could place cars in slightly different positions [than in] the real data or make them drive in a different direction than they did in the real data, to model situations that never occurred in the dataset. Prototypes of this kind of system already exist, and we are starting to participate in that kind of research with our industrial partner Zenseact.”

Felsberg continued, “The combination of model-based knowledge and data-driven knowledge is a hybrid learning approach that is very popular in many domains where predictions are required—for example, making climate predictions. But we don’t yet know how to apply this concept to autonomous vehicles. We have made some first attempts, where we demonstrated the feasibility, but there’s still a lot of research to be done.

Some people argue that large language models and large multimodal models will be a big help in manipulating internal representations, and we are certainly considering that option.”

Correcting Unbalanced Datasets

Aside from filling in the gaps in training data, a remaining challenge lies in minimizing biases—or, more accurately, adjusting the biases in ways that produce desired outcomes. “Some cases are much more common than others, and you would like to have control over the bias induced by this effect,” Felsberg said. “That requires methods and tools for adjusting biases in a trained model.”

When training data is collected for pedestrians, for example, children or wheelchair users might be insufficiently accounted for because they are encountered less frequently than pedestrians in other categories. To ensure that an AV responds appropriately to all pedestrians—as any reasonable person would expect it to do—the system has to overcome skews in the data that might otherwise lead to skews in object recognition. “It’s impossible to have fully balanced datasets for whatever you want to do,” Felsberg said. “But you can measure the unbalance and have your system adjust accordingly.”

Felsberg has been asserting that self-driving cars are at least 10 years out ever since he began working on the technology in 2007, and it remains his prediction today. But when the day does come for AVs to be sold or rented to the general public, manufacturers should be required to demonstrate that their self-driving cars have overcome potential biases that result from unbalanced data, he said.

Felsberg proposes amending Euro NCAP—a safety rating system to help consumers select cars based on their reactions to a set of real-life accident scenarios—to include tests designed to do just that. “Those changes should be made now,” he said.

Pat Brans is a contributing writer for EE Times Europe.
Formula 1 fans are accustomed to watching two cars race wheel-to-wheel at breakneck speed, but when the Indy Autonomous Challenge (IAC) driverless race car competition delivered the same experience in Las Vegas earlier this year, the cars’ skilled maneuvers took even the race’s organizers by surprise. Without human intervention, a pair of Dallara-built Indy Lights race cars, filled with processors and sensors controlled via autonomous software, sped around turns, getting as close as 1.5 meters from each other before one overtook the other at some 135 mph.

IAC chief executive Paul Mitchell was thrilled. “Historically, our passes have involved a car moving out to overtake, passing and then immediately pulling back in,” he said. “But here the car [from Korea Advanced Institute of Science and Technology] started to overtake, ’realized’ it was in a turn [and] decided to wait until it was out of that turn to finish the overtake.

“We didn’t plan that, the teams didn’t plan that—it was the car’s reaction. This was a first for high-speed autonomous racing,” Mitchell said.

Since 2021, the Indy Autonomous Challenge has attracted university software developer teams from around the world to compete in fastest-lap, obstacle avoidance and head-to-head races. Each team has worked with the IAC’s Dallara AV-21 rear-wheel drive, one-seater race car, equipped with six mono cameras, four radars, three LiDARs and GPS. Researchers develop the autonomous software stack, including localization, object tracking, prediction and control software, alongside AI modules, to pilot the race car. Barring their software, the cars are exactly the same—so, put simply, the best “AI driver” wins.

Simon Hoffmann heads up TUM Autonomous Motorsport, the Technical University of Munich team, which developed the software for the defending car that took part in the remarkable wheel-to-wheel overtake in Las Vegas and went on to win the IAC, held during CES 2024. “It was pretty nice to watch our software do this [maneuver], but the entire race was interesting, as there was a lot of interaction between cars,” he said. “As well as that long overtake, the cars tried several times to get into a gap. That gap would then close and the overtake would be aborted, so the car would try it on the other side. We got some really nice data from all of this.”

And data—at least for the competitors—is largely what the IAC is about. While the races have drawn huge numbers of spectators, with winning teams bagging US$1 million prizes, the IAC maintains the Challenge is first and foremost an applied research program to advance autonomous vehicle technology, show what it can do at extreme speeds and drive commercialization of fully autonomous urban transport.

“We are not, and never will be, a motorsports racing series; we’re not
here to compete with Formula 1,” Mitchell asserted. “We are using this platform as a way to accelerate the testing and validation of hardware and software, and to give students a learning lab experience that they otherwise can’t get.”

**TRIALS AND TRIBULATIONS**

IAC progress has been impressive. Initial time trials between university teams took place at the Indianapolis Motor Speedway in 2021, with more time trials and two-car track racing following on other U.S. oval speedways in Las Vegas and in Fort Worth, Texas. “When you’re on an oval, you may never drop below 150 mph ... and path-planning, overtaking—these vehicle dynamics are extremely hard,” Mitchell said. “Can all systems operate and make decisions continuously? This is the challenge.”

Simulations, which include details of the track and car, have featured heavily during software development to test AI driver performance before the real race. And along the way, university teams have broken autonomous world record after autonomous world record, achieving highest land speed (192.2 mph) on the Kennedy Space Center runway as well as top track speed (180 mph) and fastest on-track head-to-head overtake (177 mph), both on oval tracks.

In June 2023, the IAC veered from its usual oval circuit onto a Formula 1 road track at the Monza Circuit in Italy for a single-vehicle time-trial competition. This was one of the biggest challenges yet, said professor Sergio Savaresi, a leader of the Politecnico di Milano–University of Alabama PoliMOVE team, which holds the land speed and other records and recently added Michigan State University to its roster.

An autonomous race car uses its localization software, which largely relies on real-time kinematic positioning with RTK-GPS sensors, to navigate safely as it pelts around the track. Savaresi and cohorts had mostly developed their localization software for the open-sky oval circuits, but on certain stretches of the winding Monza road track, with trees, overpasses and tunnels, GPS dropped out.

“Being GPS-denied on some parts of the Monza circuit was difficult, as it was like being blind,” Savaresi said. “It’s one thing to localize at 50 km per hour in an urban autonomous vehicle, but doing this at 300 km per hour is much more tough; any tiny error in localization can immediately take you off the track and into a wall.”

Hoffmann concurred. “To be fast, you need to take the corners sharply, and so you need to know within a couple of centimeters where you are on the track. With the GPS [at Monza], sometimes the car didn’t know where it was within 10 meters,” he said.

To counter the intermittent GPS, university teams revised their algorithms to use a combination of data from RTK-GPS as well as additional cameras and sensors, including LiDAR, while training the software on digital maps of the racetrack environment. “This time, we were also relying on visual localization and object recognition of lines at the curb, trees, walls—essentially developing map-matching [algorithms],” Savaresi said.

In the end, each team conquered the road track. PoliMOVE scored the fastest lap, reaching a top speed of 169.8 mph—a new world record for autonomous speed on a racetrack. (Source: Business Wire)

PoliMOVE passes TUM during the competition at the Las Vegas Motor Speedway in 2022. (Source: Indy Autonomous Challenge)

The PoliMOVE team won the second annual Autonomous Challenge at CES in Las Vegas in 2023, reaching a top speed of 180 mph—a new world record for autonomous speed on a racetrack. (Source: Business Wire)
Las Vegas track at night to demonstrate the cars’ capabilities.

Working with industry players, IAC engineers drastically revamped the vehicle and designed a digital-twin simulation model. The new platform comes with 360° long-range LiDAR, advanced 4D radar, updated wireless communications and GPS, and a drive-by-wire system with independent actuation for front and rear brakes. Industry partners on the project included Japanese tire giant Bridgestone, U.S. communications heavyweight Cisco, U.S. LiDAR developer Luminar, Italian automotive tech company Marelli and German software developer dSPACE.

“We’d found that the actuators we’d had for oval racing were getting pushed to their limit on the road courses, so we put in more powerful actuation, and braking bias with front and back braking,” Mitchell said.

“The new actuation drive-by-wire system is tighter, for more precision braking and steering.”

Teams are now adjusting to these changes with the next challenge being an autonomous hill climb at the U.K.’s Goodwood Festival of Speed later this year, Mitchell said. He believes the precise control delivered by the AV-24 will eventually lead to more Formula 1-style dog-fighting, with sophisticated maneuvers and close encounters on the track—glimpses of which were seen with the Las Vegas overtake.

But is it realistic to expect an AI-driven autonomous race car will ever match the performance of a human-driven car? In a 45-minute “human versus AI” race at the A2RL event, former F1 driver Daniil Kvyat beat his autonomous competitor by just over 10 seconds.

PoliMOVE’s Savaresi reckons there’s room for improvement and predicts AI-drivers will match F1 world champion Max Verstappen’s ability “to drive and keep a car to its limit” in the next decade. Mitchell also believes autonomous race car technology will help the likes of Verstappen raise their fastest speeds from around 235 mph to perhaps as high as 300 mph.

But for Hoffmann, track interaction remains the real challenge.

“Formula 1 races have 20 cars on a track, and this is where we want to go. We’re moving in the right direction, but it’s a long road.”

Rebecca Pool is a contributing writer for EE Times Europe.

THE NEXT MOVE

The ability—or inability—to predict a nearby car’s next move has influenced autonomous racing, and competitor numbers, from the word go. The first IAC race at the Indianapolis Motor Speedway in 2021 was intended to be a 10-vehicle competition with cars running mere centimeters apart but was scaled back to a time trial, although two-car races soon followed.

A recent four-car race held by the Abu Dhabi Autonomous Racing League (A2RL) also suffered a few false starts. The competing teams were PoliMOVE; TUM Autonomous Motorsport; Unimore, from the University of Modena and Reggio Emilia; and Germany’s Constructor University. The cars didn’t reach their usual high speeds, and the race was punctuated by stops, restarts, bumps and spins. TUM eventually came in first; PoliMOVE failed to finish. Despite the setbacks, Hoffmann said the TUM team was “happy with the result” and expressed hope that the next race would “show the full potential of autonomous racing, with lots of overtaking maneuvers.”

Like the competitors in the single-car time trials at Monza, the A2RL teams grappled with GPS intermittency and localization issues. But a four-race presents different challenges than a time trial or two-car race. “There are situations where you have a car before you and behind you, and your car needs to take that into account when planning its decisions,” Hoffmann said. “This adds a lot of complexity.”

The complexities aside, Hoffmann believes putting more cars on the track is the future of autonomous racing. “Now that we’ve tried it, it is definitely the way to go,” he said.

Savaresi agreed. “The quantum leap in autonomous racing was from one to two cars, not two to four cars—this is an evolution,” he said.

“From a technology point of view, we are ready to run four cars.”

Mitchell confirmed the IAC would add more cars to its races, but he would not be drawn on when this could become the norm. “Our racing format will be defined by how we see our teams progress during testing this summer,” he said.

But more of these multicar races could arrive sooner than later, if this year’s CES launch of the IAC’s AV-24 next-generation platform is any indication. As part of the launch, three of the new cars circled the Las Vegas track at night to demonstrate the cars’ capabilities.

Working with industry players, IAC engineers drastically revamped the vehicle and designed a digital-twin simulation model. The new platform comes with 360° long-range LiDAR, advanced 4D radar, updated wireless communications and GPS, and a drive-by-wire system with independent actuation for front and rear brakes. Industry partners on the project included Japanese tire giant Bridgestone, U.S. communications heavyweight Cisco, U.S. LiDAR developer Luminar, Italian automotive tech company Marelli and German software developer dSPACE.

“We’d found that the actuators we’d had for oval racing were getting pushed to their limit on the road courses, so we put in more powerful actuation, and braking bias with front and back braking,” Mitchell said.

“The new actuation drive-by-wire system is tighter, for more precision braking and steering.”

Teams are now adjusting to these changes with the next challenge being an autonomous hill climb at the U.K.’s Goodwood Festival of Speed later this year, Mitchell said. He believes the precise control delivered by the AV-24 will eventually lead to more Formula 1-style dog-fighting, with sophisticated maneuvers and close encounters on the track—glimpses of which were seen with the Las Vegas overtake.

But is it realistic to expect an AI-driven autonomous race car will ever match the performance of a human-driven car? In a 45-minute “human versus AI” race at the A2RL event, former F1 driver Daniil Kvyat beat his autonomous competitor by just over 10 seconds.

PoliMOVE’s Savaresi reckons there’s room for improvement and predicts AI-drivers will match F1 world champion Max Verstappen’s ability “to drive and keep a car to its limit” in the next decade. Mitchell also believes autonomous race car technology will help the likes of Verstappen raise their fastest speeds from around 235 mph to perhaps as high as 300 mph.

But for Hoffmann, track interaction remains the real challenge.

“Formula 1 races have 20 cars on a track, and this is where we want to go. We’re moving in the right direction, but it’s a long road.”

Rebecca Pool is a contributing writer for EE Times Europe.
The Evolution of NVIDIA Drive

Drive is Nvidia’s computer platform for developing advanced driver-assistance systems (ADAS) and AVs. Nvidia Drive was introduced at CES 2015 and has grown through multiple generations. Table 1 summarizes Nvidia Drive’s evolution over the years.

THE EVOLUTION OF NVIDIA DRIVE

Drive is Nvidia’s computer platform for developing advanced driver-assistance systems (ADAS) and AVs. Nvidia Drive was introduced at CES 2015 and has grown through multiple generations. Table 1 summarizes Nvidia Drive’s evolution over the years.

What Is Nvidia Doing in Automotive?

By Egil Juliussen

Nvidia’s lucrative detour into AI, where it has established dominance, has yielded technologies that are steering its automotive business back into growth mode.

The first generations, Drive CX and Drive PX, used the Maxwell microarchitecture and focused on digital cockpits and ADAS applications. Even in 2015, there were 256 or 512 CUDA cores available that could be used for parallel compute operations.

The second-generation Drive PX2 targeted ADAS functions and was introduced in January 2016. Drive PX2 used the Pascal GPU architecture, and the number of Arm processors increased to 12 CPUs, all using 64-bit processors. Tesla used the Drive PX for its battery electric vehicle (BEV) Autopilot for several years.

