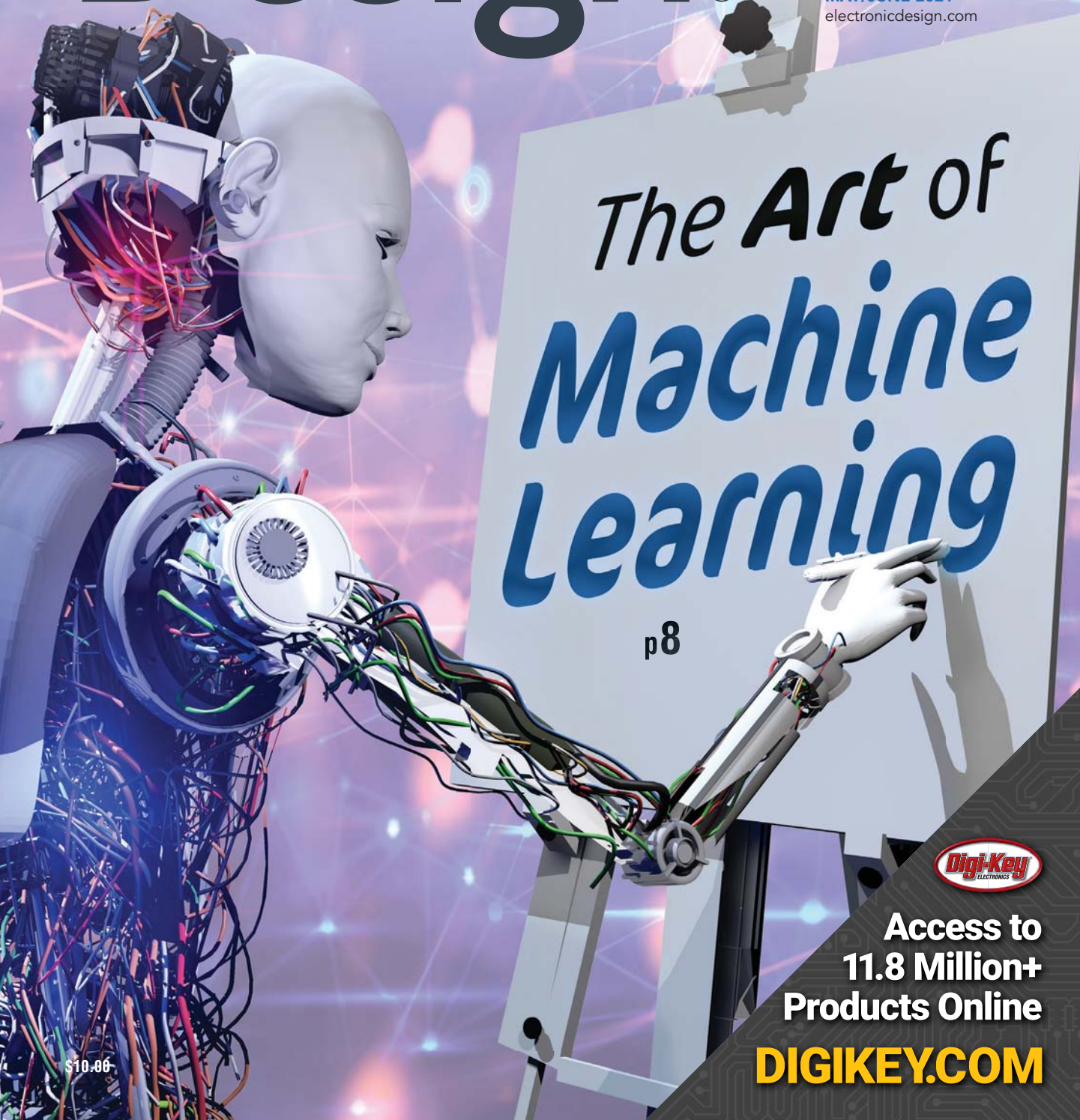
Electronic Design®

THE AUTHORITY ON EMERGING TECHNOLOGIES FOR DESIGN SOLUTIONS

MAY/JUNE 2021
electronicdesign.com



The Art of Machine Learning

p8

$10.00

11.8 MILLION+ PRODUCTS ONLINE | 1,700+ INDUSTRY-LEADING SUPPLIERS | 100% AUTHORIZED DISTRIBUTOR

Find It Here

1.800.344.4539

DIGIKEY.COM

Digi-Key ELECTRONICS

Experience the future of fan technology *with*:

Axi*Force*

- Completely redesigned housing and impeller
- Up to 6 dB(A) lower sound level
- Efficiency increase by up to 18%
- Environmental protection level up to IP68
- Powerful air performance curve

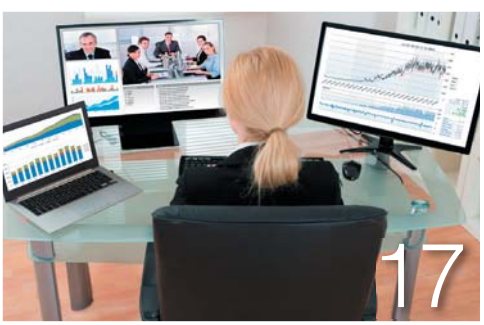
ebmpapst

engineering a better life

Contact sales@us.ebmpapst.com for more info.

IN THIS ISSUE

FEATURES



COLUMNS & DEPARTMENTS

FOLLOW US ON FACEBOOK AND TWITTER

(ElectronicDesign)

(@ElectronicDesgn)

EDITORIAL MISSION:
To provide the most current, accurate, and in-depth technical coverage of the key emerging technologies that engineers need to design tomorrow's products today.

ELECTRONIC DESIGN (ISSN 0013-4872) is published bi-monthly in Jan/Feb, Mar/Apr, May/June, July/Aug, Sept/Oct and Nov/Dec by Endeavor Business Media, 1233 Janesville Ave., Fort Atkinson, WI 53538. Paid rates for a one-year subscription are as follows: $60 U.S., $90 Canada, $120 International. Periodicals postage paid at Fort Atkinson, WI and additional mailing offices. Printed in U.S.A. Title registered in U.S. Patent Office.

Altech®

Largest Selection of DIN Rail Mount Electronics Enclosures

Features

- 8 different series with over 240 parts in many sizes
- Integral PC Board Guides
- DIN Rail and Panel Mount
- Fixed and Pluggable Terminals
- Customization available

KS4400 Series
12 to 96 Terminals

In-Rail Bus
Universal Bus System

KO4700 Series
8 to 132 Terminals

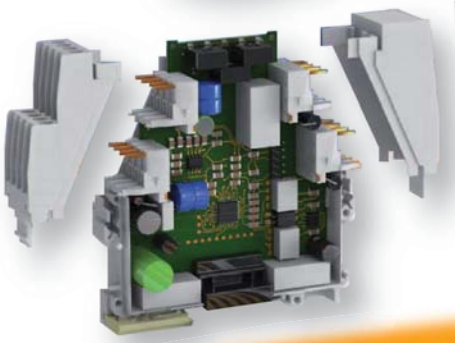
KU4100 Series
0 to 96 Terminals

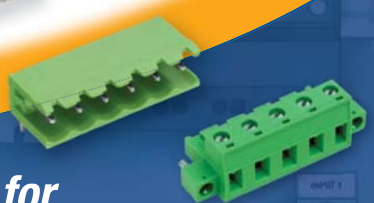
K70 Series
8 to 32 Terminals

KU4000 Series
8 to 64 Terminals

KO4300 Series
16 to 64 Terminals

KO4070 Series
4 to 16 Terminals





Time Saving Push-In Technology

Easy wire insertion

No tool required !

- Unique ergonomic design for quick and easy mounting
- 3.50, 3.81, 5.00 and 5.08mm pin spacing
- Screw flange version for additional stability
- Imprinting and coding available

Your Source for Printed Circuit Board Terminal Blocks

Fixed

- Tubular Screw Clamp
- Screw-Cage Clamp
- Dual Pin Screwless Spring Clamp

Pluggable

- Tubular Screw Clamp Plug & Pin Headers

Mini Pluggable

- Screw-Cage Clamp Plug & Headers

Standard Pluggable

- Screw-Cage Clamp Plug
- Headers
- Double Level Headers

Altech Corp.®

Your Source for Automation & Control Components

info@altechcorp.com • 908.806.9400

AltechCorp.com

ElectronicDesign

MAY/JUNE 2021

EDITORIAL

SENIOR CONTENT DIRECTOR: **BILL WONG** bwong@endeavorb2b.com
SENIOR EDITOR: **DAVID MALINIAK** dmaliniak@endeavorb2b.com
MANAGING EDITOR: **ROGER ENGELKE** rengelke@endeavorb2b.com
SENIOR STAFF WRITER: **JAMES MORRA** jmorra@endeavorb2b.com
CONTRIBUTING EDITOR: **LOUIS E. FRENZEL** lfrenzel@endeavorb2b.com

ART DEPARTMENT

GROUP DESIGN DIRECTOR: **ANTHONY VITOLO** tvitolo@endeavorb2b.com
ART DIRECTOR: **JOCELYN HARTZOG** jhartzog@endeavorb2b.com

PRODUCTION

GROUP PRODUCTION MANAGER: **GREG ARAUJO** garaujo@endeavorb2b.com
PRODUCTION MANAGER: **DEANNA O'BYRNE** dobyrne@endeavorb2b.com

AUDIENCE MARKETING

USER MARKETING MANAGER: **DEBBIE BRADY** dmbrady@endeavorb2b.com

FREE SUBSCRIPTION / STATUS OF SUBSCRIPTION / ADDRESS CHANGE / MISSING BACK ISSUES

OMEDA **T** | 847.513.6022 **TOLL FREE** | 866.505.7173 **FAX** | 847.291.4816 electronicdesign@omeda.com

SALES & MARKETING

REGIONAL SALES REPRESENTATIVES:

AZ, NM, TX: **GREGORY MONTGOMERY** gmontgomery@endeavorb2b.com

AL, AR, SOUTHERN CA, CO, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MI, MN, MO, MS, MT, NC, ND, NE, OH, OK, SC, SD, TN, UT, VA, WI, WV, WY, CENTRAL CANADA: **JAMIE ALLEN** jallen@endeavorb2b.com

CT, DE, MA, MD, ME, NH, NJ, NY, PA, RI, VT, EASTERN CANADA: **ELIZABETH ELDRIDGE** eeldridge@endeavorb2b.com

AK, NORTHERN CA, NV, OR, WA, WESTERN CANADA: **STUART BOWEN** sbowen@endeavorb2b.com

INTERNATIONAL SALES:

GERMANY, AUSTRIA, SWITZERLAND: **CHRISTIAN HOELSCHER** christian.hoelscher@highcliffemedia.com

BELGIUM, NETHERLANDS, LUXEMBURG, UNITED KINGDOM, SCANDINAVIA, FRANCE, SPAIN, PORTUGAL: **LARA PHELPS** lara.phelps@highcliffemedia.com

ITALY: **DIEGO CASIRAGHI** diego@casiraghi-adv.com

PAN-ASIA: **HELEN LAI** helen@twoway-com.com

PAN-ASIA: **CHARLES LIU** liu@twoway-com.com

ARTICLE REPRINTS: reprints@endeavorb2b.com

LIST RENTALS/ SMARTREACH CLIENT SERVICES MANAGER: **MARY RALICKI** mralicki@endeavorb2b.com

DIGITAL

SENIOR DIGITAL INNOVATION & STRATEGY DIRECTOR: **RYAN MALEC** rmalec@endeavorb2b.com

DESIGN ENGINEERING & SOURCING GROUP

EVP, DESIGN & ENGINEERING GROUP: **TRACY SMITH** tsmith@endeavorb2b.com
GROUP CONTENT DIRECTOR: **MICHELLE KOPIER** mkopier@endeavorb2b.com
VICE PRESIDENT OF MARKETING SOLUTIONS: **JACQUIE NIEMIEC** jniemiec@endeavorb2b.com

ENDEAVOR BUSINESS MEDIA, LLC

331 54th Ave N., Nashville, TN 37209 USA | www.endeavorbusinessmedia.com
CEO: **CHRIS FERRELL**
CRO/CMO: **JUNE GRIFFIN**
CFO: **WILLIAM NURTHEN**
COO: **PATRICK RAINS**
CHIEF ADMINISTRATIVE AND LEGAL OFFICER: **TRACY KANE**
EVP KEY ACCOUNTS: **SCOTT BIEDA**
EVP SPECIAL PROJECTS: **KRISTINE RUSSELL**

ENDEAVOR BUSINESS MEDIA

Electronic Design | Machine Design | Microwaves & RF | Hydraulics & Pneumatics | Source ESB | Source Today | Evaluation Engineering

Fast Pulse Test Systems from Avtech

Avtech offers over 500 standard models of high-speed pulsers, drivers, and amplifiers ideal for both R&D and automated factory-floor testing. Some of our standard models include:

AVR-E3-B: 500 ps rise time, 100 Volt pulser

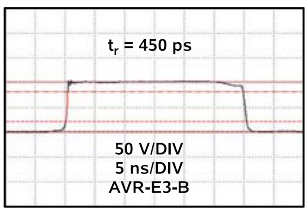
AVRQ-5-B: Optocoupler CMTI tests, >120 kV/us

AVOZ-D6-B: 1000V / 200A, 1-10 us pulser

AV-1010-B: General purpose 100V, 1 MHz pulser

AVO-9A-B: 200 ps t_r, 200 mA laser diode driver

AV-156F-B: 10 Amp current pulser for airbag initiator tests



AVTECH

Nanosecond Electronics
Since 1975

Pricing, manuals, datasheets and test results at:
www.avtechpulse.com

AVTECH ELECTROSYSTEMS LTD.
US & Canada 888-670-8729, Worldwide +1-613-686-6675
info@avtechpulse.com

High Output - Low Risetime Pulsers from Avtech

Nanosecond and faster pulsers covering a wide range of amplitudes!

GPIB / RS-232 / Ethernet standard on "-B" models.

Standard and customized models available.

Ampl	t_{RISE}	Max. PRF	Model
5 V	80 ps	1 MHz	AVP-AV-1S-C
10 V	120 ps	1 MHz	AVP-AV-1-B
20 V	120 ps	1 MHz	AVP-AV-HV2-B
20 V	200 ps	10 MHz	AVMR-2D-B
40 V	150 ps	1 MHz	AVP-AV-HV3-B
50 V	500 ps	1 MHz	AVR-E5-B
100 V	500 ps	100 kHz	AVR-E3-B
100 V	300 ps	20 kHz	AVI-V-HV2A-B
200 V	1 ns	50 kHz	AVIR-1-B
200 V	2 ns	20 kHz	AVIR-4D-B
400 V	2.5 ns	2 kHz	AVL-5-B-TR

AVTECH

Nanosecond Electronics
Since 1975

Pricing, manuals, datasheets and test results at:
www.avtechpulse.com

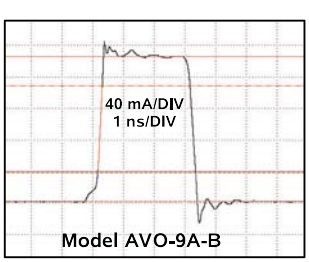
AVTECH ELECTROSYSTEMS LTD.
US & Canada 888-670-8729, Worldwide +1-613-686-6675
info@avtechpulse.com

Pulsed Laser Diode Drivers from Avtech

Each of the 18 models in the Avtech **AVO-9 Series** of pulsed laser diode drivers includes a replaceable output module with an ultra-high-speed socket, suitable for use with sub-nanosecond pulses.

Models with maximum pulse currents of 0.2A to >10A are available, with pulse widths from 400 ps to 1 us.

GPIB, RS-232, and Ethernet control available.



AVTECH

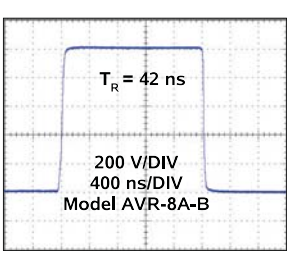
Nanosecond Electronics
Since 1975

Pricing, manuals, datasheets and test results at:
www.avtechpulse.com/laser

AVTECH ELECTROSYSTEMS LTD.
US & Canada 888-670-8729, Worldwide +1-613-686-6675
info@avtechpulse.com

100 to 1000 Volt Lab Pulsers

Avtech offers a full line of 100, 200, 500, 700 and 1000 Volt user-friendly pulsers capable of driving impedances of 50 Ω and higher. The **AVR Series** is suitable for semiconductor and laser diode characterization, time-of-flight applications, attenuator testing, and other applications requiring 10, 20, or 50 ns rise times, pulse widths from 100 ns to 100 us, and PRFs up to 100 kHz. GPIB, RS-232, and Ethernet ports are standard.



See additional test results at:
http://www.avtechpulse.com/medium

AVTECH

Nanosecond Electronics
Since 1975

Pricing, manuals, datasheets and test results at:
www.avtechpulse.com/medium

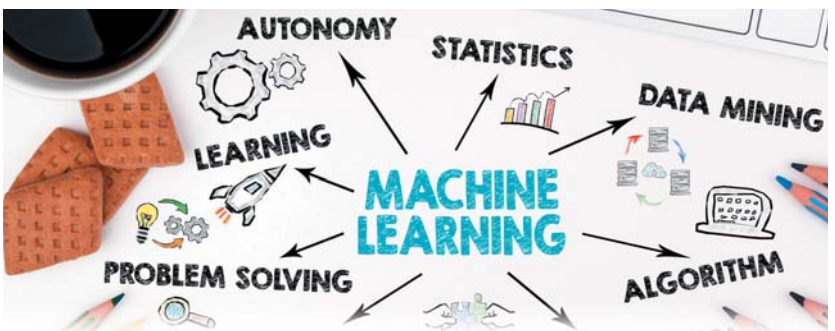
AVTECH ELECTROSYSTEMS LTD.
US & Canada 888-670-8729, Worldwide +1-613-686-6675
info@avtechpulse.com

Editorial

WILLIAM WONG | Senior Content Director
bwong@endeavorb2b.com

Embracing Machine Learning

Applications that utilize machine learning can do amazing things, but they're still just tools.



TERMS LIKE MACHINE LEARNING (ML), artificial intelligence (AI), neural networks, and deep learning abound in articles and application descriptions these days. Dreams of human-like robots and self-driving cars continue to emerge, and we're closer to those now than ever before. However, the former is still a pipe dream and even the latter is farther off than most people would like. Too many assume that the ability to recognize people, gestures, or sentences indicate that cognizant AI is just around the corner.

On the flip side, the past and current AI/ML advances have turned out to be nothing short of astounding. Hardware improvements are the main drivers, but software advances have been even more important. Actually, the parallel nature of many of these improvements has proven beneficial, because single-core processor technology has effectively hit a power barrier.

Like any new technology, lots of questions arise as one learns what technologies are now available, what they can do, how they work, and how they can be incorporated into an application. The challenge with the latest AI/ML solutions is the extremely broad range of what's available. A chart covering this technology looks more like a Banyan tree.

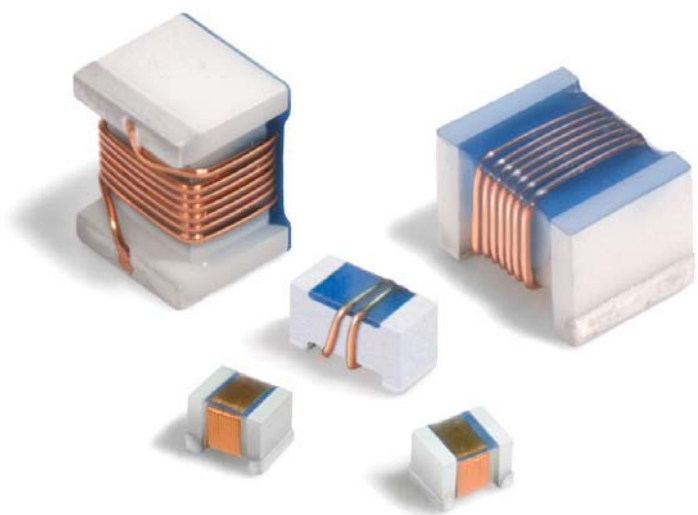
Still, the benefits and applications are significant. Likewise, the hardware and software that can be employed ranges from something running on a standard 8-bit microcontroller to a custom AI/ML chip running on hundreds of drive bays in the cloud. Of course, that little micro might only be working on predictive maintenance of a motor controller while a hardware-accelerated system might be analyzing multiple video streams.