Drive PX Xavier was introduced in January 2017 and used the Volta microarchitecture. Nvidia positioned it for use in L3 and L4 vehicles.

Drive PX Pegasus arrived in September 2017, was based on the Turing architecture and was Nvidia’s first automotive product with AI functionality. The Pegasus platform provided a performance increase of about 10x over Drive PX2. Availability was scheduled for mid-2018, and as of October 2017, Nvidia had more than 200 partners developing hardware and/or software products for its Drive platform.

The Nvidia GPU Technology Conference (GTC) has been the most important GPU conference for over a decade. Now, it may also be the most important AI hardware-software-deployment conference, as GPUs are driving much of AI training and inferencing activities.

At past GTCs, autonomous vehicles played an important role, as they were expected to be an early deployment segment for AI. The 2024 GTC presented some automotive content, but less than previous conferences. This year’s automotive content was mostly rolled into the presentations on Nvidia’s AI strategy, and in the keynote presentation, automotive AI was part of the robotics strategy. The most important automotive-related announcement was that the Blackwell GPU would be part of Nvidia’s Drive Thor centralized computer for safe and secure AVs.

This article offers perspectives on what Nvidia is doing in the automotive industry.
In December 2019, Nvidia introduced the Drive AGX Orin board family, and in May 2020, Nvidia announced Orin would use the Ampere architecture. The Drive Orin is still deployed for ADAS and L3–L4 vehicles and is likely to be in production for many years. Orin has up to 2 trillion CUDA cores, a level that enables parallel processing of complex AI models. The Drive Orin Ampere SoC has 17 billion transistors and meets ISO 26262 ASIL-D regulations.

Nvidia announced in April 2021 that the planned Drive Atlan would be based on the Ada Lovelace GPU architecture, but in September 2022, the company cancelled Drive Atlan and announced a replacement, Drive Thor. At GTC 2024, Nvidia reported that Drive Thor would use the Blackwell GPU architecture and the Arm Neoverse V3, a 64-bit CPU with up to 64 cores that was announced in February 2024.

**DRIVE THOR DEVELOPMENTS**

With its basis in Blackwell, Drive Thor represents a considerable technological advance over Drive Orin, which is based on the Ampere GPU architecture. The Blackwell GPU builds on the accumulated capabilities of three generations of Ampere and leverages a further four years of Nvidia’s AI experience. Drive Thor and Drive Orin are compared in Table 2. VSI Labs expects more details on Drive Thor to be available when the platform is ready for deployment.

The first row of the table compares the GPU capabilities of Blackwell and Ampere, as the improvements will directly advance Thor performance. Drive Thor has 12x more transistors than Drive Orin, resulting in more than 60x higher performance based on Nvidia’s own comparisons. The Blackwell calculations are 4-bit floating-point (FP4) arithmetic—much faster than Ampere’s 16-bit floating-point (FP16) calculations. FP4 calculation is a recent addition to Nvidia’s GPUs, as is FP8. FP8 and FP4 calculation accuracy is good for accelerating large language models (LLMs). Most of the millions to billions of parameters in LLMs can be expressed as FP8 or FP4 numbers, an ability that speeds up AI training and/or lowers power consumption.

All Blackwell products feature two reticle-limited dies connected by a 10-terabyte/second (TB/s) chip-to-chip interconnect in a unified single GPU. Deployment of Drive Thor vehicles is expected to start in 2025.

Blackwell adds reliability and resiliency with a dedicated reliability, availability and serviceability (RAS) engine to identify potential faults and minimize downtime. The RAS AI-powered predictive-management capabilities monitor thousands of data points across hardware and software to predict sources of potential vehicle safety issues and downtime. The RAS engine provides in-depth diagnostic data to identify areas of concern and plan for maintenance. By localizing the source of issues, the RAS engine reduces turnaround time and prevents potential vehicle safety problems that could result in crashes, injuries and fatalities.

The Transformer AI model is a neural network that learns the context of sequential data and generates new data. The AI model learns to understand and generate human-like text by analyzing patterns in large amounts of text data. The transformer AI model is a key factor in LLM growth. Blackwell and Drive Thor can leverage transformer technology to solve AV and similar automotive problems.

AV software and AI models require that vast amounts of data move from program and data memories to and from processors. Blackwell’s Decompression Engine can access large amounts of memory in the Nvidia Grace CPU over a high-speed link offering 900 GB/s of bidirectional bandwidth. This accelerates the database queries that are a large source of issues, the RAS engine reduces turnaround time and prevents potential vehicle safety problems that could result in crashes, injuries and fatalities.

**Nvidia is leveraging its ongoing learning and experience from its AI leadership position to rapidly add new functions and features to its GPUs and CPU chips.**

Decompression Engine can access large amounts of memory in the Nvidia Grace CPU over a high-speed link offering 900 GB/s of bidirectional bandwidth. This accelerates the database queries that are a large part of all AI LLMs and software platforms.

CUDA is the source of Nvidia’s success in GPU-centric applications. By year-end 2023, CUDA downloads surpassed 48 million. CUDA is a parallel computing platform and application programming interface (API) that allows software to use GPUs for many programming tasks simultaneously. This made CUDA the leader in AI applications, because AI models are all about exploiting as many GPU cores and

<table>
<thead>
<tr>
<th>Drive Version</th>
<th>Introduction Year</th>
<th>Microarchitecture</th>
<th>GPU Chip(s)</th>
<th>CPU Chip(s)</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive CX</td>
<td>January 2015</td>
<td>Maxwell</td>
<td>2× SMM Maxwell 256 CUDA cores</td>
<td>4× Cortex A57 4× Cortex A53</td>
<td>Digital cockpit</td>
</tr>
<tr>
<td>Drive PX</td>
<td>January 2015</td>
<td>Maxwell</td>
<td>4× SMM Maxwell 512 CUDA cores</td>
<td>8× Cortex A57 8× Cortex A53</td>
<td>ADAS</td>
</tr>
<tr>
<td>Drive PX2</td>
<td>January 2016</td>
<td>Pascal</td>
<td>1× 2 SM Pascal 256 CUDA cores</td>
<td>2× Denver; 64-bit 4× Cortex A57</td>
<td>ADAS</td>
</tr>
<tr>
<td>Drive PX2</td>
<td>January 2016</td>
<td>Pascal</td>
<td>2× 2 SM Pascal 512 CUDA cores</td>
<td>4× Denver; 64-bit 8× Cortex A57</td>
<td>ADAS</td>
</tr>
<tr>
<td>Drive PX Xavier</td>
<td>January 2017</td>
<td>Volta</td>
<td>1× Volta iGPU 512 CUDA cores</td>
<td>8× Arm Carmel Arm 64-bit</td>
<td>AV L3–L4</td>
</tr>
<tr>
<td>Drive PX Pegasus</td>
<td>October 2017</td>
<td>Turing</td>
<td>2× Volta iGPU 512 CUDA cores 2× Turing GPUs</td>
<td>16× Arm Carmel Arm 64-bit</td>
<td>AV L3–L4</td>
</tr>
<tr>
<td>Drive AGX Orin</td>
<td>December 2019</td>
<td>Ampere</td>
<td>2× Ampere iGPU 2K CUDA cores</td>
<td>12× Cortex A78</td>
<td>AV L3–L4</td>
</tr>
<tr>
<td>Drive Atlan</td>
<td>April 2021</td>
<td>Ada Lovelace</td>
<td>Cancelled</td>
<td>Cancelled</td>
<td>Cancelled</td>
</tr>
<tr>
<td>Drive Thor</td>
<td>March 2024</td>
<td>Blackwell</td>
<td>Arm Neoverse V3</td>
<td>Arm Neoverse V3</td>
<td>AV L3–L4</td>
</tr>
</tbody>
</table>

SMM = Maxwell's streaming multiprocessor design; SM = streaming multiprocessor; iGPU = integrated GPU
(Source: VSI Labs, April 2024)
What Is Nvidia Doing in Automotive?

### Table 2: Drive Thor Information

<table>
<thead>
<tr>
<th>Key Information</th>
<th>Other Information</th>
</tr>
</thead>
</table>
| Blackwell GPU vs. Ampere GPU | • Blackwell: 208 billion transistors  
• Blackwell: Unknown CUDA core count; 4× to 8× higher than Ampere?  
• Blackwell: 40,000-teraFLOPS FP4 |
| • Ampere: 17 billion transistors  
• Ampere: 2,048 CUDA cores  
• Ampere: 620-teraFLOPS FP16 |
| Two chip dies | • Blackwell has two chip dies  
• Each chip die has 104 billion transistors |
| RAS engine | • Reliability, availability and serviceability  
• Key for automotive crash safety |
| Transformer AI | • Blackwell Tensor Core technology |
| Data handling | • Blackwell Decompression Engine |
| CUDA base | • CUDA the secret to Nvidia’s success |
| | • Also key to Nvidia Drive’s success |

(Source: VSI Labs, April 2024)

### Table 3: Nvidia Drive Customers: Thor and Previous Drive Version Customers

<table>
<thead>
<tr>
<th>Key Information</th>
<th>Other Information</th>
</tr>
</thead>
</table>
| Drive Thor customers | • BYD Auto: Expanding use of Nvidia technology  
• DeepRoute.ai: Robotaxis in China  
• GAC Aion: Multi-ECU uses likely  
• Hyper: Luxury vehicles for 2025  
• Nuro R2: Goods-only AVs  
• Plus: Robotrucks—U.S., China and Europe  
• Waabi: Robotruck startup—North America  
• WeRide: Robotaxi, robotruck, robobus  
• Xpeng: Uses Drive Orin in G6 SUV  
• Zeekr: Future BEVs |
| • To use Thor, Isaac and Omniverse  
• Drive Thor for future robotaxis  
• Selected Drive Thor in 2022  
• Drive Thor for next-gen EVs  
• Future version of Nuro Driver  
• Future version of Plus SuperDrive  
• Future version of Waabi Driver  
• Coop. with Lenovo Vehicle Computing  
• Drive Thor for next-gen EVs  
• Selected Drive Thor in 2022 |
| Other Drive customers | • Aurora: Robotrucks  
• BMW Group  
• Ford  
• Jaguar Land Rover (JLR): Using Drive Orin  
• Li Auto: L9 SUV using Drive Orin  
• Lucid Air: Gravity SUV  
• Mercedes-Benz: Project from June 2020  
• Nio: BEVs  
• Polestar 3: BEV SUV  
• Pony.ai: Robotaxis use Drive Orin  
• QCraft: Robotaxis in China  
• Stellantis  
• TuSimple: Robotrucks  
• Volvo EX90 |
| • Long-term Drive customer  
• Listed as Nvidia partner at GTC 2024  
• Listed as Nvidia partner at GTC 2024  
• Listed as Nvidia partner at GTC 2024  
• Future use of Drive Thor  
• Using Drive Orin  
• Partner in M-B’s SDV & AV architecture  
• Using Drive Orin  
• Using Drive Orin  
• Selected Drive Thor in 2022  
• Using Drive Orin  
• Listed as Nvidia partner at GTC 2024  
• Using Drive Orin  
• Long-term Drive customer |

ECU = electronic control unit (Source: VSI Labs, April 2024)

Other accelerators as possible to run tasks in parallel.

This summary shows that Nvidia is leveraging its ongoing learning and experience from its AI leadership position to rapidly add new functions and features to its GPUs and CPU chips.

**DRIVE THOR CUSTOMERS**

Table 3 lists Drive Thor customers that were publicly known at the beginning of April 2024, but many more can be expected. The table also includes a partial list of customers of previous Drive versions. Most customers from earlier versions of Drive will be upgraded to Drive Thor as improved vehicles are introduced.

Other car companies using Drive—at least for research and testing—include Audi, Chery, Hyundai, Tesla, Toyota and VW. Truck companies using Drive but not listed in the table include DAF, Einride, Kenworth, Navistar and Peterbilt. Additional AV startups using Nvidia Drive include 2getthere, AutoX, Cruise, Didi, Navya, Optimus Ride and Zoox.

**NIM IN AUTOMOTIVE?**

Another strategy that strengthens Nvidia’s software leadership is Nvidia Inference Microservices (NIM), a way of packaging and delivering CUDA-based software that increases GPU-centric software availability. NIM services also create an opportunity for developers to reach hundreds of millions of GPUs with their custom AI software.

NIM services are built from Nvidia’s accelerated computing libraries and generative AI models. There is expected to be a growing NIM software base—courtesy of the standard NIM APIs—comprising third-party software developers who are attracted to the large CUDA installed base as future customers. The NIM will be most important for AI applications, especially business-centric AI.

Automotive NIM software is likely to be a growth market over the next decade and will apply to software-defined vehicles (SDVs), AVs and infotainment applications.

In summary, NIM is a set of optimized cloud-native microservices designed to shorten time to market and simplify deployment of generative AI models. The microservices work across cloud platforms, data centers, GPU-accelerated workstations and Nvidia Drive vehicles. NIM expands the AI developer pool by abstracting away the complexities of AI.
model development and packaging for production using industry-standard APIs.

NIMs will help expand AI models and applications across the automotive industry as a NIM software base becomes available for use across automotive segments and use cases.

**OMNIVERSE IN AUTOMOTIVE?**

In 2012, Pixar developed an interchange framework called Universal Scene Description (USD). Nvidia integrated the USD framework with its Omniverse, including technologies for modeling physics, materials and real-time path tracking.

In 2016, Nvidia released open-source USD software for generating 3D worlds using OpenUSD applications. OpenUSD provides a rich, common language for defining, packaging, assembling and editing 3D data to create virtual worlds across many industries, including automotive, architecture, construction, engineering, entertainment, media and telecom.

Nvidia Omniverse is now a platform of APIs, software development kits and services that enables developers to integrate OpenUSD and RTX rendering technologies into existing software tools and simulation workflows for building AI systems. (RTX is Nvidia’s ray tracing platform for designing complex visual models.) Nvidia’s view is that Omniverse brings AI into the physical world; in essence, Omniverse can mimic and simulate the real world.

A key announcement at GTC 2024 was the Nvidia Omniverse Cloud, which will be available as APIs. The cloud APIs are expected to extend the reach of Omniverse as the leading platform for creating industrial digital twin applications and workflows.

The five new Omniverse Cloud APIs let developers integrate core Omniverse technologies directly into existing design and automation software applications for digital twins. Developers can also integrate their simulation workflows to test and validate autonomous machines like robots or AVs.

Expect an expanded market presence for Omniverse in the automotive industry as digital twins become increasingly important for most automotive developers. AV and SDV digital twins are likely to become key Omniverse applications.

**SUMMARY**

Nvidia is a strong technology supplier to the automotive industry, and its importance is set to grow over the next decade. The automotive industry has become less significant to Nvidia, as the AI explosion has overwhelmed most other industry segments. However, the AI technology that Nvidia is developing will have a substantial impact on multiple automotive segments, from AVs and ADAS to SDVs and infotainment. Both Nvidia's chip and software technologies could advance automotive products and services.

Nvidia Drive continues to grow as an ADAS and AV platform. Nvidia Drive has over 25 design wins, with at least 50 models using or expected to rely on Nvidia Drive technology. Drive Orin is reaching volume use among multiple automotive OEMs and will likely continue as a production technology for many years. Drive Thor is gaining design wins, with first deployment expected in 2025 and production projected to extend through 2030 or longer.

Nvidia’s software strategy is as important as its hardware technologies in the automotive industry. The CUDA platform for using parallel processing software is a core reason for Nvidia’s success and is gaining strength as AI expands in automotive applications. The NIM software technology is set to extend Nvidia’s reach in automotive AI software systems.

Omniverse is another part of Nvidia’s software strategy that will benefit automotive development. The Omniverse platform is projected to be a key technology for creating digital twins across automotive activities, from product creation and testing to simulation and operation.

GTC 2024 was a milestone for Nvidia and showed its growing product portfolio and market reach across technologies and industries. The company is best known for its GPU chips, which are the core of its product strategy. The Blackwell GPU chip architecture is the latest advance in a long string of success stories. The strengths of Nvidia’s software platforms and accompanying chips and hardware will make the company a long-term leader and tough competitor.

**Figure 1: Nvidia’s automotive revenue** (Source: VSI Labs research, April 2024)

Nvidia’s automotive revenue reached nearly US$700 million in fiscal year 2024. This is due to the exceptional growth of Nvidia’s data center revenue for AI training and related applications. The data center segment accounted for 78% of Nvidia’s revenue in fiscal year 2024 compared with 27% in fiscal year 2020.

**Leader in Autonomous Driving**

NVIDIA DRIVE is our end-to-end Autonomous Vehicle (AV) and AI Cockpit platform featuring a full software stack and is powered by NVIDIA (systems-on-a-chip) SoCs in the vehicle.

- DRIVE Orin SoC ramp began in FY23
- Next-generation DRIVE Thor SoC ramp to begin in FY26

Drive Thor is gaining design wins, with first deployment expected in 2025 and production projected to extend through 2030 or longer.

Growth Drivers
- Adoption of centralized car computing and software-defined vehicle architectures
- AV software and services: Mercedes-Benz, Jaguar Land Rover

**What Is Nvidia Doing in Automotive?**

Nvidia is a strong technology supplier to the automotive industry, and its importance is set to grow over the next decade. The automotive industry has become less significant to Nvidia, as the AI explosion has overwhelmed most other industry segments. However, the AI technology that Nvidia is developing will have a substantial impact on multiple automotive segments, from AVs and ADAS to SDVs and infotainment. Both Nvidia’s chip and software technologies could advance automotive products and services. Nvidia Drive continues to grow as an ADAS and AV platform. Nvidia Drive has over 25 design wins, with at least 50 models using or expected to rely on Nvidia Drive technology. Drive Orin is reaching volume use among multiple automotive OEMs and will likely continue as a production technology for many years. Drive Thor is gaining design wins, with first deployment expected in 2025 and production projected to extend through 2030 or longer.

Nvidia’s software strategy is as important as its hardware technologies in the automotive industry. The CUDA platform for using parallel processing software is a core reason for Nvidia’s success and is gaining strength as AI expands in automotive applications. The NIM software technology is set to extend Nvidia’s reach in automotive AI software systems.

Omniverse is another part of Nvidia’s software strategy that will benefit automotive development. The Omniverse platform is projected to be a key technology for creating digital twins across automotive activities, from product creation and testing to simulation and operation.

GTC 2024 was a milestone for Nvidia and showed its growing product portfolio and market reach across technologies and industries. The company is best known for its GPU chips, which are the core of its product strategy. The Blackwell GPU chip architecture is the latest advance in a long string of success stories. The strengths of Nvidia’s software platforms and accompanying chips and hardware will make the company a long-term leader and tough competitor.

**Figure 1** shows that Nvidia’s automotive revenue reached nearly US$1.1 billion in the fiscal year ending in January 2024. As a share of revenue, automotive declined from 6.4% in fiscal year 2020 to 1.8% in fiscal year 2024. This is due to the exceptional growth of Nvidia’s data center revenue for AI training and related applications. The data center segment accounted for 78% of Nvidia’s revenue in fiscal year 2024 compared with 27% in fiscal year 2020.

Egil Julisessen is a principal analyst at VSI Labs and a contributing writer for EE Times Europe.
Nvidia's Developer Conference has been a major event in the GPU industry for more than a decade. With the company’s GPUs having become the leading AI chips over the past two years, the 2024 Nvidia GPU Technology Conference (GTC) attracted not just the traditional, engineering-centric attendees but many Wall Street analysts as well.

The AI GPU trend accelerated Nvidia’s market value and revenue tremendously in the past two years, and it is now one of the three most valuable companies in the world based on market capitalization. Nvidia’s revenue grew from US$10 billion in 2018 to US$61 billion in 2023, a spike that explains the influx of Wall Street analysts at GTC 2024.

The growth of generative AI since the launch of ChatGPT has accelerated the importance of GTC presentations and product announcements. This article summarizes Nvidia’s key announcements at GTC 2024, offering a perspective that considers Nvidia’s history, notably the evolution of its GPU chip architecture and technology; the importance of Nvidia’s CUDA software technology; and why and how Nvidia’s GPUs became a key technology for the AI industry.

**HISTORY OF NVIDIA’S GPU ARCHITECTURE**

The microarchitectures that Nvidia designs for each GPU generation offer significant improvements with each new version. Multiple GPU chips are designed based on each microarchitecture over a period of a few years. Each GPU chip has functionality that improves performance for specific application segments—initially for PC graphics and game consoles and recently for autonomous-vehicle functions and AI-centric software and systems.

Thus far, the company has introduced a total of 16 GPU microarchitectures, all named after famous inventors and scientists across multiple disciplines. The first was the Fahrenheit architecture, released in 1995. Nvidia’s latest GPU microarchitecture, Blackwell, debuted in March 2024.

Table 1 shows how Nvidia’s GPUs have improved since 1995. The table includes all 16 GPU microarchitectures, with a focus on the latest versions. The first GPU microarchitecture used in the automotive industry, in 2015, was Maxwell, Nvidia’s ninth GPU architecture, released in 2014.

The first Fahrenheit-based GPU had 1 million transistors and was based on a 500-nm manufacturing process technology. By 1999, the technology enabled GPU chips with 15 million transistors. Tesla, the sixth Nvidia GPU microarchitecture, was capable of 210 million transistors in 2006 and 1.4 billion transistors in 2010. During its four-year manufacturing run, the Tesla GPU chip shrank from 90 nm to 40 nm, even as the transistor count grew 6.7×.

The first Nvidia GPU microarchitecture to offer more than 1 billion transistors at introduction was Maxwell, which provided 2.9 billion transistors in 2014. Maxwell was also the first Nvidia GPU to be used in automotive applications. Each new generation added more computing power through computing accelerators for specific calculation in parallel or simultaneously, which increased the performance of GPU-based chips.

---

**Table 1: Evolution of Nvidia’s GPU Microarchitecture**

<table>
<thead>
<tr>
<th>Microarchitecture Name</th>
<th>Introduction Year</th>
<th>Microarchitecture Generation</th>
<th>Transistor Range</th>
<th>Manufacturing Technology</th>
<th>Fab Partner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fahrenheit</td>
<td>1995</td>
<td>1st</td>
<td>1M to 15M</td>
<td>250 nm to 500 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Celsius</td>
<td>1999</td>
<td>2nd</td>
<td>17M to 25M</td>
<td>150 nm to 220 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Kelvin</td>
<td>2001</td>
<td>3rd</td>
<td>29M to 63M</td>
<td>150 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Rankine</td>
<td>2003</td>
<td>4th</td>
<td>45M to 135M</td>
<td>130 nm to 150 nm</td>
<td>TSMC, IBM</td>
</tr>
<tr>
<td>Curie</td>
<td>2004</td>
<td>5th</td>
<td>75M to 302M</td>
<td>130 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Tesla</td>
<td>2006</td>
<td>6th</td>
<td>219M to 1,400M</td>
<td>40 nm to 90 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Fermi</td>
<td>2010</td>
<td>7th</td>
<td>260M to 3,000M</td>
<td>40 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Kepler</td>
<td>2013</td>
<td>8th</td>
<td>292M to 7,080M</td>
<td>28 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Maxwell</td>
<td>2014</td>
<td>9th</td>
<td>2.94B to 8B</td>
<td>28 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Pascal</td>
<td>2016</td>
<td>10th</td>
<td>1.8B to 12B</td>
<td>16 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Volta</td>
<td>2017</td>
<td>11th</td>
<td>21.1B</td>
<td>12 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Turing</td>
<td>2018</td>
<td>12th</td>
<td>4.7B to 18.6B</td>
<td>12 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Ampere</td>
<td>2020</td>
<td>13th</td>
<td>8.7B to 28.3B</td>
<td>7 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Hopper</td>
<td>2022</td>
<td>14th</td>
<td>80B</td>
<td>4 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Ada Lovelace</td>
<td>2022</td>
<td>15th</td>
<td>18.9B to 76.5B</td>
<td>4 nm</td>
<td>TSMC</td>
</tr>
<tr>
<td>Blackwell</td>
<td>2024</td>
<td>16th</td>
<td>208B</td>
<td>4 nm</td>
<td>TSMC</td>
</tr>
</tbody>
</table>

(Source: VSI Labs, April 2024)
As AI became the driving force behind GPU improvements, the transistor count rose to more than 28 billion for the Ampere GPU in 2020 and to 80 billion for the Hopper GPU in 2022. More parallel and specialized computing functions are key to improving performance for AI software. The Hopper GPU microarchitecture had especially extensive improvements for accelerating AI computing. Hopper was the first GPU with a transformer engine for acceleration of AI training models.

In March 2024, Nvidia announced Blackwell as the latest GPU generation, with 208 billion transistors. Product shipment is due later this year. The 2024 Blackwell GPU chip will more than double the Hopper performance and transistor count. However, much of the transistor increase is due to the use of two chip dies connected by a 10-TB/s chip-to-chip interconnect in a unified single GPU. Each die has 104 billion transistors—a 30% increase over Hopper’s 80 billion transistors. The Blackwell architecture can have up to 576 GPU processors on a chip. Blackwell includes a second-generation transformer engine.

TSMC has manufactured the vast majority of Nvidia’s chips, though IBM and Samsung have manufactured some chips for Nvidia in the past 20 years. Samsung was a fab partner from 2016 to 2022. SGS-Thomson Microelectronics was an early Nvidia fab partner.

CUDA PLATFORM

GPUs were originally designed for image manipulation and the calculation of local image properties. The mathematical foundations of neural networks and image manipulation are similar, which created a big opportunity for GPUs as AI chips. In the mid-2010s, GPUs evolved to enable deep learning for training and inference in many applications, including autonomous vehicles.

Nvidia defined a GPU programming interface called CUDA (which stands for Compute Unified Device Architecture, though Nvidia uses the acronym only) and released the first version in mid-2007. CUDA version 12.4 was released in March 2024. CUDA is a parallel computing platform and application programming interface that allows software to use GPUs for accelerated general-purpose processing. Nvidia has used “accelerated computing” to define its product strategy for about a decade.

The CUDA API is an extension of the C programming language that adds the ability to specify software-level parallelism in C and to specify GPU-device-specific operations. CUDA is also a software layer that gives direct access to the GPU’s virtual instruction set and parallel computational elements. CUDA is designed to work with programming languages like C, C++, Fortran and Python. This accessibility makes it easier to develop parallel processing via the many processors and accelerators provided in GPU chips. Many corporations program their AI applications in CUDA to get maximum performance from Nvidia GPU systems. This has built a broad CUDA software base and a large population of programmers with expertise in developing software for Nvidia’s GPUs.

Intel’s continued leadership in processor chips for the PC market is attributable to the software base that was built around IBM’s PC standard in the 1980s and 1990s. Intel has leveraged that software to maintain dominance in PC chips for more than 40 years, but it did not transfer this software advantage to later computer segments like smartphones and tablets.

Nvidia has a similar software base, built around CUDA, that has given the company a dominant share of the GPU market for over a decade. Nvidia retained its leadership during the graphics card era and through the video processing era as well. Now, the CUDA software base and parallel processing capabilities are the leaders for AI model development and AI application deployments. This AI software base advantage will persist for some time, even with increased competition in the future.

The question is whether Nvidia can extend its software base advantages to new AI applications and other new application segments as they emerge. Nvidia has embraced and transitioned to new CUDA opportunities before and is likely to do so again. Another question is whether the Unified Acceleration (UXL) Foundation will have an impact. The UXL Foundation was launched in September 2023 to develop a CUDA alternative and has an impressive membership roster, including Arm, Fujitsu, Google Cloud, Intel, Qualcomm, Samsung and others.

The formation of UXL looks like an acknowledgement that CUDA is a clear leader and that no single company can compete with CUDA’s software base and momentum. Only a foundation with major company participation can or may build a CUDA competitor.