Another challenging aspect with respect to AI/ML is the technology's perpetual expansive growth. It's not uncommon for a new version of a machine-language model compiler to double the performance of an application with no change to the model.

Most applications will not benefit from AI/ML, and it's not a great idea to force-fit a fancy new technology into an established application. On the other hand, taking a new look at an application with an understanding of the various AI/ML methodologies can reveal ways to provide significant enhancements to an application.

It's worthwhile learning more about AI/ML to at least identify hardware and software that may be useful in developing an application. It takes significant effort to get to the point where the technology can be incorporated in an application, though, including determining the advantages to be attained as well as the costs involved. ed

Blue Chip Performers



The industry's highest Q, high SRF, and a wide range of inductance values make our wirewound chip inductors a sure bet!

Engineers love our high-performance, low-cost, wirewound ceramic chip inductors; considered by most to be the best performing wirewounds available.

For example, our 0201HL Series is offered in seven inductance values ranging from 22 to 51 nH – the highest currently offered in an 0201 (0603) package – making them ideal for impedance matching in 700 MHz band LTE and 5G applications.

Our 0402DC, 0603DC and 0805HP all provide the industry's highest Q factors in their respective sizes for super low loss in high-frequency circuits. And the 0402DC offers 112 values from 0.8 to 120 nH, including 0.1 nH increments from 2.8 to 10 nH.

Find out why our customers are so bullish on our wirewound ceramic chip inductors. Order your free samples today at **www.coilcraft.com.**

Coilcraft®

WWW.COILCRAFT.COM

WILLIAM WONG | Senior Content Director

Understanding the Art of MACHINE LEARNING

Though neural-network-based machine learning is escalating in popularity, the mechanics behind it tend to be misconstrued or simply not known at all.

Ekkasit919 | Dreamstime.com

Machine learning (ML) is a hot topic when it comes to almost anything related to computing, from analyzing data in the cloud, to self-driving cars recognizing people and things, to detecting defective PCBs or chips. Like artificial intelligence (AI), ML is a very broad subset of AI that's often mischaracterized even by people with technical backgrounds.

Deep learning is a term that's bandied about these days, but what does it really mean? Typically, it's shorthand for a discussion about deep neural networks (DNNs). We will get into more detail about neural networks, but first a comment about current ML use.

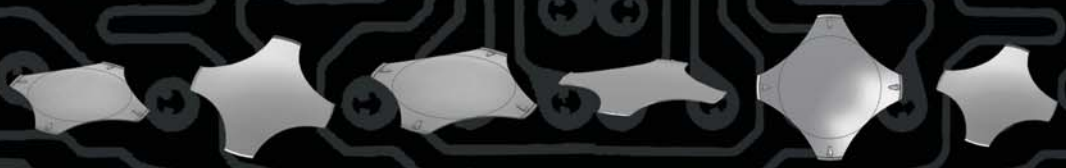
We recently finished up our local Mercer Science and Engineering Fair, which I help manage. As you might guess, a lot of projects employed ML models for various tasks. This is a challenge from numerous points of view.

All of the students were using a predefined ML model to perform a specific task, like identifying a problem. Some actually trained a model, but the general knowledge about ML, their model, and its implications were something neither the students or judges really had a handle on. I wound up giving a number of judges an overview on ML and how they should view it with respect to the projects.

Essentially, the student's use of ML was as a tool and not some cutting-edge AI advance. In general, the understanding of the tool, how it worked, and how to use it was significantly lacking. However, that should not be unexpected given how much students are learning when working on a project that's very new to them. It's great to discover how they can get things done. Still, developers and engineers who plan on building commercial solutions need much more insight because of the responsibility they have for the resulting products and their use.

What I hope to accomplish here is to reveal the types and capabilities of some of the more important machine-learning technologies so that you can figure out what to investigate and what level of interaction is needed to take advantage of DNN ML technology. It can range from using preconfigured models to creating new models for an application.

THERE'S A KEYSTONE IN EVERY GREAT INVENTION.



MODERN BURGLAR ALARMS & THEIR ELECTRONICS

Featuring Metal Key-Pad Dome Switches

Since the 1850s, homeowners have been protecting their homes with automated security systems. Burglar alarm systems are considerably more sophisticated now than they were in the 1850s, offering a wide range of protection from focused burglary protection to ultrasonic movement detection. Today precision electronic components and cloud integration build on the original simple design to give alarm owners maximum peace of mind.

You'll find many of our products in modern burglar alarm systems such as our featured Metal Key-Pad Dome Switches, as well as • Fuse Clips & Holders • LED Hardware • Panel & Computer Hardware • Battery Clips, Contacts & Retainers

IT'S WHAT'S ON THE INSIDE THAT COUNTS

KEYSTONE ELECTRONICS CORP.

Designers & Manufacturers of Electronic Components and Hardware

View our Dynamic Catalog M70 at www.keyelco.com

(516) 328-7500 • (800) 221-5510

See our Blog

The amount of effort between those two is significant, from an afternoon's worth of work to decades of man hours. Many will simply use a product like a smart camera that employs ML without ever dealing with the technology directly. Nonetheless, it helps to know how the technology works and what the limitations are because we don't have positronic brains around the corner or the robots to match—yet.

Machine-Learning Basics

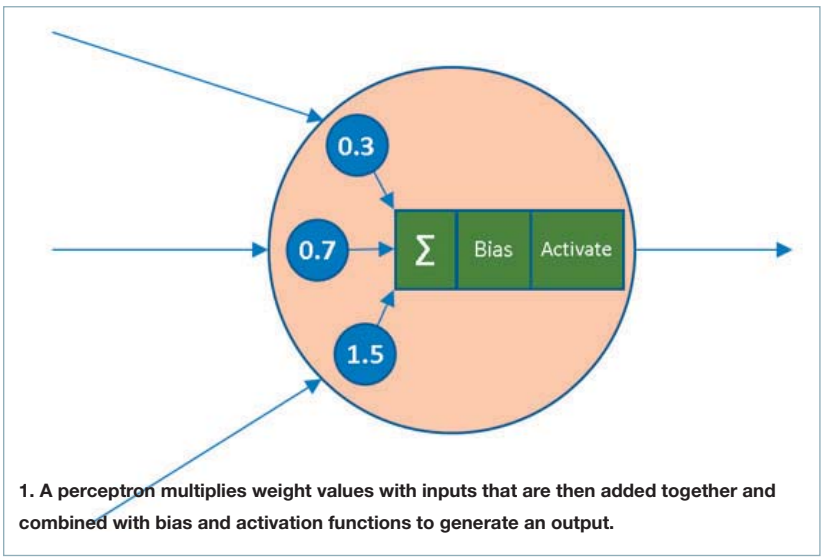
Two of the main ML technologies include rule-based expert systems that are sort of state machines on steroids, and neural networks.

Expert systems are still in use and a viable solution for many problems. They're normally created using explicit rules to build behavior-based systems that respond to input. Self-trained systems like this haven't worked well, but expert systems can be very fast, accurate, and efficient when properly programmed. Unfortunately, translation of expertise into an expert system is often time-consuming.

On the other hand, neural networks (NNs) are modeled after the brain, although at a logical level. The approach has been around for decades. However, new hardware made this approach more practical, allowing for the implementation of complex AI/ML models.

A perceptron is the model of a biological neuron *(Fig. 1)*. In general terms, input values are combined with weights, summed with a bias and activation function included in the mix to generate an output. This is kind of like discussing logic circuits, op amps, and transistors. We're doing a bit of hand waving here, but it's enough to provide the basics for our discussion.

The next step is to combine perceptrons *(Fig. 2 on page 12)*. Typically, multiple layers are involved in a model that needs to be designed based on the number of inputs, outputs, and functionality of the system. A model is the combined architecture along with the weights used within the system. These weight values



1. A perceptron multiplies weight values with inputs that are then added together and combined with bias and activation functions to generate an output.

A perceptron is the model of a biological neuron *(Fig. 1)*. In general terms, input values are combined with weights, summed with a bias and activation function included in the mix to generate an output. This is kind of like discussing logic circuits, op amps, and transistors.

are usually obtained by training a model. The number of layers can be large, hence the term deep neural networks, or DNNs, that really describes the general magnitude of the models rather than a specific approach, which we will get to later.

The outputs are probabilities. A system will often have thresholds for acceptable identifications depending on the purpose of the model. The level of detail in the analysis of an image, for example, can range from an object to an animal to a cat to a Persian cat.

Dealing with images can greatly increase the complexity of a model. Even a small, 320- × 240-pixel image translates to 76,800 inputs. If color instead of grayscale is used as inputs, then a red-green-blue (RGB) encoding bumps this by a factor of three.

On the other hand, working with fewer inputs enables simpler models to be implemented, often in software on a microcontroller. For example, a motor-control application may have half-a-dozen inputs to support a preventive-maintenance model.

Neural Networks Aplenty

There are many different types of NNs, including artificial neural networks (ANNs) and spiking neural networks (SNNs). ANNs use a parallel approach to presenting the inputs to the network with outputs available after all of the data flows through the network.

SNNs are more akin to biological neurons—they are asynchronous working off time-related blips of data *(Fig. 3)*. An SNN neuron emits a spike when a membrane threshold is reached. An SNN can be implemented in many ways, but the most popular is a leaky integrate-and-fire approach that mimics its biological cousins.