HOW NVIDIA BECAME THE LEADER IN AI

We have already discussed two reasons for Nvidia’s rise to AI dominance: GPU technology advances for AI calculations and the CUDA software platform for writing code for parallel execution of many streams of information. The technology advances of GPUs have been very rapid—from 1 million transistors on a GPU in 1995 to 208 billion transistors in 2024. This functionality and performance growth was key to enabling the rapid expansion and complexity of AI models. There are at least two more factors: building momentum in graphics- and video-related GPU applications, and understanding and reacting to the potential of AI as a major future GPU market.

On all counts, Nvidia’s management gets credit for crafting and implementing a successful strategy.

Nvidia made a foresighted investment in the late 2000s to set up CUDA programming classes at most of the major universities worldwide—at Nvidia’s cost. The bold gambit paid off; it was part of the creation of a market concept and enabled the later growth of the CUDA software base.

Nvidia’s early realization that AI had great potential was critical. The company first focused on AVs as an AI opportunity and included GPU functionality for speeding up AI calculation. The AV market was delayed, but generative AI came along to take AVs’ place—and Nvidia was ready to ride this market growth. The rest is a great history for Nvidia.

Another key factor is that Nvidia became as much a software company as a chip design company. Today, Nvidia is more of a software company with highly skilled AI GPU chip designers. Nvidia has developed a large AI software base, including software development tools, AI libraries and AI foundational models for a large spectrum of AI applications.

Through its AI software development activity, Nvidia gained an understanding of what GPU hardware architecture and processing accelerators were needed for AI performance, especially for the AI training phase. This knowledge base was reinforced and amplified by building AI systems for cloud system operators and similar customers. All of this has given Nvidia the expertise to rapidly improve its AI-focused GPUs and the systems for AI training and inferencing. In the last two GPU architecture upgrades, the focus was on improving AI performance for training and inferencing, AI application improvements will be the focus of future GPU chip architectures.

Apple is known for gaining major advantages from designing both its hardware and software into a better system experience for its users.

The AI GPU trend accelerated Nvidia’s market value and revenue, and it is now one of the three most valuable companies in the world based on market capitalization.
Nvidia is reaping similar advantages from its activities in chip and hardware design and the CUDA software platform with additional AI software development.

**KEY ANNOUNCEMENTS AT GTC 2024**

GTC 2024 was primarily a technical conference for GPU developers, as evidenced by its more than 900 technical sessions and over 20 workshops. More than 300 exhibitors showcased their hardware, software and services focused on GPU-based market opportunities.

The many announcements at GTC 2024 covered all of Nvidia’s product lines. Table 2 summarizes most of the announcements.

**Blackwell GPU**

Topping the list of announcements was the Blackwell GPU, which Nvidia claims is the most powerful processor chip available. Nvidia is positioning Blackwell as the next generation of accelerated computing and as the processor for the generative AI era. Blackwell has a second-generation transformer engine with support for FP4–FP6 data types in the Tensor cores. Most generative AI models need primarily low-precision calculations, and Blackwell performance is greatly improved under such conditions.

The new GPU architecture is named after David Harold Blackwell. A University of California, Berkeley mathematician specializing in game theory and statistics, he was the first Black scholar inducted into the National Academy of Sciences.

Blackwell GPUs include a dedicated engine for reliability, availability and serviceability (RAS). The RAS feature is especially important for automotive and other systems in which a failure can lead to a loss of life. Blackwell also adds chip-level capabilities for using AI-based preventative maintenance to run diagnostics and forecast reliability issues. This improves system uptime and resiliency for massive-scale AI deployments to run uninterrupted for weeks or months at a time and to reduce operating costs.

Nvidia said Blackwell is being adopted by every major global cloud services provider, pioneering AI companies, system and server vendors and regional cloud service providers. Nvidia believes Blackwell will be its most successful product launch in its history. The cloud players need high-performance systems and connections based on Blackwell. Some of these are summarized below.

**DGX SuperPOD**

Nvidia announced its next-generation AI supercomputer: the Nvidia DGX SuperPOD. It is powered by Nvidia DG GB200 Grace Blackwell superchips and can process trillion-parameter models for large-scale generative AI training and inference workloads. It provides 11.5 exaFLOPS using FP4 data types and 240 TB of fast memory. It can scale to higher performance with additional racks of DGX systems.

**NVLink Switch chip**

NVLink Switch chip is Nvidia’s high-speed network link for connecting multiple GPU processors and systems. It is a complex chip that has 50 billion transistors and is manufactured by TSMC using 4-nm design rules. Each NVLink Switch can connect four NVLink interconnects at 1.8 TB/s.

NVLink Switch and GB200 are key components for creating giant GPUs. The Nvidia GB200 NVL72 is a multi-node, liquid-cooled, rack-scale system that harnesses Blackwell to offer supercharged compute for trillion-parameter models with 720 petaFLOPS of AI training performance and 1.4 exaFLOPS of AI inference performance in a single rack.

**NIM runtime software**

Nvidia Inference Microservices (NIM) are secure software packages built from Nvidia’s accelerated computing libraries and generative

---

<table>
<thead>
<tr>
<th>Table 2: Announcement Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic</strong></td>
</tr>
<tr>
<td>Blackwell GPU</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Nvidia's positioning of Blackwell</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Blackwell vs. Hopper</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>DG GB200 systems</td>
</tr>
<tr>
<td>DGX SuperPOD</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>NVLink Switch</td>
</tr>
<tr>
<td>Omniverse Cloud API</td>
</tr>
<tr>
<td>NIM runtime software</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>6G Research Cloud</td>
</tr>
<tr>
<td>Weather prediction</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Blackwell-based Thor</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

FP4 = floating point with 4-bit accuracy

(Source: VSI Labs, April 2024)
AI models. The microservices support standard APIs to connect and work across Nvidia’s CUDA installed base. They are re-optimized for new GPUs and are scanned for security vulnerabilities and exposures.

Nvidia is launching a new type of NIM-based biosystem software. The company rolled out more than two dozen microservices that will allow healthcare enterprises to leverage the latest advances in generative AI.

Omniverse Cloud API
The main goal of Omniverse is to bring AI to the physical world. Nvidia announced that Omniverse Cloud will be available as APIs to extend the reach of the Omniverse platform for creating industrial digital twin applications and workflows across many software ecosystems.

The five new Omniverse Cloud APIs enable developers to integrate core Omniverse technologies into existing design and automation software applications for digital twins. This provides simulation workflows for testing and validating autonomous machines like robots and self-driving vehicles.

6G Research Cloud
In telecom, Nvidia announced the Nvidia 6G Research Cloud, a generative AI and Omniverse-powered platform to advance the next communications era. It’s built with Nvidia’s Sionna neural radio framework, the Nvidia Aerial CUDA-accelerated radio access network and the Nvidia Aerial Omniverse Digital Twin for 6G. This offering should help telecom developers test and simulate the many options for defining the technologies and features of 6G.

Weather prediction
Nvidia announced the availability of its Earth Climate Digital Twin, a cloud platform that enables interactive, high-resolution simulation to accelerate climate and weather predictions at a 2-km scale.

Nvidia also announced new Earth-2 cloud APIs on Nvidia DGX Cloud. These will allow users to create AI-powered emulations to speed the delivery of interactive, high-resolution simulations of phenomena ranging from global atmospheric conditions and local cloud cover to storms and other events. Earth-2’s APIs offer AI models and a new Nvidia generative AI model, called CorrDiff, using state-of-the-art diffusion modeling to generate 12.5× higher-resolution images than are possible with current numerical models. Compared with the current models, the Earth-2 API models are 1,000× faster and 3,000× more energy-efficient.

Blackwell Thor
For the automotive industry, Drive Thor was the key announcement at GTC 2024. Drive Thor incorporates the Blackwell GPU architecture, which is designed for transformer and generative AI applications. These capabilities will be important for future AV and ADAS functionality.

There is also a Jetson Thor version, which targets robotics applications and is also likely to be used in some automotive segments.

SUMMARY
Nvidia made more announcements at GTC 2024 than at any other of its conferences. The Blackwell GPU is the most important of those introductions, as there will be many new products—some announced at the conference and others likely to appear next year.

Nvidia’s GTC 2024 was an important event that shows Nvidia’s strategy and product direction for the next couple of years. It looks like Nvidia will continue its domination in GPU-based hardware and CUDA-based software for existing AI software segments and use cases over that time period. Nvidia will face more competition and will lose some market share, albeit of a much larger pie. Expect Nvidia to retain its AI leadership for a long time to come as it leverages its synergistic chip-hardware-software products and services.
In this article, we will discuss how Advantech is helping customers take advantage of AOI for the production of PCBs and ICs. We will highlight the new capabilities of Advantech’s AMD-powered AIMB-723 industrial motherboard, which is enabling a new era of AOI-based visual inspections.

AOI COMPUTING REQUIREMENTS

The ideal AOI solution requires high CPU frequencies and industrial-grade durability. It demands that GPU processing be complemented by the appropriate amount of compute power from an embedded CPU. To date, computing power has been too expensive for AOI machine makers to utilize properly.

More specifically, the ideal solution is a single board that integrates both vision (GPU) and compute (CPU) processing capabilities, and that ideal is now a reality. Advantech has brought to market the AIMB-723, powered by the AMD Ryzen Embedded 7000 Series processor, the first industrial-grade motherboard integrated with an AMD Socket AM5 chipset. The new board enables manufacturing equipment suppliers and electronics manufacturers to use AOI to perform visual inspections of PCBs and ICs during manufacturing in which a camera is used to scan the board meticulously for any defects or failures. AOI monitors the quality of PCB and IC production and corrects them in the process flow, which is a key to success in today’s competitive production environment.

ADVANTECH AIMB-723 FEATURES

The board offers the industry’s most highly integrated and capable AOI solution. In addition to accelerating performance-demanding AOI applications, it features an optimized form-factor design through PCIe placement for GPU card installation, thus supporting the high performance that is demanded for vision- and compute-intensive industrial applications. It is scalable to support a three-slot–width GPU card and an additional three-frame grabber card for machine-vision applications, enabling it to cater to legacy requirements for PCI expansion while accommodating a wide range of communication protocols and I/O interfaces. The Advantech AIMB-723 also provides extraordinary heat dissipation and industrial-grade durability.

The board satisfies the requirements of lightning-fast memory and streamlines data transfer and processing. Its specific features include the following.

Optimized for AOI applications
- Four dual in-line memory module (DIMM) sockets, for up to 128-GB DDR5 at 5,200 MHz—the first AIMB motherboard to embrace DDR5 unbuffered DIMM (UDIMM)
- PCIe ×16 Gen4 with ample space for a graphics (GFX) card
- Flexible storage solutions with four SATA ports and an M.2 slot
- Support for legacy devices on COM ports and a PCI slot

Incorporates powerful AMD Socket AM5 + Ryzen Embedded 7000 Series processors
- Fast “Zen 4” performance on up to 105-W embedded processors
- AMD PRO technologies built for modern business
- Abundant I/O: up to 12 USB connectors (8× USB 3.2)
- SuperSpeedPlus USB 3.2 Gen2 10 Gbps (4× ports)

Developed by global leader in IPC manufacturing
- Ultra-durable; proven when tested under 65/–5°C operating environment
- Long-lasting, with MTBF up to 668,283 hours
- Longevity: product support up to seven years

As industrial applications continue to grow in capability and complexity, they are driving the need for the use of increasingly powerful processors in manufacturing. One such application, automated optical inspection (AOI), is an important aspect of a robust industrial manufacturing strategy, used in the production of all kinds of products but especially critical for monitoring printed-circuit boards and integrated circuits for defects and accurate measurements.

While PCBs and ICs are shrinking in size, they are packing more complex capabilities. Even a relatively simple board can be made up of literally thousands of soldered components, making detecting and monitoring for defects by the human eye virtually impossible and therefore making AOI-based systems an absolute necessity in today’s production environment.

In this article, we will discuss how Advantech is helping customers take advantage of AOI for the production of PCBs and ICs. We will highlight the new capabilities of Advantech’s AMD-powered AIMB-723 industrial motherboard, which is enabling a new era of AOI-based visual inspections.
A POWERFUL PUNCH: THE AMD RYZEN EMBEDDED 7000 SERIES PROCESSORS

Building on its leadership in next-generation 5-nm technology, the AMD Ryzen Embedded 7000 Series processor family is optimized for the high-performance requirements of industrial markets. The Ryzen Embedded 7000 Series processor is the first embedded processor to use 5-nm technology with a seven-year manufacturing availability commitment, driving increased product longevity.

By combining “Zen 4” architecture and integrated Radeon graphics, Ryzen Embedded 7000 Series processors deliver performance and functionality not previously offered for the embedded market. With their expanded features and integration, Ryzen Embedded 7000 Series processors were an ideal solution for the new Advantech AIMB-723.

A thermal design power (TDP) spanning 65 to 105 W allows system designers to choose a power level that meets their requirements. In addition, the AMD Ryzen Embedded 7000 Series delivers significant gains in performance and efficiency over the Ryzen Embedded 5000 Series. At the 65-W TDP setting, the Ryzen Embedded 7000 Series provides up to a 74% performance increase.

In third-party testing by PassMark Software, the AMD Ryzen devices were found to have a better price/performance ratio over competing CPU products, representing more user performance value for deployments. Additionally, in competitive performance tests, the AMD Ryzen 5 PRO 7645 exhibited up to a 121% performance increase in PCMark 10 benchmark testing and a 148% in PCMark 10 productivity assessments. AMD industrial customers are now able to leverage this powerful family to meet the needs of their next-generation product performance requirements.

CUSTOMER TRACTION

The AMD-powered Advantech AIMB-723 is currently being tested by customers for AOI, edge computing and semiconductor equipment. One is a world leader in the supply of semiconductor assembly and packaging equipment and surface-mount technology solutions. Based in Taiwan, it produces high-quality equipment for all major steps in the electronics manufacturing process, from carriers for chip interconnection to chip assembly and packaging to semiconductor manufacturing. The supplier initially produced some of its equipment using consumer-grade components and parts, a practice that resulted in customers struggling with product reliability and other issues.

As a long-time Advantech partner, the supplier requested an industrial-grade hardware solution, and Advantech initially developed a motherboard based on technology from a competing silicon company. However, because of the performance enhancements offered by Advantech with the new AIMB-723, the customer plans to use the new motherboard for production deployments.

ADDITIONAL INFORMATION

For more information on the AMD-powered Advantech AIMB-723 and its capabilities, visit tinyurl.com/3ww9e6nj.

<table>
<thead>
<tr>
<th>AMD Ryzen Embedded 7000 Series Processor Model Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>AMD Ryzen™ Embedded 7945</td>
</tr>
<tr>
<td>AMD Ryzen™ Embedded 7745</td>
</tr>
<tr>
<td>AMD Ryzen™ Embedded 7700X</td>
</tr>
<tr>
<td>AMD Ryzen™ Embedded 7645</td>
</tr>
<tr>
<td>AMD Ryzen™ Embedded 7600X</td>
</tr>
</tbody>
</table>

CPU Value Comparison:
AMD RYZEN™ 7000 vs. 12th/13th Gen Intel® Core™i

As of Dec. 26, 2023—higher results represent better value (PassMark Software—PC Benchmark and Test Software). Test methods may vary, yielding different results.
Thanks in large part to EU funding, European supercomputers are making a strong showing on the world stage. Three of the top 10 entries in the November edition of the TOP500 list of the world’s fastest supercomputers are located in EU member states. The first is Lumi, based in Finland and taking fifth place in the November 2023 edition of the TOP500. The second is Leonardo, which operates in Italy and takes sixth place on the same list. And the third is MareNostrum, based in Spain and appearing as the world’s eighth fastest computer on the TOP500.

Each of these three systems was financed through a combination of investments from national agencies and matching funds from the EuroHPC Joint Undertaking, which began operating in November 2018. It’s interesting to note that during that same month, only one supercomputer in an EU member state made it to the Top 10. That was SuperMUC-NG, in Germany, which took eighth place.

EE Times Europe spoke with leading executives in the organizations running Lumi, Leonardo and MareNostrum. Each told us about the kinds of applications their supercomputers run and how they have adapted to meet the challenges of AI and other applications. We also spoke about future architectures that will meet the growing requirements for faster processing of ever-larger data sets while minimizing power consumption.
Lumi: CSC’s Manninen on Managing Europe’s Biggest Supercomputer—and AI’s Expectations

By Pat Brans

Housed in the Advanced Computing Facility at CSC’s IT Center for Science in Finland, Lumi has been in operation since 2021 and reached its full capacity in 2023. According to the most recent edition of the Top500 list, it’s now the fifth-largest supercomputer in the world and the biggest in Europe.

To find out more about the creation of Europe’s fastest supercomputer, its management and the applications it runs, we spoke to Pekka Manninen, who, as director of science and technology at the Advanced Computer Facility, is responsible for the underlying technology.

**EE TIMES EUROPE: What is the story behind Lumi, and how did it get its name?**

**Pekka Manninen:** Obviously, many people were involved in the decision-making and preparation in many countries and at many levels. But if I had to name a single person, it would be CSC’s managing director, Kimmo Koski, who has relentlessly advocated for pan-European investments in HPC [high-performance computing], starting well before the European Commission introduced plans for a European HPC infrastructure in the form of the Europe HPC Joint Undertaking [EuroHPC JU].

Once EuroHPC JU was launched, Kimmo harnessed the interest of our organization and of our partners. We all got behind the idea of proposing CSC’s data center in the city of Kajaani, which was already the home of Finnish supercomputers.

Back then, there was a consortium of nine countries that supported the idea and agreed to contribute funding. [These were Finland, Denmark, Estonia, Sweden, Norway, Belgium, Czech Republic, Switzerland and Poland.] In Finland, the key contributors were the Ministry of Education and Culture and the Ministry of Economic Affairs and Employment, which both helped with funding. Thanks to the efforts of all these groups, the EuroHPC JU granted the Hosting Agreement to the Lumi consortium. Two other countries [Iceland and the Netherlands] have since joined the consortium, which pays half of the total cost of over €202 million. The EU contributes the other half.

As for the name Lumi, I was actually the one who came up with that. It means “snow” in Finnish. I thought a white supercomputer would look cool, and then figured that snow would be a strong reference to northern Europe and white at the same time. After we named it Lumi, it was retrofitted with a slightly clumsy “backronym”: Large Unified Modern Infrastructure.

Lumi is housed in what used to be a paper mill, which was owned by UPM, one of the biggest pulp and paper producers in the world. UPM decided to close the plant in 2008 because of a worldwide overproduction of paper. Because the paper mill required over 230 MW of energy, the power capacity we needed was already there when we moved in.

By the way, the local district heating plant is in the same area, so we push excess heat to that system. The result is that Lumi heats 15% to 20% of Kajaani’s homes—and during certain times, such as when heating plant maintenance breaks, we can heat the whole city.

CSC Finland operates the facility and takes care of system administration. Keeping the data center open requires only around 10 people. Then there are a lot of other people doing other work to ensure the usefulness and productivity of the supercomputer. Our consortium countries run user support, and special interest and collaboration groups address things like public relations, AI and cybersecurity.

**EE TIMES EUROPE: What are the main architectural features of Lumi, and how does it compare with other world-class supercomputers?**

**Manninen:** Lumi consists of around 3,000 GPU nodes, each with four GPUs, and around 2,000 CPU nodes, with two CPUs each. Its sustained compute capability toward a single binary is around 380 petaFLOPS at fp64 precision. Due to ambitious rack engineering, it’s possible to condense this and all the necessary storage into a rather compact space, which in terms of square meters is around the size of two tennis courts.

Lumi uses all-AMD node technology, with AMD MI250X GPUs and AMD Milan 64-core CPUs. Both computing partitions and all the auxiliary resources—including data analytics partitions plus all the storage and data management solutions—are tightly interconnected in the same HPE Slingshot network. Slingshot is based on high-radix switches, which enableexascale and hyperscale data center networks with at most three switch-to-switch hops. It uses an optimized Ethernet protocol, which allows it to interoperate with standard Ethernet devices while providing high performance to HPC applications.
Lumi: CSC’s Manninen on Managing Europe’s Biggest Supercomputer—and AI’s Expectations

About 50% of Lumi’s capacity now goes to AI. … This is clearly a higher share of the workload than we anticipated in back in 2019, when we were expecting simulations to take the biggest share.

— PEKKA MANNINEN

EE TIMES EUROPE: How do you decide who gets to use Lumi, and what are some of the applications that are run? Manninen: The EuroHPC JU grants 50% of the access time, and the Lumi consortium countries grant the other 50%, in proportion to their contributions to the total cost of ownership. Each of them has a slightly different role for Lumi in their national computing infrastructures.

Scientists and companies use Lumi for tens of different computing problems and related applications. Obviously, the main reason for building large systems is to tackle computing problems that are intractable with smaller systems. For instance, Lumi has run the most accurate climate simulations to date, comprising 50 years of coupled Earth system model scenarios at a 5-km global resolution. Compare that with the previous state of the art in climate modeling; which was at around a global resolution of 100 km. We have seen similar leaps in accuracy and fidelity in solar magnetosphere modeling and plasma physics simulations run with Lumi.

Very often, the science communities develop and maintain their own applications and use the supercomputer as a platform-as-a-service in a cloud manner. Lumi has around 3,000 user accounts and a couple of hundred projects. There’s a very big spectrum of different applications that run on the system. But the distribution is very top heavy, in the sense that a handful of applications consume 95% of the resources.

About 50% of Lumi’s capacity now goes to AI, which here refers to training deep neural networks for different purposes, especially for large language models but also for things like image recognition. This is clearly a higher share of the workload than we anticipated back in 2019, when we were expecting simulations to take the biggest share. Traditional HPC would be running things like molecular dynamics and CFD [computational fluid dynamics], so that’s what we thought would be the big use cases.

Fortunately, the system we built turned out to be a very suitable system for AI. I think AI will impact all fields of science, becoming complementary to simulations—or even replacing them in some domains. This use case clearly needs a lot of computing power and ultra-fast data access, which can only be delivered by large supercomputers.

We didn’t need to make so many hardware adjustments, but there were some cultural differences we had to address. The AI community has different expectations than what we were used to in the HPC world—for example, in the way the system is accessed.

In traditional HPC, people tend to submit batch jobs; they put a job in a queue, the machine executes it, and [they] go and check the results. AI people, on the other hand, want much more interactivity, and they would like to have a big part of the machine for themselves for several weeks. That’s a cultural clash that takes quite a lot of expectation management to make sure everybody benefits from the machine.

EE TIMES EUROPE: Will supercomputers as they are currently designed be able to keep up with future demand? If not, what needs to change? Manninen: Clearly, the demand for compute is in a rapid increase, especially given the surge of AI in science and in commercial applications.

Not everything lends itself to parallel processing. But we hit the wall a long time ago on how fast we can make one serial execution unit. Nobody even thinks about 10-GHz processors anymore. In fact, we have had to gun down a lot on the clock frequency to build multiple execution units. This is most heavily manifested by today’s GPUs, which are essentially big parallel processing units.

There are many other factors that determine the speed of processing of a non-parallelizable workload. It’s not necessarily things like the clock frequency; it can be things like memory access speed, or how different components are accessed—for example, file I/O. And bad code will slow things down no matter how powerful the computer is. It’s very often an algorithmic problem and then a software problem.

While there is a set of computational problems that won’t ever be run on anything other than serial computing, it’s not a very big set. For many problems, we will still need to scale workloads over tightly connected nodes and even connected systems. I’m not sure if it’s necessary—or even possible—to keep building larger supercomputers in terms of node counts. One solution may be to take several large supercomputers and have them interoperate in a federated fashion to speed up suitable workflows.

AI is a good example of something that’s very nicely parallelizable. It relies on dense linear algebra with many layers of parallelism we can exploit. The AI computing demand is something we can handle with the multicore CPUs and GPUs—no problem there. And the reason GPUs are so well-suited is that they just get more FLOPS for the same power budget—that is, better performance per watt or per dollar than traditional x86 CPUs. The recent GPU surge happened because they are very good for AI, but in fact, the adoption of GPUs in scientific computing started well before the foundational models and other extreme computing needs of AI.

An interesting story worth mentioning is about floating-point operations—IEEE floating-point arithmetic. In the HPC world, we are used to working with 64-bit arithmetic, but the GPUs were actually never built for that. They originate from computer games, where you don’t really care if a pixel is slightly off, and a pixel can be represented by only 4 or 8 bits. To bring GPUs to high-performance computing, it took 10 years for system vendors and the HPC community to get GPUs excelling in 64-bit precision. Too. But now, with AI, there is a big workload that doesn’t need the 64-bit precision, so there is a trend toward lower-precision arithmetic, which puts pressure on traditional simulation software to work with the lower-precision arithmetic.

One more thing on the topic of meeting future demands: I mentioned the importance of algorithm software in serial processing, and the same is true in distributed processing. One of the biggest bottlenecks in supercomputing is in the application software.

The scientific community loves tested and proven legacy software. But the programming solutions that were developed 50 years ago are often suboptimal on present-day hardware. Code modernization and good software engineering are needed to keep up with developments in supercomputing.
Sanzio Bassini: Leonardo is housed in what used to be a large tobacco factory in Bologna that was constructed in 1952. The factory was designed by Pier Luigi Nervi, who was a star architect at that time—like Gaudi was a star in Spain. In fact, because it had been designed by Pier Luigi Nervi, it’s considered an industrial archaeological site, under the surveillance of the Ministry of Cultural Heritage in Italy. The design was such that the production part of the factory consisted of a very large collection of barrels, each 30 meters in diameter and 100 meters long.

EE TIMES EUROPE: Could you explain how Leonardo started?
Sanzio Bassini: Leonardo is housed in what used to be a large tobacco factory in Bologna that was constructed in 1952. The factory was designed by Pier Luigi Nervi, who was a star architect at that time—like Gaudi was a star in Spain. In fact, because it had been designed by Pier Luigi Nervi, it’s considered an industrial archaeological site, under the surveillance of the Ministry of Cultural Heritage in Italy. The design was such that the production part of the factory consisted of a very large collection of barrels, each 30 meters in diameter and 100 meters long.

In 2006, the British American Tobacco Company, which owned the factory at the time, decided to move all production to Naples. Ownership was then handed off to the regional government, which in 2019 turned the building over to ECMWF [European Centre for Medium-Range Weather Forecasts], which built a data center with two large HPC systems in three of the large barrels. During the same year, a pre-exascale computer that was partially funded by EuroHPC was built in another barrel. The year 2019 was the 500th anniversary of the birth of Leonardo da Vinci, so we called that new supercomputer Leonardo. Leonardo is co-funded by EuroHPC and by the Italian Ministry of Universities and Research.

The site is now one of the largest concentrations of high-performance computing systems in the world. Two large HPC systems are managed by ECMWF, Leonardo is managed by Cineca and another large hyperscale HPC system is housed in another barrel and is managed by INFN [the National Institute of Nuclear Physics].

EE TIMES EUROPE: How big is Leonardo?
Bassini: The current configuration of Leonardo consists of two partitions. One is for general purposes, and we call it the Data Centric module. It consists of 1,536 compute nodes based on Intel Sapphire Rapids CPUs. The other partition is for heavier tasks and is called the Booster module. It consists of 3,456 compute nodes, which are based on Nvidia A100 GPUs.

We are in the process of adding a third partition, specifically for AI. This is part of a major upgrade that we expect to be open for production at the beginning of 2025. The new partition will be called LISA, an abbreviation for Mona Lisa. We’ve retrofitted a name to fit the acronym: Leonardo Improved Supercomputer Architecture.
The motivation for this upgrade is driven by an increasing demand for resources to train LLMs [large language models]. We are currently working on two big projects. One is for a text-based LLM for Italian; the other is for a multimodal LLM for Italian, which would include text, sounds and images. Leonardo has also been used to train other European languages—for example, Mistral, the French-language LLM.