FREE ARCHIVED WEBINAR

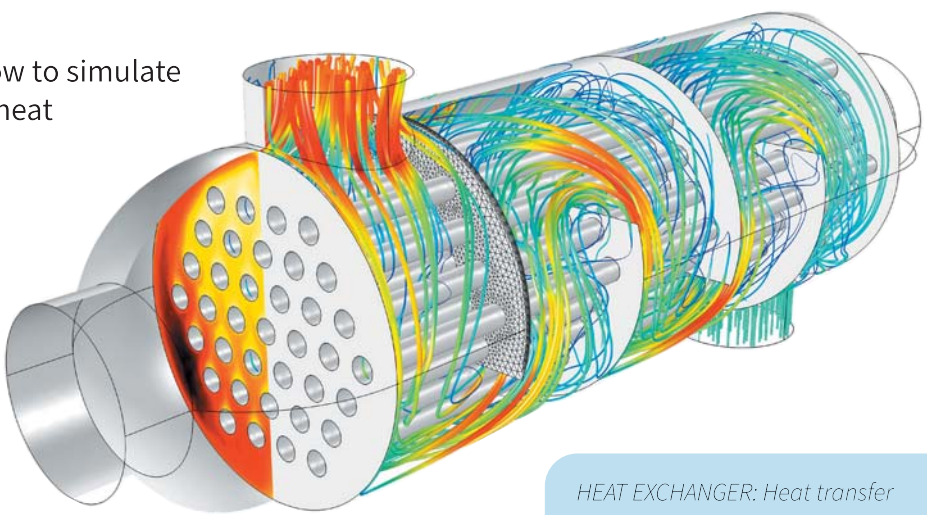
Registration URL: https://www.electronicdesign.com/resources/webcasts/webinar/21160324/simulation-of-heat-exchangers

Simulation of Heat Exchangers

Heat exchangers are used to transfer thermal energy, usually between two fluids. Several types of heat exchangers are available, and each type of heat exchanger can be sized to suit the desired application.

In this webinar, you will see how to simulate fluid flow and heat transfer in heat exchangers.

We will study how simulation can be used to evaluate and improve several properties, such as the overall heat transfer coefficient, pressure losses, total power exchange, compactness, and more.



HEAT EXCHANGER: Heat transfer between air (shell side) and water (tube side) through steel walls in a shell-and-tube heat exchanger. Nonisothermal flow simulation results reveal flow velocity, temperature distribution, and pressure within the fluids.

You will also get a live demonstration of the model configuration using COMSOL Multiphysics®. Finally, specialized simulation applications will be used to compare the performance of heat exchangers as a function of, among other things, geometrical properties and the choice of a counterflow or parallel flow configuration.

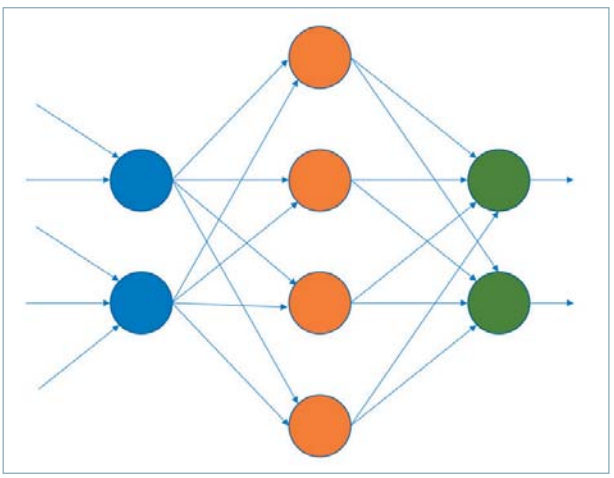
SPEAKER: Nicolas Huc, Product Manager, Heat Transfer module, COMSOL

Nicolas Huc joined COMSOL France in 2004 and is currently the head of their development team. He is also the manager of the Heat Transfer Module. Nicolas studied engineering at ENSIMAG before receiving his PhD in living system modeling from Joseph Fourier University.

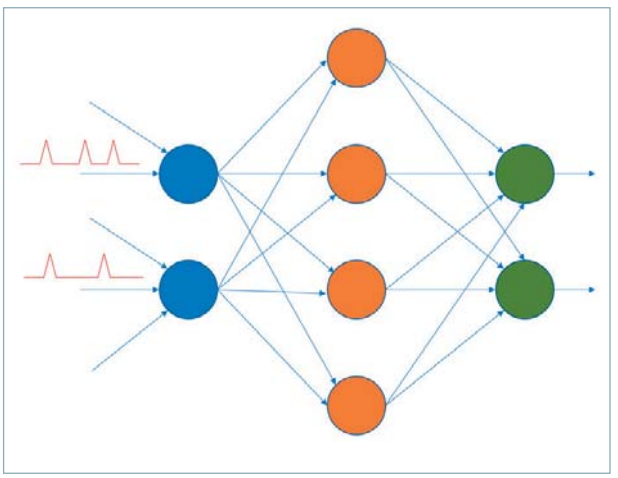
Hosted with

ElectronicDesign

COMSOL



2. Perceptrons are combined in multiple layers to form machine-learning models. The middle layers are often referred to as hidden layers, and there may be multiple hidden layers.

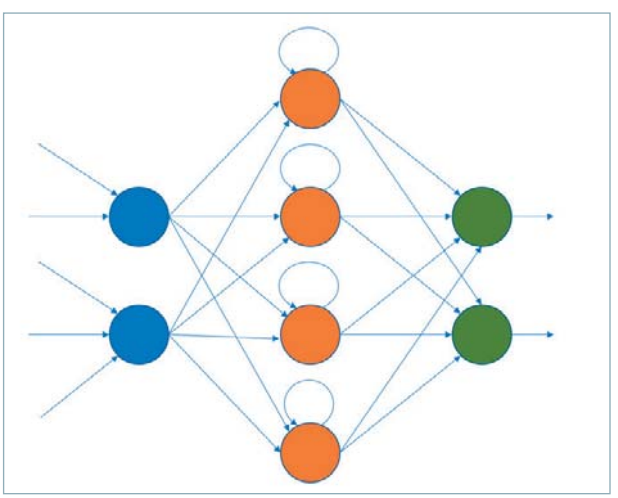


3. SNNs are similar to biological neurons because input data is a series of time-related blips of data.

SNNs tend to have less overhead than other neural networks, but their accuracy is usually lower. Keep in mind that all of these networks are dealing with thresholds and probabilities and 100% accuracy is only a goal. Also, keep in mind that SNNs can often be trained in the field, while training for other neural nets is typically done in the cloud where additional data is accumulated. SNNs are able to do this in part because of their lower implementation overhead and due to the technology itself.

A number of other neural-network types are out there, and many models can fit under these broad categories. They can perform functions such as identification and classification. Let's take a closer look at three major categories:

- Convolutional networks (CNNs)
- Recurrent neural networks (RNNs)
- Generative adversarial networks (GANs)



4. Recurrent neural networks (RNNs) provide each perceptron with a feedback loop.

A CNN is a DNN that's typically used for image analysis. It can be applied in facial-recognition systems, parsing documents, and image segmentation. CNNs also are a type of space invariant artificial neural network (SIANN).

A CNN model essentially implements filters via the training process. The bias and activation operations include a Frobenius inner dot product. The model includes convolutional and pooling layers. The number of inputs and outputs of a convolutional layer are typically the same. The pooling layers reduce the number of outputs. This is often accomplished by getting an output, which involves taking the maximum input value or average of a group of inputs.

An RNN's perceptrons take their output as the input to the next set of data *(Fig. 4)*. A data-flow implementation would typically include latched inputs so that the output could be used in the subsequent calculation. Oftentimes, RNNs are used to analyze audio streams, text, and time series information.

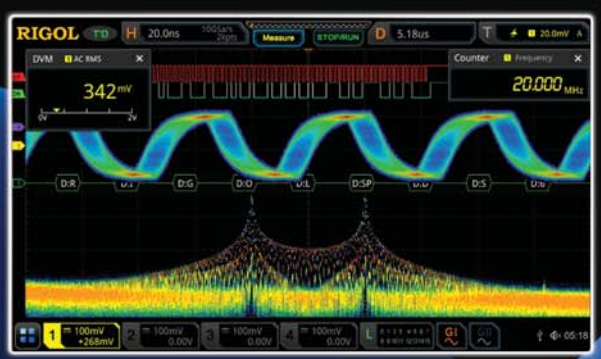
The GAN is a neural network designed to generate an output that's similar to the input. For example, a GAN model could change a cat into a dog that would have the same pose, etc. This type of model is able to unpaint pictures or upscale video.

A GAN consists of a generator network and a discriminative network. The former creates candidates based on the inputs and the latter evaluates the new data. Each must be trained with the discriminative network being trained with inputs, much like how a CNN would be trained. The generator is typically a deconvolutional neural network that's trained to increase the error rate of the discriminative network.

World's Most Powerful & Affordable Bench Oscilloscopes

UltraVision II SCOPES

Category Leading Performance for Complex Embedded Debug



- UVII Scopes from 70MHz to 2GHz
- All Feature High Sample Rates, Deep Memory & Next Gen UI
- 7 in 1 Instrument Functions
- Advanced Analysis Tools
- Starting at $909

RIGOL

Check Out Our Debugging IoT Electronics Video Series

www.RigolNa.com/Scopes

Multicore vs. Data-Flow Acceleration

Neural-network models are like applications. They're specifications that can be implemented in a number of ways, from a software application to a hardware implementation. Hardware implementations are typically faster and use less power, thus providing more performance such as the ability to analyze video streams in real-time.

A very, very large number of calculations is the reason that hardware acceleration is often required in AI applications, but this is related to the input size and model complexity. Video applications typically benefit from hardware support.

Hardware-acceleration approaches can be divided into data-flow or application-specific support, targeted multicore, and instruction set augmentation. Most ML/AI chips go with data-flow or targeted-multicore implementations. The latter is often a DSP or single-instruction, multiple-data (SIMD) architecture that's customized for AI/ML models. An example of instruction set augmentation is Arm's Cortex-M55, which adds instructions that improve support for AI/ML-related numeric encodings and computations such as matrix manipulation.