We also have requests to train domain-specific LLMs. One is for weather forecasting, based on radar images. Another is for an LLM related to the "Beyond 1 Million Genomes" [B1MG] project, which is a European initiative.

All these projects—the general LLMs and the more domain-specific ones—require very, very significant resources, including both compute and data storage and transfer. We think that one way of improving how we accommodate AI applications is by making it easier to access huge amounts of data. To that end, we plan to make available a large data lake repository based on multiprotocol S3 technology both for training and inference phases.

The third partition—the one designed for AI—will be fully integrated with the bigger system. We will also integrate quantum computing technology.

**EE TIMES EUROPE: What are some of the applications Leonardo runs now?**

**Bassini:** It’s easy to get a broad overview because we show the distribution of the workload in our annual report. The big application areas we reported for the year 2023 are condensed-matter physics, computational chemistry, computational fluid dynamics, nuclear fusion, computational engineering, astro and plasma physics, earth and climate, life sciences and computational biology, life sciences and bioinformatics, particle physics, and AI and machine learning.

Some of the life sciences applications support the idea of personalized and precision medicine and require a lot of data, some of which is highly personal. To protect data privacy, very high standards of security are needed. Cineca created a working environment that complies with standards for information security management in full compliance with GDPR [General Data Protection Regulation].

Some of the recent use cases involved analyzing the suitability of land for vegetation. This is particularly important in the Piedmont region of Italy, where agriculture makes up a large portion of the economy. With climate change, policymakers and scientists need a new set of tools to make predictions. This kind of analysis requires the power of a supercomputer.

Then we’ve had users who wanted to assess agricultural risk associated with climate change for insurance purposes. This analysis required big data and artificial intelligence to find patterns that could be used to make predictions. A lot of data and a lot of computing power are needed.

Many of the scientific users want to run sophisticated simulations. For example, some have used Leonardo to simulate the behavior of new materials that might be used to improve computers. Another example is to run simulations of the behavior of black hole binaries in astrophysics; there is also a big need to run simulations to predict plasma behavior for nuclear fusion.

Of course, AI is a big use case for several different domains, including automotive. Both training and inference require huge amounts of data and computing power and will be a big part of what drives the evolution of Leonardo.

**EE TIMES EUROPE: Can you tell us more about the planned upgrades and the evolution of Leonardo?**

**Bassini:** One of the first things we plan to do is connect Leonardo with the two S3 data lake repositories we plan to set up—one in Bologna and the other in Naples. Since we’re embracing AI for science and innovation, we need a huge amount of data, and we do need to provide easy access to the data. Leonardo has 100 petabytes [PB] of scratch storage capacity. We want to provide a permanent home for data that can be used for different applications. We are planning at least 100 PB of data lake repository, multiprotocol S3, for managing data of any kind.

Another step will be to implement what we call AI factory, a system that will support AI workloads. The idea is that a cluster of SMPs [shared-memory multiprocessors] would be better than a cluster of the server nodes we use under the current architecture. By the way, Leonardo was ahead of its time when it was built. It was the first system designed with four GPUs per server node. Before [Leonardo], the maximum was two GPUs per server node. We designed it with Atos, the system integrator, and Nvidia.

AMD technology has two core GPUs per socket, which means that two sockets would have four GPUs. But it’s different from the point of view of data movement and access to shared memory. So it’s very effective for LINPACK but not for AI, because data movement becomes a very big issue with AI.

We are beginning to think about the design of a post-Leonardo system. And we are of the mind that this cluster of SMP GPU nodes would be a good foundation. We would want to go beyond the eight GPUs currently available from Nvidia, for example, toward maybe a cluster of 16 GPUs per server node, something like that. Of course, that would require us to overcome the challenges with respect to memory storage layers. We would probably need HBM [high-bandwidth memory] capacity and eventually even some CLX [Compute Express Link] kind of capacity.

The new system would add to Leonardo’s current design—the general-purpose partition and the booster partition. We might complement those partitions with an AI-LLM partition and an inference service partition, for example.

As I mentioned before, we also plan to connect quantum computers to Leonardo. In fact, we plan to integrate two quantum computers. One is an educational system, which will be installed soon. This will be around 10 qubits, and the procurement is in progress. The idea is that the technology for the educational system will be based mostly on superconducting qubits.

The other will be a production quantum computer, which will be co-funded by EuroHPC. The total investment will be around €200 million. The procurement will be open soon. The idea is to use neutral atom technology.
How did the MareNostrum series begin?

Sergi Girona: In 2004, Mateo Valero, the current director of BSC, established a cooperative agreement with IBM to run a joint research center. IBM had established two tracks to build two different supercomputers that would both be counted among the top 100 in the world. Its accelerated partition appears in the most recent Top500 list as the world’s eighth-biggest supercomputer, and the second-biggest in Europe. And its general-purpose partition is listed as the nineteenth-biggest supercomputer in the world.

EE Times Europe spoke to Sergi Girona, BSC director of operations and responsible for MareNostrum 5 operations, to find out more about how it started, the types of applications it runs and future updates.

For the other track, they wanted to build and run a commodity cluster, and they wanted that to be outside the U.S. They had already been collaborating with Mateo and his research group since 2000—and so they chose to put the commodity cluster in Barcelona. That agreement was formalized on March 10, 2004.

The objective was to install the system in the vicinity of the Politecnical University of Catalonia campus, because that’s where Mateo’s group was—and they wanted to do it within four months to allow enough time to get into the top 100 by November 2004. The Torre Girona chapel had been desacralized in 1985 and was in the vicinity. So that’s where we built the data center, and we did it in four months. (By the way, the name of the chapel has no relation to my last name. That is merely a coincidence.)

We named our cluster MareNostrum, which is the Latin name for the Mediterranean Sea. IBM loves the blue theme and Barcelona is on the Mediterranean, which is blue. So that name joins the two cultures.

MareNostrum 1 was completed in 2004 and achieved the fourth ranking in the Top500 list in November 2004. And as predicted, Blue Gene/L displaced Earth Simulator as the world’s top supercomputer in November 2004.

We are now on MareNostrum 5, which was inaugurated in 2023 and is 23× more powerful than its predecessor, MareNostrum 4—and nearly 10,000× more powerful than MareNostrum 1. The new MareNostrum 5 also has more storage capacity—650 PB [peta-bytes], a big increase from the 15 PB that were available in MareNostrum 4.

With an interconnection network based on InfiniBand NDR200, MareNostrum 5 allows more than 8,000 nodes to exchange information very efficiently to solve more complex problems.
problems. More services are needed for the operation of MareNostrum 5—including refrigeration and electrical transformers. These new services occupy almost 5x as much space as they did in MareNostrum 4: about 2,000 square meters.

By the way, BSC does much more than just MareNostrum. It’s a research and service center, with more than 1,000 people working in different domains—including computer science, life sciences, Earth sciences and engineering. We created the supercomputer within BSC not only to provide access to the users at the Spanish and European level but also to support the development of special applications that have a worldwide impact. BSC is a joint public consortium, which includes the Polytechnical University of Catalonia, the Catalan government and the Spanish Ministry of Science, Innovation and Universities.

EE TIMES EUROPE: What are the main architectural features of MareNostrum 5?
Girona: For MareNostrum 5, we decided to continue with what we started in MareNostrum 4, which is to base the system on several connected clusters. No single architecture solves all user problems, and this concept gives us the flexibility we need to address a range of application domains.

But while MareNostrum 4 was designed mostly for general-purpose processing, MareNostrum 5 is a bigger system with more capacity to support the most intensive processing. At the same time, it continues to serve commodity requirements. It has a general-purpose partition, which consists of 90 racks with 72 nodes. Each node has two sockets of Sapphire Rapids 56 cores. And each of those 90 racks uses about 60 kW. We made the decision to continue using a significant part of our investment in money, power and space for this commodity cluster, because there is still a big demand to run legacy code that requires these processors.

But we have a smaller partition in terms of physical size, which consists of 35 racks, with four Nvidia Hoppers [H100s] in each node. Each rack uses about 80 kW. So the general-purpose partition uses a total of about 4.5 MW, and the accelerated partition uses about 2.5 MW. The compute capacity is significantly different between the two partitions.

MareNostrum 5’s general-purpose partition is the world’s largest based on the popular x86 computing architecture, with a peak performance of 45.4 petaFLOPS. The accelerated partition, the third most powerful in Europe and eighth in the world, has a peak performance of 260 petaFLOPS. It has 4,480 state-of-the-art Nvidia Hopper100 chips, each about 8 cm² in size. To give you a sense of how far we’ve come, each of those chips is more powerful than the entire MareNostrum 1 installed in 2004, which occupied the entire 180-square-meter Torre Girona chapel and was the fourth-most-powerful in the world.

EE TIMES EUROPE: How do you decide which applications to run, and what are some of those applications?
Girona: The total cost of ownership for MareNostrum 5 is about €205 million over a five-year period. This includes both the initial investment [capex] and the cost of operations [opex]. EuroHPC contributes 50% of the total costs, and Spain, Portugal and Turkey together contribute the remaining 50%. Access time is allocated based on financial contribution, so EuroHPC decides on half the access time, and Spain, Portugal and Turkey decide on the other half.

In each case, an access committee is used. People submit applications, which are ranked and evaluated for approval. The system is mostly devoted to open science, but there is remaining capacity for industrial applications. So EuroHPC has an access committee—as for Spain, Turkey and Portugal, they each have their own.

In the case of Spain, decisions are made based on external advice. It’s not the center that makes access decisions, but external users who do peer reviews to decide on access time. We let the scientists guide us on how MareNostrum is used. But we do make sure it is not oversubscribed, so people who are granted time don’t have to wait.

The system is huge, so we don’t have to limit the use to one domain at the expense of others. For example, it’s used for large language models, as well as biomedical, energy, industrial and automotive applications. But
The second challenge is that the systems are more complex today, requiring more specialized components. Supercomputers are water-cooled, which means they need hoses and pipes that withstand high temperatures and need to be certified. And the components are very dense, which makes it difficult to build the bigger system and to remove and replace parts.

Today’s supercomputers are unique. The cost of the systems is very high, and the supplies are very scarce. MareNostrum 1 was based on the PowerPC 970, which was the processor used for Apple Macintosh at the time. It was very easy to buy those general-purpose processors.

Today, we’re using Intel Sapphire Rapids with 56 cores, which, because it isn’t sold to the general market, is in limited supply—and the same is true for the Hopper 100. We have a total of 4,480 H100s, which is a big part of what’s available on the market. When one is damaged and you need a new one, it might not be easy to get a replacement quickly.

Cables are also a challenge. The number of cables and the different lengths make it very challenging to keep what you need in stock. You have to find a reliable supplier to get the parts you need.

Another obstacle that slows things down is the software that’s used on supercomputers. We have been too focused on maintaining legacy code, when instead we should be rethinking the ways we solve problems. If you could change some of the algorithms to take advantage of the newer architectures, you could create new opportunities.

The major objective is to support science and innovation in Spain and the rest of Europe. Many of the users do applied research in collaboration with industrial players to develop new methods and algorithms to make advances in science and technology.

We’ve supported many applications that achieve breakthroughs in science and technology. For example, we dedicated parts of the system to help develop vaccines for the coronavirus and to solve distribution problems to get the vaccines to people. Another example is the work the ESA [European Space Agency] has been doing to build a 3D map of the universe.

By increasing the computing power, system memory and number of cores, MareNostrum 5 will help solve more and more complex problems. For example, climate change simulations will be able to have higher resolution, moving from representing phenomena on spatial scales of hundreds of kilometers to include processes occurring on scales of a few kilometers, making predictions much more accurate and reliable.

It will also be possible to tackle much more complex problems of artificial intelligence and big data analysis. For example, it will be possible to generate massive language models by training much larger neural networks with hundreds of billions of parameters, using infinitely larger datasets than today.

MareNostrum 5 specifically targets European medical research, which includes the design of new drugs, the development of vaccines and the simulation of virus propagation. We also expect it to be an essential tool for materials science and engineering. For example, it will run simulations for new aircraft designs.

We’re seeing increasing demand in several other domains. This includes energy in general and nuclear fusion in particular. For example, for wind farms, you have to model the behavior of the wind, which is very complex; and for nuclear fusion, you have to analyze the behavior of plasma inside the reactor. The demand also includes large language models and physical and chemical simulations and modelization at the highest level. For example, personalized medicine requires more complex analysis to adapt treatment to a specific patient. To really analyze the behavior of an individual patient’s heart or to design drugs that are tailored for specific patients, exascale systems will be required.

EE TIMES EUROPE: What are some of the obstacles slowing down the evolution of supercomputers?

Girona: Today, it’s really hard to get a supercomputer up to the state of production on time. From my experience with systems in the past, we set a date and managed to have the system in operation on that date. That is nearly impossible now.

The first challenge is that materials and capacities are lacking. Supply chains still haven’t recovered since the pandemic, and different events around the world continue to produce cascading effects. We saw this recently with the Suez Canal crisis and with the earthquake in Taiwan in early April, when TSMC and other manufacturers temporarily stopped production of semiconductors.
With cybersecurity high on everyone’s to-do list, it is no surprise that the question of memory-safe embedded systems keeps popping up. The Biden-Harris administration has been particularly vocal on this topic, sounding the alarm in 2021 and again in early 2024. Its latest communication, titled “Back To The Building Blocks: A Path Toward Secure and Measurable Software,” highlights the increasing number of memory-safety vulnerabilities and recommends that developers swap out memory-unsafe languages, notably C and C++, for memory-safe alternatives.

According to the U.S. National Security Agency (NSA), memory-safe languages include C#, Go, Java, Ruby, Rust and Swift. In collaboration with partners, the NSA has published a guidance document for top-level executives to assist in eliminating memory-safety vulnerabilities from their companies’ products.

An often-quoted research statistic on memory-safety common vulnerability exposures (CVEs) comes from Microsoft, which found that memory-safety issues accounted for more than 70% of CVEs from 2006 to 2018. Although the report covers only Microsoft C/C++-based products and not embedded systems, it highlights the extent of the cybersecurity challenge.