Hardware-acceleration approaches can be divided into data-flow or application-specific support, targeted multicore, and instruction set augmentation. Most ML/AI chips go with data-flow or targeted-multicore implementations.

AI/ML models can be implemented using double-precision floating point, but that tends to be very inefficient for various reasons. Using more compact formats like 8- and 16-bit floating point, and 8-bit fixed point, significantly reduces storage requirements, calculation hardware, and calculation time, thereby lowering power requirements while improving performance.

Another aspect deals with sparse networks. It turns out that in many cases, weights are often zero or very close to zero. This enables those calculations and the data movement to be eliminated by optimization. Much of it is hidden behind model compilers that accept models from standard frameworks like TensorFlow and Caffe2.

Another consideration with respect to hardware involves multiple models and splitting models across different types of hardware. Many times, multiple models are used to provide more functionality based on the same input or to address different aspects of the overall application.

For example, an electric car may have an advanced driver-assistance system (ADAS) or fully autonomous-driving system that uses object recognition and other models to handle preventative maintenance for the electric motors. Targeted or sophisticated system-on-chip (SoC) solutions are often designed to handle this mixture.

The splitting of models makes sense because the hardware support for different parts of the model can vary. Multicore microcontrollers like Eta Compute's ECM3532 includes a Cortex-M4 and NXP CoolFlux DSP. The latter is augmented to handle AI/ML models, but the Cortex-M4 can help.

Making Decisions for Applications

Choosing models and hardware platforms to support them can be complex. They could probably use an AI model if someone would make it. On the plus side, this type of support can be approached in different ways if your application would benefit from AI/ML support.

The first way is easy: Use a product with AI support that requires minimal or no training. Many robotic sensors targeting the educational market fall into this category. They are essentially smart sensors that provide feedback to a host, such as "red obstacle 15 cm ahead."

The majority of uses fall into the next category: Utilizing predefined and often pretrained models on hardware of your choice. This may require choosing a model, and usually the available models will be based on the hardware choice or vice versa. This is the advantage of using platforms like TensorFlow or Caffe, since hardware vendors have chosen to support these. Likewise, performance numbers are often available, so you can gauge what platform may be suitable for your application.

If a model is available and you have the data to train it, then the problem often reverts to a standard engineering challenge of balancing and optimizing cost, performance, power, and capacity. You may have much more work to do if the models that exist don't quite match your needs.

Creating a new model is only for those with big bucks, lots of time, or lots of expertise. This is where the experts come in—even companies and organizations are using AI to create models. You could also wait until someone develops a model that you can use, although working with an unsupported, open-source model may not be the best choice for a commercial product. Like many open-source software projects, you may have to support yourself.

The challenge these days is that the technology and availability of AI/ML solutions is changing so quickly. Compiler improvements can sometimes double performance, making a lower-end platform capable of handling your application when it couldn't do so with the earlier version. Likewise, many new chips with AI enhancements or dedicated chips are becoming available, and new processor designs are regularly including instruction set extensions and data formats amenable to AI/ML models. ED

Microwave Synthesizer Offers Multioctave Frequency Coverage and Excellent Phase Noise Performance

Marty Richardson, Senior Applications Engineer

Introduction

System frequency and modulation rate requirements continue to escalate with the need for more bandwidth and higher data rates. Low power consumption has become critical as applications once relegated to military and defense enter the commercial sector. These demands come with the caveat that there be no sacrifice of electrical performance or functionality. In order to meet these requirements, including an improved signal-to-noise ratio (SNR), bit error rate (BER), and the high quality of service (QoS) that users are accustomed to, the phase noise of the local oscillator (LO) must improve as well.

The newly released ADF5610 is an integrated phase-locked loop (PLL) and voltage controlled oscillator (VCO) that highlights Analog Devices' efforts to provide a solution that addresses each of these concerns and more.

Frequency Coverage

A total of eight octaves are covered by the ADF5610 with the VCO fundamental frequency ranging from 3.65 GHz to 7.3 GHz, which is fed back to the PLL to minimize phase noise. A single-ended output (RFOUT) doubles the fundamental frequency range to provide 7.3 GHz to 14.6 GHz while the differential output simultaneously allows the full operating range of 57 MHz to 14.6 GHz through the use of divide by 1/2/4/8/16/32/64 and 128 settings.

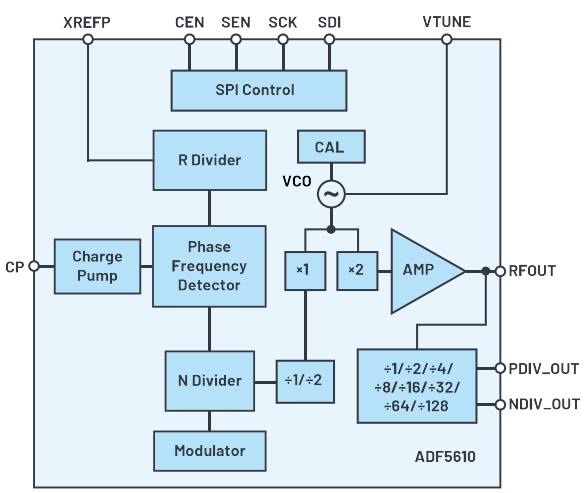


Figure 1. Block diagram of the ADF5610.

The architecture of the ADF5610 VCO allows wideband synthesizer performance while retaining industry-leading phase noise performance with a nominal open-loop phase noise at 10 GHz of −114 dBc/Hz at a 100 kHz offset. An internal state machine allows frequency settling times of under 40 µs using just a passive loop filter; no need for additional circuitry or lookup tables (LUTs) unless faster settling times are required.

Leading PLL Performance for Converter Clocks

While the phased-locked loop (PLL) inside of the ADF5610 boasts a modest figure of merit (FOM) of −229 dBc/Hz (−232 dBc/Hz high current mode), when combined with exceptional 1/f noise (−129 dBc/Hz) and state-of-the-art VCO phase noise, rms jitter numbers less than 38 fs (1 kHz to 100 MHz integration limit) are possible. This makes the ADF5610 suitable for use in the most demanding converter clock applications. Loop filter resistor values should be kept at a minimum to reduce their thermal noise and a high frequency (100 MHz). An ultralow noise reference source is essential in order to achieve this level of performance.

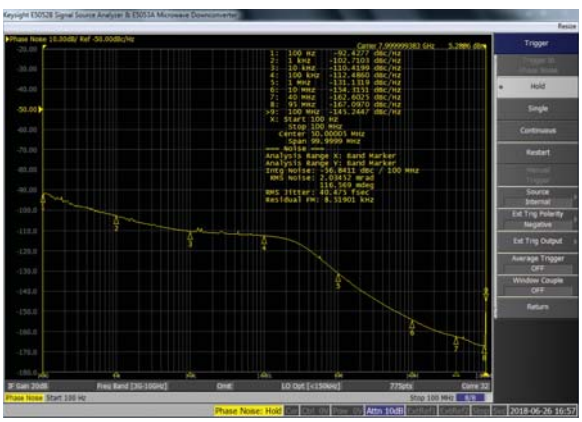


Figure 2. RMS jitter: 8.0 GHz.

Figure 3. RMS jitter: 14.4 GHz.

Communications and Instrumentation LOs

In addition to its broad frequency coverage, industry-leading phase noise, and exceptionally fast lock times, the ADF5610 has additional features that make it attractive for wireless and instrumentation applications where it will most often serve as the local oscillator.

24 bits of fractional resolution are modest, but when paired with the ADF5610's exact frequency mode functionality, frequency generation with 0 Hz error is possible. Using the ADF5610 as the local oscillator allows the active mixer to be driven directly from the RFOUT port due to the nominal 5 dBm of output power, eliminating the need for additional amplification and saving valuable board space. The output power on the differential divider (PDIVOUT/NDIVOUT) is nominally 2 dBm when used single ended, but it can be combined through a low loss balun or hybrid coupler for narrow-band applications to achieve an additional dB or two of output power.

Low power dissipation is essential today and the ADF5610 does its part by sipping less than 700 mW (low current mode with the output divider disabled) to just over a watt under worst-case conditions (high performance mode with the output divider set to divide-by-128). Even in low current mode, the ADF5610's phase noise performance leads in its class, increasing by just 2 dBc/Hz.

The ADF5610 has good spurious performance, with PFD spurious as low as −105 dBc, and in-band unfiltered integer boundary spurs nominally at −45 dBc.

Small Size

The ADF5610 PLL/VCO is available in a 7 mm × 7 mm, 48-lead land grid array (LGA) package. Minimal additional decoupling is required, meaning that exceptional performance exists in a small footprint solution. To achieve the best performance, the use of high quality low dropout (LDO) regulators such as the ADM7150, LT3045/LT3042, or HMC1060 are recommended. The VCO requires a 5 V supply, while the remaining circuitry is powered from a 3.3 V rail. The ADF5610 can be simulated in ADIsimPLL™ to assist the user in designing the appropriate external component circuitry required to implement a complete PLL synthesizer.

Conclusion

Industry-leading frequency coverage and phase noise performance coupled with high output power and low power dissipation in a small form factor combine on the ADF5610 to address the stringent demands of new communication and instrumentation systems.

PONTUS LIDMAN | Principal Software & Security Architect, Synaptics Inc.

AI Can Tame Conferencing's HUNGER FOR BANDWIDTH

Video conferencing requires high-bandwidth, low-latency connections, but can AI help codecs deliver this connectivity more efficiently?

Andrey Popov | Dreamstime.com

Video conferencing for virtual meetings, distance learning, or socializing has exploded with the onset of the coronavirus pandemic. Some experts suggest that even after the virus recedes, our reliance on virtual gatherings will remain part of our new normality. If so, the huge bandwidth hunger that ubiquitous video conferencing imposes on the internet—from the core out to the thinnest branches—is here to stay.

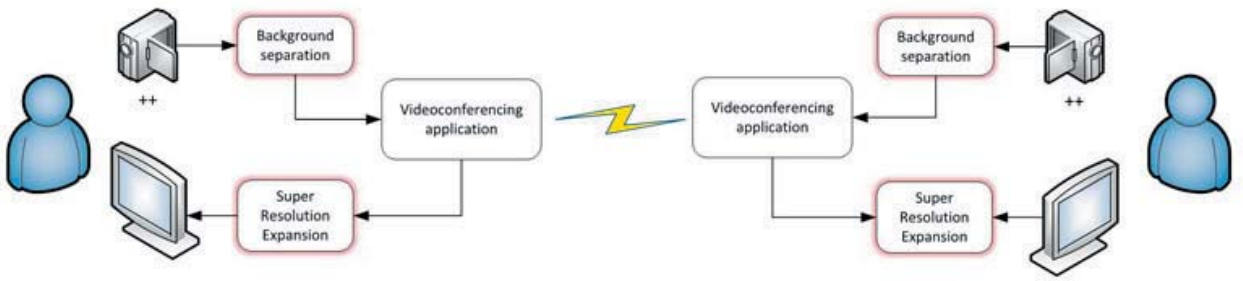
Even using modern video codecs, a video conference can be demanding on bandwidth: 1 to 2 Mb/s per participant just to keep those thumbnail images on the screen. And there's growing evidence that with experience, users become more critical of image quality, longing to see fine details of facial expressions, gestures, and posture that carry so much information in an in-person meeting. This trend limits the ability of apps to use higher compression ratios to reduce bandwidth needs. The fine detail the compression algorithm throws out contains just the cues a skilled negotiator needs most.

AI to the Rescue?

Help with this dilemma can come from a surprising source: artificial intelligence (AI), or more specifically, a branch of AI called deep-learning networks. Today, a machine-learning application called Super-Resolution image expansion—already being explored for delivering high-quality video to 4K UHD television screens—can cooperate with existing video-conferencing apps to significantly reduce their required bit rates.

Here is how it works. Each video-conferencing device in the conference that intends to display high-resolution images would maintain two machine-learning "inference models." Each model is a block of code and data that's been trained beforehand, through an exhaustive process in a data center, to perform a particular function. One of the models processes the video from a user's HD camera before sending it the conferencing app, and the other processes video coming out of the conferencing app before displaying it *(Fig. 1)*.

The first model takes in video from the camera frame by frame and isolates the image of the user from the background, reducing the number of pixels that later stages will have to deal with. This simplified video stream then flows on to a normal video-conferencing app, where it's sampled down to a lower resolution like 480P and then compressed using an industry-standard algorithm such as H.264. The compressed video is subsequently exported. Except for that first step



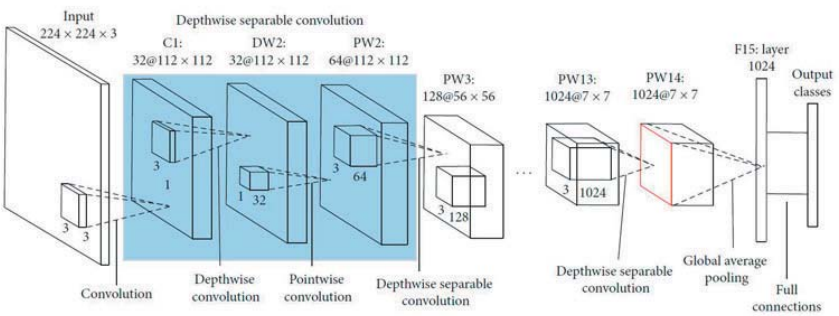
1. AI-enabled Super Resolution can reduce the bandwidth of a video stream in video-conference applications. It uses two machine-learning inference models: One processes the video from a user's HD camera before sending it the conferencing app, and the other processes video coming out of the conferencing app before displaying it. *(Source: Synaptics)*

of isolating the user's image, everything so far has gone as it would in any other video-conferencing scenario.

On the receiving end, the video-conferencing app receives the compressed bit stream, decompresses it into low-resolution video, and sends the video stream to the display subsystem. If the image is to be displayed as a thumbnail—in a multi-party conference, for instance—the video will be displayed directly. The image quality will be good enough for the small size on screen. But if the image is to be larger, the decompressed video is diverted into the second deep-learning model, the Super-Resolution expander.

The Super-Resolution model has been trained using myriad images with different faces, lightings, and poses to selectively add back the information that's been lost when going to low resolution and in compression. The result is a high-quality image of the original user that closely resembles her image in the original HD camera video.

Note that this isn't the same as decompression. The AI model is adding features into the low-res image that aren't there, but that human subjects would expect to see, completing the high-resolution picture frame by frame in real-time.



2. A typical deep-neural-network structure, illustrating the regular structure of computations that makes it the problem domain suitable for dedicated hardware acceleration. *(Source: A Novel Image Classification Approach via Dense MobileNet Models Wei Wang et al)*

What it Takes

Deep-learning networks, like most kinds of AI, are notorious for their huge computing appetites. Fortunately, most of the computing goes into training the models in the first place—a task done in a data center before the model is shipped to users. Once a deep-learning model is trained, it's just a reasonably compact block of code and some data files. Both the user-extraction model and the Super-Resolution expander model can be run comfortably on a GPU or a reasonably fast notebook computer.

But as video conferencing becomes more common, the need will grow to use much more modest devices, such as dedicated conferencing appliances, tablets, smart TVs, or set-top boxes. Work on special deep-learning hardware accelerators—chips that vastly increase the number of computations done at the same time while sharply reducing power consumption—has brought these deep-learning models within the range of low-cost, low-power devices.

One example of this work is the Synaptics VS680 system-on-chip (SoC). This multimedia processor SoC combines Arm CPU cores, a GPU, video- and audio-processing subsystems, extensive provisions for security and privacy, and a deep-learning accelerator called the Neural Processing Unit. That latter block can run both the user-extraction and the Super-Resolution expander models at the same time at full video frame rates.

The result is a single chip that substantially reduces bandwidth requirements for video conferencing while maintaining high-quality images, at a price and power consumption suitable for even low-cost displays, streamers, and set-top boxes. And the service is compatible with existing video-conferencing apps.

As use of video conferencing continues to rise, and as it's utilized by more people in areas marginally served by broadband access—often people without expensive notebook computers—the ability to significantly reduce bandwidth requirements without harming image quality, and to do so on inexpensive devices, will become increasingly important.

The Many Faces of Deep Learning

A deep-learning network model, once designed and trained, can only do what it was trained to do: identify flowers, say, or in our case, pick a person out of her surroundings in a video frame. But the underlying software and hardware that execute the trained model can often handle a wide variety of different kinds of machine-learning models, trained in different ways to perform very different tasks.

For example, the firmware and the Neural Processing Unit hardware in the Synaptics VS680 can perform a wide variety of jobs in a multimedia system. These include recognizing objects, sensing the user's location and surroundings, or detecting objectionable content or malware in incoming data streams.

The calculations performed by a deep neural network are massive, but fundamentally simple ones. *Figure 2* shows the structure of one of the most popular neural networks: MobileNet. It contains a series of convolutions that require massive amounts of multiply-and-accumulate operations.

This makes the problem highly amenable to optimization by custom hardware implementations. MobileNet is a typical network that can be used for multiple image-processing applications. Networks for other tasks are built using similar primitives; that's why the dedicated Neural Processing Unit in the Synaptics VS680 can deliver high performance to any deep-learning AI task in video, audio, or analytics applications, to name a few.

One recent proposal by a GPU vendor illustrates the not necessarily desirable lengths to which this flexibility can be taken. There's a category of deep-learning networks called generative adversarial networks, or GANs. They are used, most notoriously, to create deep-fake videos.

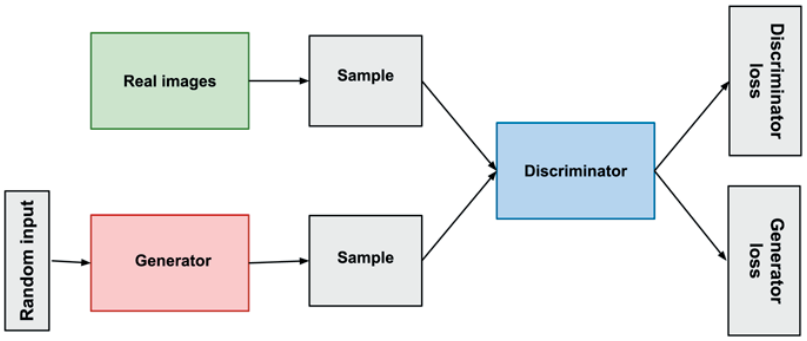
Given a detailed photo of a person and a set of parameters specifying the location and orientation of key facial features and body parts, a well-trained GAN will generate a photorealistic image of the person. That image could be in surroundings not present in the original photo, and gestures and expressions could be different from those in the original. String a sequence of such generated images together, and you have a video of the person doing or saying things they never did or said, in a place they may have never been.

Training a GAN involves two neural networks: a generator, and a discriminator *(Fig. 3)*. The generator will generate random images, which the discriminator will try to tell from real images. The discrepancy between the generated image and the real image gets fed back to the generator during training. Eventually the generator will be able to generate images that the discriminator can't distinguish

from the real ones. The discriminator network is performing image classification and could be based on MobileNet or another network.

Despite the unfortunate obvious use of this technology, it can also be utilized to slash the bandwidth consumed in video conferencing. By using a GAN to generate an image of the user at the receiving end of the connection, you need only send an initial static image and then a stream of data specifying the location and shape of the key features. This stream of data can be significantly smaller than the original high-res compressed video stream.

There are practical issues. For one, because the technique sends a stream of abstract data rather than a stream of standard compressed video, it's not compatible with existing video-conferencing apps. For another, the security risks of operating a video-conferencing network full of GANs, any of which could be hijacked to create deep-fake images instead of reconstructed ones, would require careful consideration. But the idea illustrates how, once a video-conferencing device is able to execute deep-learning models, imagination is the only limit on the functions it can perform.