C DOMINATES EMBEDDED DEVELOPMENT

Written in the early days of the UNIX operating system for Digital Equipment’s PDP-11, C has been around for more than half a century. Though it took a while for it to gain popularity among the embedded development community, C/C++ dominates the embedded landscape today. Computer scientists Brian W. Kernighan and Dennis M. Ritchie—the latter created the language in 1972 at Bell Labs—described C as a “general-purpose programming language which features economy of expression, modern control flow, data structures and a rich set of operators.”1 Performance was undoubtedly a key consideration, but the initial goals for the language had little to do with security.

There is no doubt about the need for memory safety in an embedded system, but the implications and ramifications of switching to another language are dramatic.

The half-century-old programming language is showing its security vulnerabilities, but swapping out the embedded community’s common tongue for an alternative like Rust would not be straightforward.

Poor coding practices and limited code testing can result in embedded systems with security vulnerabilities. There’s an argument to be made, however, that with improved attention to coding standards, more extensive testing, and application of static and dynamic code analysis tools, there may be no need to suffer the trauma of a language change.

WHAT MAKES C MEMORY-UNSAFE?

Buffer overflows and poor memory-management routines are the most obvious and frequently exploited vulnerabilities. An overflow, such as when an application accesses an array outside its defined bounds, is easy to replicate. An adversary with sufficient skills can use the resulting memory failure to inject malicious code.

The code in Figure 1 dimensions an array, ary, with five elements and writes an index counter to each element. The readback, however, reads back one element beyond the sized array (Figure 2). The code

```c
#include <stdio.h>

int main(void) {
    int ary[5];
    for(int count = 0; count < 6; count++) {
        ary[count] = count;
        printf("writing %d: %d\n", count, ary[count]);
    }
    for(int count = 0; count < 6; count++) {
        printf("reading %d: %d\n", count, ary[count]);
    }
    return 0;
}
```

Figure 1: Example C code that reads beyond the boundaries of an array (Source: Robert Huntley)

```bash
[ubuntu@ubuntu ~]$ ./memoryunsafe
writing 0: 0
writing 1: 1
writing 2: 2
writing 3: 3
writing 4: 4
reading 0: 0
reading 1: 1
reading 2: 2
reading 3: 3
reading 4: 4
reading 5: 43699
[ubuntu@ubuntu ~]$ 
```

Figure 2: The results from running the code, highlighting the arbitrary data (Source: Robert Huntley)
Presented by:

Hosted by:
MAURIZIO DI PAOLO EMILIO
Editor-in-Chief
Power Electronics News

Visit the Power Electronics Forum at Electronica on

Tuesday, November 12th
Wednesday, November 13th
Friday, November 15th

Key Topics:
- Wide Band Gap Devices
- Electrification/Renewable Energies
- Smart Energy & Smart Grid
- And More!

This forum is free for all registered electronica attendees. It will be located in hall A5 - 351 at the Power Efficiency Stage.

For more information visit:
https://www.electronica.de/foren
Is C/C++ Memory-Safe?

Bluefruit Software’s Paul Massey

compiles using GCC and executes without error, but in a real-world application, the arbitrary data 43690 might cause unpredictable application behavior or provide an adversary with confidential information.

Commercial toolchains may offer compile-time options to trap such errors, but at the penalty of compiler performance. Interestingly, if the code is edited to write beyond the array size, compilation fails with a stack smash error.

IS RUST A VIABLE CONTENDER FOR EMBEDDED DEVELOPMENT?

Of all the memory-safe languages identified by the NSA, Rust is the only contender for embedded systems. The argument for its use is that a memory-safe language doesn’t require the developer to implement special memory protection measures; the coding example highlighted above would fail during execution. But while Rust could potentially move embedded development to another level of memory protection and CVE mitigation, the impact is massive for the development community. There is also the fear that entrusting the language with identifying such errors further removes developers from following best coding and testing practices.

EE Times Europe asked three embedded development experts to assess the reality of memory safety in C/C++ and the viability of NSA recommendations to use a memory-safe language like Rust.

ARE THERE WORKAROUNDS FOR C/C++ VULNERABILITIES?

Bluefruit Software (Redtruh, U.K.) is well aware of the memory-safety challenges, and “the way that we’re working around it is with best practices—something that is very important to us,” said director Paul Massey. “We use a test-driven development approach on all our projects. It’s a bare-minimum hygiene practice. In itself, it doesn’t ensure your memory is safe, but there’s a set of sub-practices within how we do the test-driven development that can help mitigate the risks around memory safety.”

Massey said that Bluefruit Software tends to use C++ rather than C and that well-architected object-oriented code can mitigate memory-safety problems. He added that C++ makes code more testable, supporting the company’s use of unit testing together with a comprehensive set of practices.

FEASIBILITY OF A MEMORY-SAFE C COMPILER

Camden Dixie O’Brien, software engineer at Witekio (Lyon, France, and Bristol, U.K.), posited that C’s unsafe reputation might not come down to the language itself. “A lot of the unsafe aspects of C actually have to do with the way that C is specified and standardized,” she told EE Times Europe. “The C standard is full of things with undefined behavior, and that’s to allow for highly performant implementations, like highly performant compilers, to generate really efficient code.

"Unless you stick precisely to the defined semantics of the language and standardized behaviors, you can easily come to a point where you are using undefined behaviors," O’Brien said. “Such behaviors, along with the fact that everyone today uses optimizing compilers, is what really makes C unsafe.”

You could make a memory-safe compiler if you wanted to, O’Brien added, but it would be far slower than the C compilers available today.

“Definitely, and for good reasons. There are always lots of factors to consider, and we view language to be an architectural choice.”

If a customer decides to go with Rust, there are longer-term implications. Massey said customers must consider how they will support the project after the Bluefruit team has delivered it. There are fewer Rust developers than C++ programmers available for hire, and “we are honest with customers about this,” he said, as “it might be a dominating factor for them. Also, the tools and libraries for Rust are not as mature as they are for C++.”

Witekio’s Camden Dixie O’Brien

CODE REUSE, REWRITE CHALLENGES

Another perspective on the complexities of considering a replacement language for C came from Graeme Wintle, co-founder of ByteSnap (Birmingham, U.K.): “Memory-safe languages like Rust are obviously different [but] not that different. Someone with a good grasp of C would appreciate the basic mechanics of a language like Rust. But you will probably need a clean sheet of paper. You’re not going to be able to write anything without going through retraining.”

Wintle pointed to the wealth of resources available for a 52-year-old language like C. “There is so much code there, and a lot of it you don’t really want to reuse, but there is probably a good amount you do want to reuse,” he said. “You’re potentially throwing a lot away, or you’re getting the worst of both worlds by bridging a new, safe environment with some unsafe code on the basis that it’s already there, and we’d need to use it rather than rewrite it all again. It’s a difficult one.”

CUSTOMER TRANSPARENCY ON LANGUAGE DECISIONS

Massey told EE Times Europe, “We are programming in Rust more frequently. It’s becoming more common, but the way that we work with our customers is that they are outsourcing their embedded software to us. We maintain a high degree of transparency in the way that we work with customers, so decisions like language, the approach to memory safety, are things we would talk with them about. We would discuss whether Rust is a good choice on a project-by-project basis.”

Asked whether some customers were more amenable to Rust, Massey answered, “Definitely, and for good reasons. There are always lots of factors to consider, and we view language to be an architectural choice.”

ByteSnap’s Graeme Wintle
CODING STANDARDS
Wintle cautioned developers against overdependence on a memory-safe language. “It’s like having a cushion—the fact that something else is looking out for problems for you. I think that’s a bit dangerous. You have to be careful with what some of the software tools can provide, too,” as the tools’ promised functionality can likewise provide a false sense of security.

EE Times Europe asked Wintle whether the diligent use of formal coding practices, such as the MISRA C standard, would be beneficial. “I think coding styles change between different environments, organizations and systems,” he said. “The general practices are the same, so making sure that things are modular, and are able to be unit-tested as a separate entity rather than as a whole system, is something that everyone would agree on.

“If you apply at least a basic level of coding practices, I think it really helps, but you’ve got no control over someone else’s coding practice or style,” Wintle added.

Massey called MISRA C a good set of standards but cautioned that there’s still a human element to consider, noting that “under MISRA, you’re allowed to exempt yourself from any rule you like, as long as you give a justification.”

For example, the standard says not to use dynamic memory allocation but adds that if you do, you should do so in a particular way. “Engineers can still make the decisions to use riskier memory techniques,” Massey said.

A HYBRID APPROACH
O’Brien said the wisdom of adopting a hybrid approach based on C/C++ and Rust “depends on the project,” adding, “One thing that isn’t talked about enough with embedded development, and specifically with languages like Rust, is that a lot of embedded code is inherently unsafe.

“This is because when you’re talking to hardware peripherals, for example, you are writing data into memory addresses that are inherently unsafe,” she explained. “There is no way that it can be recognized as safe by a compiler, since it is unable to systematically decide what’s safe and unsafe. For example, in Rust, if you are writing firmware for a sensor, you’re going to have to wrap everything in unsafe blocks, so the compiler allows it.

“I’d say that Rust is actually less safe than C, because in C, the language is all designed around writing unsafe code,” O’Brien said.

MAKING THE COMMERCIAL CASE
Don’t expect the embedded community to ditch C/C++ for a memory-safe alternative anytime soon. Clearly, embedded developers are not of one mind about adopting a language like Rust. The scale of change required by stakeholders is staggering. However, Rust is clearly a viable contender for embedded systems, as Bluefruit’s Massey observed.

Wintle looked to inject a dose of reality into initiatives to improve the messaging around a language swap: “Where there is a commercial reasoning behind swapping, other than [just] the ‘need’ to, it would make sense. But the benefits are not guaranteed, because if you follow good programming practices, you should end up in the same place” either way.

“As for initiatives, I don’t see how it would work, other than some form of financial encouragement, but that feels the wrong way around,” Wintle said. “It feels like politics is driving industry, which is not the right way to do this.”

REFERENCE

The Tiobe Index has collected nearly 25 years of data about programming languages and tracked their popularity over time. (Source: The Tiobe Index, May 2024)
Three MCU Innovations at Embedded World 2024

By Robert Huntley

Embrolled world, now in its 22nd year, is the main event for the European embedded community, but it also has global reach. This year’s conference drew 1,100 exhibitors from almost 50 countries and 32,000 visitors from roughly 80 countries, underscoring "the significance of embedded world as the key meeting place for the embedded community," as the event’s director, Benedikt Weyerer, said in his comments at the show’s conclusion.

Microcontrollers continue to be the bedrock of the embedded industry, providing the essential building blocks for any design. EE Times Europe spoke with three MCU companies that made announcements at embedded world 2024.

Low-Latency Bluetooth Channel Sounding Drives Ranging Applications

Sujata Neidig, marketing director for wireless connectivity at NXP Semiconductors (Eindhoven, Netherlands), spoke with EE Times Europe about the launch of its MCX W series of wireless MCUs. Adding to the recently announced MCX A and MCX N series devices, the multiprotocol MCX W is claimed to be the first wireless MCU to support Bluetooth Channel Sounding.

"Bluetooth Channel Sounding is available on the MCX W71x family and offers a more accurate ranging capability, and it is designed for secure access, digital key and proximity services," Neidig said. "With Bluetooth Channel Sounding, the MCX W can achieve ±0.5 m accuracy, utilizing time-of-flight and round-trip phase ranging modes and a fast response time. A dedicated on-chip localization compute engine accelerator keeps latency to less than 10 ms."

The MCX W features a dedicated Arm Cortex-M3-based, 2.4-GHz 802.15.4 radio subsystem complete with a +10-dBm power amplifier and a low-noise amplifier, freeing the primary Arm Cortex-M33 core from running wireless stacks. It is over-the-air upgradable to support future wireless standards, performance features and security safeguards. Supported protocols include Matter, Zigbee, Bluetooth and Thread. The device also supports Bluetooth Low Energy (BLE) 5.3 high-speed, long-range and advertising extensions. Wi-Fi and BLE coexistence are built-in to achieve a robust dual PAN capability.

Security is an omnipresent requirement in any embedded design, and the MCX W series integrates NXP’s Edgelock secure enclave features across the range.

Next-Generation 8-Bit MCU Integrates USB Power Delivery

Microchip (Chandler, Arizona) launched its latest 8-bit USB MCU family at the show. "The AVR DU family of 8-bit microcontrollers is the latest AVR-based microcontroller to integrate USB connectivity and now features a higher power delivery capability than previous devices. It’s the sixth 8-bit MCU we've introduced in the last 12 months," Odd Jostein Svendsli, Microchip's business development manager for 8-bit MCUs, told EE Times Europe.

"The series delivers up to 15 W—a capability not found in similar-class MCUs—making it suitable for USB-C charging with currents up to 5 A at 5 V [and therefore] making it an excellent choice for products like portable power banks and rechargeable toys," Svendsli said.

The AVR DU family incorporates Microchip's established program and debug interface disable security feature, allowing developers to strengthen the device’s defense capabilities against malicious attacks. When enabled, the programming interface blocks unauthorized attempts to change or erase the firmware. Svendsli said the MCU’s six-channel core independent peripheral (CIP) event system "provides a deterministic response without using the core and interrupt service routines," adding that the available functions include three 16-bit timer/counters; an analog comparator; a 10-bit, 170-kfps A/D converter; and up to 25 digital I/O pins. "The CIP blocks provide a flexible and configurable set of functions that require little or no code, consume minimal power and require less RAM and flash than implementing them in software," he said.

Focusing on Matter

Matt Maupin, senior product marketing manager for Bluetooth and 802.15.4 technologies at Silicon Labs (Austin, Texas), told EE Times Europe that the company’s low-power xG26 series of SoCs and MCUs is focused on meeting the increasing market demand for Matter-based connectivity. "Our new xG26 family of multiprotocol 2.4-GHz wireless SoCs and microcontrollers targets a broad range of IoT applications, from home automation and security to industrial IoT, smart cities/buildings and commercial use cases," Maupin said.