3. Two neural networks are used to train the system to fill in missing detail from low-bandwidth representations: A generator and a discriminator work together in an adversarial manner to learn how to produce believable missing detail. *(Source: Google)*

Deep-learning inference acceleration hardware like the VS680's Neural Processing Unit can deploy AI to provide this bandwidth reduction. Such a solution is able to work with existing conferencing services and fit within the cost and power budgets of inexpensive consumer devices. Distance learning and working from home needn't force us to choose either that users learn to accept terrible image quality, or service providers make deeper investments in network bandwidth. With intelligence, we can have our cake and eat it, too. ED

WILLIAM WONG | Senior Content Director

How to Quickly Create SMARC Module Solutions

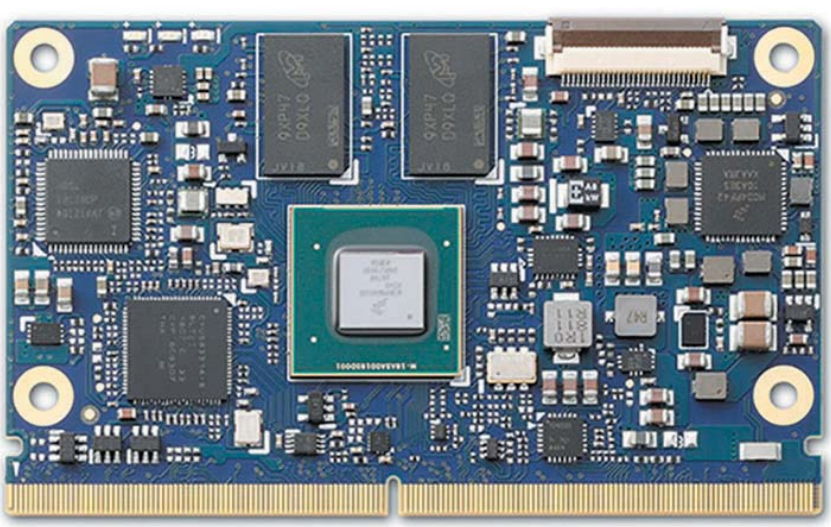
Adlink put the spotlight on SMARC and COM Express at Embedded World 2021.

One of the products given special attention by Adlink Technology at Embedded World was its I-Pi SMARC IMX8M Plus Starter Kit. The kits are complete systems that include a SMARC module like the LEC-iMX8M *(Fig. 1)*, along with a carrier board and software.

SMARC is a module standard from the Standardization Group for Embedded Technologies (SGeT). There are two modules sizes: 82 × 50 mm and 82 × 80 mm. Each has a 314-pin edge connection. The SMARC 2.1 specification is the latest with support for PCI Express and multiple display options like LVDS and HDMI. The standard defines interfaces for typical module interfaces such as Gigabit Ethernet, serial interfaces, and USB.

The LEC-iMX8M SMARC module family is based on the NXP i.MX8M SoC family that's built on Arm's 64-bit Cortex-A53 and 32-bit Cortex-M4. Multiple configurations exist of each type, which is an advantage to using SMARC since you simply replace the module in a system with another.

The modules usually include support for one or more MIPI-CSI-2 camera inputs, 4K UltraHD HDMI 2.0a, and LVDS displays. USB 2.0/3.0 and OTG support is in the mix as well. The PCI Express interface allows access to peripherals on the required carrier board and can incorporate devices like an M.2 PCI Express NVMe memory stick.



1. Adlink's LEC-iMX8 SMARC module sports a quad-core Cortex-A53 and an optional neural processing unit.

Among other useful features are cryptographic co-processor support and some modules, like the i.MX8M Plus, have a neural processing unit (NPU). The NPU can deliver up to 2.3 TOPS of performance for running machine-learning inference models.

As noted, SMARC modules require a carrier board. The Adlink kits come with one that exposes all of the peripheral interfaces.

Designing a Carrier Board

I've written a good deal about Gumstix Geppetto. It's a neat online design tool that creates a complete board design without requiring extensive training and expertise. The pick-and-place interface hides all of the printed-circuit-board (PCB) routing and guarantees a good design by requiring all necessary connections to be specified by the designer *(Fig. 2 on page 32)*. This is done by providing prompts and selections while showing the design status using color coding.

For example, I started a design by dragging a SMARC connector for an i.MX8M module to the board layout section; it had the color red. The minimum connections are power, which required grabbing a power connector and a pair of voltage

(Continued on page 32)

RAFAEL TAUBINGER | Technical Marketing Specialist, IAR Systems

Defeat Defects with SAFETY CODING TECHNIQUES

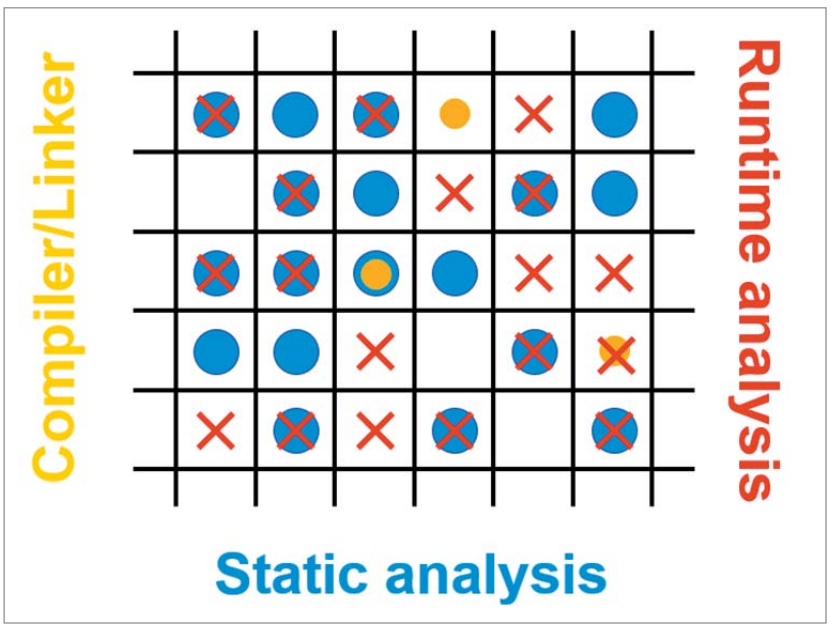
The move from mechanical processes to high-level programming languages has ushered in numerous unknown side effects leading to defects. Thus the need for coding standards to help future-proof your code and ensure ease of reuse.

The modernization of programming languages and the importance of better coding techniques is directly related to the evolution from mechanical computers to modern software-development processes. We moved from a highly specialized, mostly mathematical, notation to high-level programming language close to human syntax.[1] While this was made possible by compiler technologies, it opened the door to defects.

High-level programming languages like C and C++ contain an enormous number of undefined behaviors that various compilers may interpret slightly differently. This can cause unknown or unwanted side effects that will translate into defects.

Detecting and fixing defects can take up to 80% of the development time, depending on the maturity of the development organization. Obviously, code quality is a huge concern. Why not avoid the defects and significantly reduce debugging time?

On a side note, the terms "bugs" and "debugging" in software originally came from Harvard University's mechanical computers. A moth became stuck in a relay and the event was documented as the first system defect or "bug" in computer history.



1. This matrix best represents the complete defect coverage when combining the different tools.

The terms "bugs" and "debugging" in software originally came from Harvard University's mechanical computers. A moth became stuck in a relay and the event was documented as the first system defect or "bug" in computer history.

Our Definition of Insanity: Repeating the Same Mistakes Again and Again

A number of respectable institutions like NASA, Bell Labs, and MITRE conducted several surveys and studies that led to the same conclusion: Developers in web, apps, desktop, or embedded tend to inject, accidentally, the same kind of mistakes into their source code repeatedly.

Examples of these common errors include allocations without deallocations in C++ code (or even in C code) and functions used without prototyping, so you don't get rigorous type-checking at compile time. In addition to identifying these commonalities, the research identified lists of best or recommended programming practices that pinpoint and help prevent risky and bad coding behavior.

Some best practices have been defined as well-known standards, e.g., MISRA C and CERT C. These are especially in critical in ensuring code safety and security in applications used in industries like transportation and healthcare.

Functional safety standards such as IEC 61508,[2] EN 50128,[3] and ISO 26262[4] recommend (or highly recommend, depending on the safety integrity level (SIL) or Automotive Safety Integrity Level (ASIL)) the use of static- and runtime-analysis tools to be compliant with the standards. Defects in safety-critical systems could lead to severe consequences such as loss of lives or environmental damage.

Focus on Reliability

Safety coding techniques represent the combination of code quality, code safety, and code security. Code safety focuses on the reliability of the software, while code security is all about preventing unwanted activity and keeping the system secure during an attack. Both rely heavily on the code quality that's the foundation of every solid application.

While safety coding techniques and standards drive software safety to ensure the needed reliability, it's crucial to also enhance readability and maintainability of the source code. More efficient and readable code means future-proofing the source code with less defects and making the reuse of code possible.

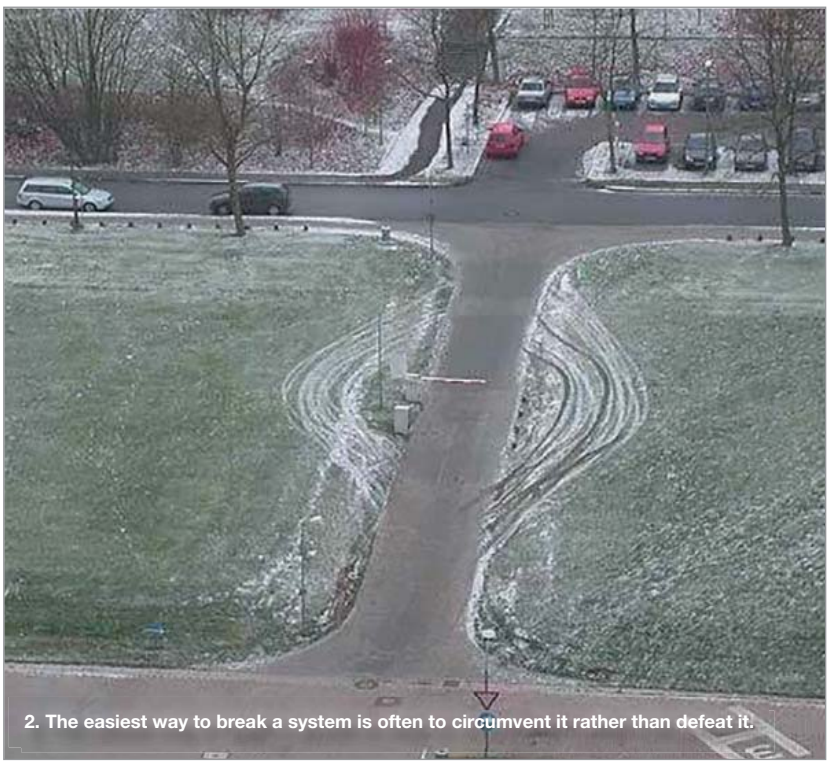
MISRA C is one of the most well-established software-development standards designed to aid in avoiding common pitfalls and vulnerabilities. However, additional guidelines like CWE (Common

Weakness Enumeration) and the CERT C coding standard are highly recommended for any embedded application.