"We're all-in on Matter, an evolving standard that is only 18 months old, so with this new series, we're meeting the demand for more overhead of resources, anticipating new features and long-life-cycle product use cases," he said.

"As the standard evolves, developers might be concerned that their devices will be unable to accommodate new product features and security updates."

The xG26 series comprises the multiprotocol MG26 SoC, the BLE BG26 SoC and the PG26 MCU. "Compared to the xG24 family, for the xG26, we've doubled the amount of flash and RAM," Maupin said. "The increased memory helps developers future-proof their devices, providing more firmware headroom."

Silicon Labs also doubled the number of general-purpose I/O pins, he said. "Developers now have access to up to 64 GPIO pins, providing the ability to interface to more peripherals [and thereby] significantly easing system integration challenges. The devices offer a higher-performance compute, with the tri-core device featuring a 78-MHz Arm-Cortex-M33 application code and dedicated cores for the radio and security subsystems. Also, an integrated AI/ML accelerator capable of 8+ faster algorithm processing and using as little as a sixth of the power is available to offload compute-intensive tasks from the application core."

The MG26 supports Matter over Thread as well as Zigbee, Thread, BLE 5.3 and Bluetooth Mesh. The MG26 series offers up to 3,200 kB of flash and 512 kB of RAM. The radio subsystem of the BG26 and MG26 provides up to +19.5-dBm transmit power and has a receive sensitivity of ~97.6 dBm for BLE 1-Mbps communications. Silicon Lab's Secure Vault functionality is incorporated across the xG26 family.
A Blueprint for the Integrated Photonics Industry?

By Rebecca Pool

A roadmap from PhotonDelta and MIT Microphotonics Center is set to help PIC players navigate the tech complexities and build a global supply chain.

Earlier this year, Netherlands-based integrated photonics industry hub PhotonDelta and Massachusetts Institute of Technology's (MIT's) Microphotonics Center published the latest Integrated Photonics Systems Roadmap—International (IPSR-I). Put together with input from more than 400 industry players, including Airbus, Meta, NASA, Dupont Electronics, General Motors, the European Space Agency and VodafoneZiggo, the roadmap identifies technology challenges and lays out how to build a global supply chain to drive volume photonic integrated circuit (PIC) manufacturing forward. As PhotonDelta CTO Peter van Arkel put it, "Reaching a consensus was challenging … [but] has definitely been worth it."

The last IPSR-I before this version was delivered in 2021. According to van Arkel, the new version follows rapid PIC development and provides insight into many industries, including aerospace, datacom and the up-and-coming agrifood and 3D sensing-LiDAR sectors. Looking at the roadmap, van Arkel also said technology bottlenecks "are everywhere", and it is clear that heterogeneous integration remains a thorny, and pervasive, issue.

“We’re talking lasers, detectors, waveguides and modulators, but also photonics and electronics in general; these all need to come closer together—and that’s across all applications,” he said. "This is one of the core messages of the roadmap."

According to van Arkel, the roadmap also shows that tackling wafer-level testing and packaging challenges will be instrumental to driving manufacturing volumes forward, and the maturation of silicon photonics process design kits (PDKs) is necessary to accelerate chip and packaging developments.

CHARTING THE FUTURE

The IPSR-I roadmap will be welcomed by many in the industry, as tech analysts expect robust growth for silicon photonics and PICs. IDTechEx, for example, predicts the global market will more than double and top US$22 billion across the next decade.

IDTechEx technology analyst James Falkiner believes such a roadmap provides critical insight for any organization deciding whether to invest in the photonics industry. "It’s also a good idea to try to bring everyone together globally [in a roadmap]," Falkiner said. "We have design houses in the U.S., talent in Europe and photonics manufacturing in Asia—with, say, GlobalFoundries and Samsung. At the end of the day, this industry is global."

In line with van Arkel, and as outlined in the roadmap, Falkiner pointed out the need to ensure that PDKs and the manufactured product actually match and to address the photonics industry’s relatively long manufacturing turnaround times.

“If you can reduce a one-year cycle time to three months—like the electronics industry—then you can start getting improvements in your chips much, much faster,” he said. "We really need more cooperation between electronics manufacturing and foundries and photonics. Industry needs to find a solution together, and having a roadmap ... is important here."

So where are the big market drivers for PICs? Van Arkel pointed to datacom and AI data centers and noted that "every telecoms provider is currently working with integrated photonics" for these applications. He also reckons that the rise of autonomous vehicles and accompanying demand for LiDAR sensors could trigger significant market growth for PICs in the next few years.

Falkiner also expects that growing demand for LiDAR will drive PIC manufacturing, as will 5G telecom. However, he believes the real market wins center on today's unprecedented demand for high-performance transceivers, which are based on silicon photonics and PICs and can support the massive data rates required by AI accelerators and data centers. Key industry players, such as Jabil-Intel, Coherent and Infinera, are using PICs in their transceivers, while China-based Innolight has already reached blisteringly fast, 1.6-Tbps speeds with its PIC-based transceivers.

"These transceivers can hit the top end of 800G and by early next year will transmit data at speeds of 1.6 Tbps, reaching 3.2 Tbps by 2026,” Falkiner said. "They are facilitating efficient, high-bandwidth communications between the large racks of AI accelerators that power ChatGPT [and] Microsoft Copilot—and are seeing billions and billions of dollars of investment."

Falkiner expects PIC technologies to continue to be used in high-performance transceivers, the demand for which will persist into the future. As he noted, Nvidia’s GPUs are all oversubscribed, with each one requiring approximately two 800G transceivers to convey data.

"We’re looking at around 15,000 data centers globally, and each one will use hundreds of thousands of units of accelerators, culminating in a massive increase in transceiver demand," he added.

Given these industry developments, scaling the manufacture of PICs surely lies ahead. As van Arkel put it, "We already have foundries full of [photonic] wafers [and] manufacturing at full capacity; I consider this to already be quite high-volume ... You can easily imagine that there's going to be many, many billions of dollars of revenue coming from photonic chip manufacturing soon."

And of course, the IPSR-I global roadmap can help to drive industry in the right direction. "We know there's been a lot of interest in this," van Arkel said. "It has sparked a lot of discussion on integrated photonics ... and at the heart of the roadmap is a global approach for the photonics industry to rally behind to meet the core challenges.”
Delft Quantum Ecosystem Addresses Connectivity Challenges

By Robert Huntley

As the quantum industry advances, there are growing signs of grassroots collaboration forming value chains and establishing all-important ecosystems. Ecosystems are essential in any technology field, from embedded systems to semiconductor manufacturing. They have many characteristics, but two typical identifying attributes are that they foster innovation and collaboration. As quantum organizations, whether creating qubits or developing sensors, start their journey through the forming, storming and norming of team development and innovation, many realize they can’t do it all and can’t do it alone. However, how do you start an ecosystem?

Delft Circuits (Delft, Netherlands), specializing in quantum I/O hardware, has actively contributed to this all-important step. EE Times Europe spoke to Daan Kuitenbrouwer, co-founder and chief commercial officer, to learn how the ecosystem of Delft-based quantum companies started and about Delft Circuits’ involvement in the quantum value chain.

Establishing a Quantum Value Chain

Kuitenbrouwer said the formation of the Dutch quantum scene owes much to the investments made by the Dutch government in the early 2010s. “This drove a lot of academic research with people gaining Ph.D.s, and then at a certain point, my co-founder, Sal [Jua Bosman], and I decided to look deeper and begin with one point of a value chain: the cabling. [In forming our company], we picked a specific part of the value chain, emphasizing it as a key part, not just the cabling. Then other nascent Delft quantum companies realized we had a clear vision, so they focused on solving different pieces of the value chain.”

Kuitenbrouwer noted that discussions with like-minded individuals, such as a friend who had started a quantum control electronics company, Qblox, shaped how the Delft quantum community formed. One aspect of this was wondering what the community might look like in 10 years. “We knew we would need a [cryogenic] fridge, but there are plenty of manufacturers, so we wouldn’t need to do that,” he said. “We realized what we needed would be chips and characterization. In time, QuantWare has become the chip company and Orange Quantum Systems the characterization company.”

Collaborative Ecosystem to Advance Interface Standards

Kuitenbrouwer said the four companies formed the ImpaQT consortium to foster a collaborative ecosystem to develop interoperable and industry-ready components. “With so many different interfaces currently available, it is paramount for the development of utility-scale quantum computers that these interfaces are standardized. Such standardization work should happen not only on paper but also on the ground with actual hardware. ImpaQT is our vehicle to achieve this.”

Quantum Going Cold

EE Times Europe asked Kuitenbrouwer whether he saw the quantum industry heading in any direction regarding qubit creation. “Yes, we see a trend that most quantum applications are going cold,” he said. “Broadly speaking, there are five approaches. Google and IBM use superconducting methods. Then you have semiconductors or spin and photonic quantum computing, where everything happens at room temperature except the readout of the system, which happens to be cold. Neural atoms use lasers in a big vacuum, which is not necessarily always cold; then you have trapped ions, which also tend to be cold.”

Delft Circuits’ Daan Kuitenbrouwer

would be chips and characterization. In time, QuantWare has become the chip company and Orange Quantum Systems the characterization company.”

Collaborative Ecosystem to Advance Interface Standards

Kuitenbrouwer said the four companies formed the ImpaQT consortium to foster a collaborative ecosystem to develop interoperable and industry-ready components. “With so many different interfaces currently available, it is paramount for the development of utility-scale quantum computers that these interfaces are standardized. Such standardization work should happen not only on paper but also on the ground with actual hardware. ImpaQT is our vehicle to achieve this.”

Quantum Going Cold

EE Times Europe asked Kuitenbrouwer whether he saw the quantum industry heading in any direction regarding qubit creation. "Yes, we see a trend that most quantum applications are going cold," he said. "Broadly speaking, there are five approaches. Google and IBM use superconducting methods. Then you have semiconductors or spin and photonic quantum computing, where everything happens at room temperature except the readout of the system, which happens to be cold. Neural atoms use lasers in a big vacuum, which is not necessarily always cold; then you have trapped ions, which also tend to be cold."
I/O TAKES THREE FORMS
EE Times Europe wondered how Delft Circuits keeps its products in step with innovations in a constantly evolving industry. "When we talk about I/O, we consider it three things," Kuitenbrouwer said. "It’s signal transmission, so bringing a signal from A to B; it’s signal conditioning, filtering the signals and adjusting them; and then it’s interfacing, which is getting the signal onto the cables and from the cables to something else. When we look at the three approaches that require the most cables, which are superconducting, spin and photonics, what they have in common is the need for signal transmission from a hot place to a cold place."

As the quantum industry advances, there are growing signs of grassroots collaboration forming value chains and establishing ecosystems. Delft Circuits has contributed to this step.

SIGNAL TRANSMISSION CHALLENGES
Kuitenbrouwer told EE Times Europe about signal transmission challenges between places of significantly different temperatures. "At the cold place, you have a limited cooling power, so it’s hard to subtract the heat. And since cables are metal, they bring heat into the system. You need to bring in the signal but not the heat. That’s one of the major tricks. You also don’t want to lose your signal integrity."

To compound the connectivity challenges, Kuitenbrouwer said, "If you need to wire up a quantum computer with two qubits, you need six cables. That’s not an issue. But if you start to go through high numbers of qubits, say 100 qubits, you need 300 cables. That’s already a lot more, and you begin to get more issues of space—where does it fit, how does it fit in the system? Then there is the thermal question. How can we avoid bringing heat in and maintain signal integrity?"

STRIP-LINE INTEGRATED FLEX COMPONENTS
EE Times Europe asked Kuitenbrouwer how Delft Circuits overcame the challenges of quantum computing connectivity. "Our technology uses a strip-line structure, and the parts we add are integrated into flex components," he said. "These are low-pass filters, attenuators and IR filters."

He observed that, typically, many quantum applications send a signal via a connector to another component and then another connector and component arrangement. However, Delft has a more integrated approach: "We use the structure and special geometries to make our low-pass filters, or we make resistant networks to make our attenuators, or we adjust the dielectric to make IR filters. Our in-flex integrated components are one of the major innovations that achieve two things. First, it allows you to scale up because you need much less space for all the components. The second thing is that it reduces the number of cable problems you get, because at every interruption, at every interface, you have an interruption that might break. And as we integrate everything in one monolithic flex, you don’t have these interruptions, so you’re much less prone to failure."

UNDERSTANDING MATERIALS SCIENCES IS CRUCIAL
Kuitenbrouwer said Delft’s products are designed for operation to at least 10 GHz and up to 15 GHz. He sees some demand for operation up to 20 GHz but noted that it varies from customer to customer. Most of the primary superconducting companies have qubits operating at about 6 to 8 GHz. They cite physical resonant frequency reasons for operating at those frequencies, with no benefit in performance achieved by attempting to operate at higher frequencies.

When asked about the innovation challenges that Delft Circuits faced creating its product lineup, Kuitenbrouwer noted, "Our superconducting flex is one of our key innovations. A superconductor is needed to reduce the amount of heat you bring in and its resistance. Our normal cables consist of a stack-up of different materials, the most important being silver and the dielectric Kapton. We use titanium and Kapton for the superconducting cables. Getting the stacking of all the materials and sizes precisely right, it’s a materials science question."

Kuitenbrouwer noted that microwave testing at such low temperatures compounds the challenges. Performing short, open and load VNA testing requires cycling the fridge between each test, which takes a lot of time and is not precise. Commenting on the need to build systems, research material properties and develop test mechanisms and protocols, he said, "None of it is straightforward; we had to develop everything so that we could make sure that we could reliably make something."
Webinar series: Trends in xEV power conversion

Discover how our innovative Si, SiC, GaN power switches revolutionize the OBC market by addressing challenges in automotive charging.

Key takeaways:

- Overview of disruptive changes in xEV power conversion
- Introduction of CoolSiC™ Automotive MOSFETs 750 V
- How to select the right technology & package
- Demonstrators exemplifying key trends such as top-side cooling

Scan QR code for more details & registration!