The MISRA C Standard

MISRA C was developed by the Motor Industry Software Reliability Association. Its aims are to facilitate code safety, portability, and reliability in the context of embedded systems, specifically those systems programmed in ISO C.

The first edition of the MISRA C standard, "Guidelines for the use of the C language in vehicle based software," was produced in 1998, and is officially known as MISRA C:1998. It was updated in 2004 and again in 2012 to add more rules. There's also the MISRA C++ 2008 standard based on C++ 2003. Lately, MISRA C:2012, Amendment 1 adds 14 additional rules with a focus on security concerns highlighted by the ISO C Secure Guidelines. Several of these rules address specific issues pertaining to the



2. The easiest way to break a system is often to circumvent it rather than defeat it.

use of untrustworthy data, a well-known security vulnerability in many embedded applications.

MISRA helps you find issues before you check code into a formal build, so finding a bug this way makes it like the defect never happened. MISRA rules are—again—designed with safety and reliability in mind, but they also make code more portable to other tools and architectures.

CWE and CERT C/C++

CWE is a community-developed dictionary of software weakness types. CWE provides a unified, measurable set of software weaknesses to better understand and manage them as well as enable efficient software security tools and services that can find them.

The CERT C/C++ Secure Coding Standards are standards published by the Computer Emergency Response Team (CERT). They provide rules and recommendations for secure coding in the C/C++ programming languages.

Enforcing Safety Coding Techniques

As a general recommendation, every embedded application minimally should follow the CWE and CERT C/C++ standards. MISRA C is mandatory for safety-critical systems.

Following the same concept, during runtime you can still be susceptible to arithmetic issues, buffer overflow, bounds issues, heap integrity, and memory leaks. Such errors can be detected by inserting specific instrumentation code or asserts at all places where a potential error can happen. However, adding instructions manually to check the condition and somehow report the issue at runtime is a very time-consuming task.

Applying all guidelines and standards means that you need to be compliant with almost 700 rules and requirements in addition to instrumenting your code. So, how can you enforce the safety coding techniques and keep up with all of the rules?

Use Automated Tools

The best way to enforce software quality, safety, and security is to use automated tools. This can be achieved with the use of a high-quality compiler and linker that are preferably functional-safety certified and combined with automated static analysis and runtime analysis.

The compiler and linker should support a modern programming language like the latest C (ISO/IEC 9899:2018) and C++ (ISO/IEC 14882, known as C++ with the latest C++17 revision). Therefore, they can generate warnings for suspicious situations or for syntax weaknesses, e.g., volatile memory access whose order of

evaluation could affect the logic of the application.

Warnings are your first-pass static-analysis check and should never be ignored, particularly in a functional-safety setting. The best recommendation is to turn the warnings into errors by changing the compiler settings to treat all warnings as errors. This will force the developers to fix all ambiguities in the code, since all issues will be handled as genuine problems.

Static-analysis tools help you locate the most common sources of defects in your code. However, they also help find problems that developers tend to not think or worry about when trying to write their code, especially when they're just putting up scaffold code to get something working.

Such tools really aid in developing better code because they enforce the coding standards. In addition, dynamic or runtime-analysis tools catch and trigger defects that only pop up during runtime. A runtime-analysis tool can find real and potential errors in the code while executing the program in a software debugger.

3. Coding standards help to future-proof your code and ensure ease of reuse.

So, when you look at all of the defects that could be in your system, static analysis is good at finding some defects and runtime analysis is good at pinpointing others. Sometimes there's overlap, but in other instances a defect can only be detected in one domain or the other. To get the best possible code analysis, you need to use both in conjunction with another and integrate them with your top-notch build tools. The matrix *(Fig. 1)* best represents the complete defect coverage when combining the different tools.

When you look at all of the defects that could be in your system, static analysis is good at finding some defects and runtime analysis is good at pinpointing others. Sometimes there's overlap, but in other instances a defect can only be detected in one domain or the other.

Catching Loopholes

This effect can best be explained in *Figure 2,* taken from University Bielefeld in Germany.[6] This photo was taken by an anonymous contributor and circulated widely in 2005.

The easiest way to break a system is often to circumvent it rather than defeat it. This is mostly the case with software vulnerabilities connected to insecure coding practices. The image in *Figure 2* is a great analogy for this—the gate is in place per the recommendations and is functioning appropriately according to the specifications. However, the security measure was easily bypassed and until a runtime-analysis tool was in place (in this case, snow), it was probably difficult to spot the flaw in the security system. Automated runtime analysis that scans your code for potential loopholes is a great way to detect these types of issues.

In this case, the safety vulnerability was fixed in hardware, namely the installation of concrete bollards according to Google Street View in 2020 *(Fig. 3).*

Coding standards help to future-proof your code and ensure ease of reuse. This means the code's quality affects the code's reusability and this is what mature organizations have in their culture when developing new products. It's a virtuous cycle to enforce the safety coding techniques and it reaffirms our premise: Everything simply starts with code quality. ed

REFERENCES

1. Hopper (1978) p. 16.
2. https://www.iec.ch/functionalsafety/standards/page2.htm
3. https://standards.globalspec.com/std/14256883/EN%2050128
4. https://www.iso.org/standard/43464.html
5. https://wiki.sei.cmu.edu/confluence/display/seccode/Top+10+Secure+Coding+Practices?focusedCommentId=88044413
6. Google Map view

VINCENT KORSTANJE | CEO, Kigen

What's the Difference Between **SIM, eSIM, and iSIM?**

Including a SIM to help achieve cellular connectivity amongst the rising tide of connected IoT devices presents opportunities and challenges for designers. Kigen's Vincent Korstanje looks at a fresh approach to device connectivity, cost, and security.

The typical smart home in 2020 had, on average, 50 connected IoT devices—and that's contributing to the 35 billion devices we as consumers are expected to own by the end of 2021. The scale of industrial deployments that allow enterprises to deliver the services we rely upon, such as our electricity, water, gas, or the reliability of supply infrastructure of food, medicines, etc., is significantly larger.

The adoption of 5G is going to accelerate this deployment of connected devices. But as with most disruptive new technology waves, 5G brings its own challenges, causing designers to consider how they support new radio bands and adding complexity to IoT device design.

Increasingly, the devices in question are low power, long-lived in the field, low cost, and afford low levels of physical access, plus they're deployed across further distances and remote locations. Cellular IoT's ubiquitous connectivity is a key draw for designers in fast-growing markets, such as infrastructure, for more intelligent transport, smart consumer and smart city devices, and connected health.

Cellular IoT has long been the choice of secure, large-scale, and resilient deployments due to the robust subscriber identity module (SIM) that authenticates a device. As our devices shrink, and as remote devices must endure a much wider range of environmental conditions, securing the device's identity requires new and broader solutions.

It's not just the physical design that's a consideration when leveraging SIMs and cellular IoT. Scale is both the IoT's most significant opportunity and hurdle, and organizations need to consider how they will manage and secure the device throughout an extended lifecycle.

Think about some of the use cases that 5G will enable in smart cities and logistics. The deployment, management, and swapping of physical SIMs to update user profiles or switch network providers in these remote locations presents a considerable cost to OEMs and end users alike.

Introducing the eSIM

Embedded SIM (eSIM) technology is still a hardware-based SIM, but this elegant, robust, and scalable technology is soldered permanently into the device and was designed to address some of the challenges impeding true scalability in cellular IoT *(see figure on page 29)*. eSIM allows devices to be deployed anywhere with existing cellular coverage. Operator profiles or network providers can be updated over the air, based on standards that offer a frictionless experience for device manufacturers and service operators.

iSIM Builds on eSIM

But there's more: An integrated SIM (iSIM) takes all of the benefits of eSIM and embeds them into the device's permanent hardware array by combining the SIM with the system-on-chip (SoC) architecture and cellular modem. Fusing the secure locations into the chipset

itself offers a low footprint. It also introduces extra layers of security through a hardware-based secure enclave (a dedicated processor for security operations) that maintains the integrity of all cryptographic and key managed operations.

Security and Scale Go Hand-in-Hand

As the number of devices ramps up, the IoT attack vector will grow exponentially—security can't be an afterthought. IoT continues to move closer to core processes, and businesses should ensure that both devices (i.e., the endpoint itself) and data exchange (in technical terms, the chip-to-cloud security) have strong identity and trust foundations.

To achieve scale, enterprises need simpler ways of manufacturing while being able to keep the robust security benefits derived from SIM capabilities. This has typically proven to be a challenge as new market needs for IoT are emerging.

The fastest growth of eSIM and iSIM deployments is coming from markets that haven't traditionally operated at the same cost points as cellular or smartphone industry. Markets such as fleet management of e-bikes or e-scooters, or connected health wearables, didn't exist in their current capabilities a few years ago.

Within these, enterprises need simplification of both the bill of materials and that of the supply chain. iSIM fundamentally changes the way device makers can access cellular capabilities for devices that couldn't be served before. iSIM offers the highest protection for subscription credentials and isolates processing in a secure enclave. An additional authentication layer serves as a root of trust for secure communications while reducing the bill of materials.

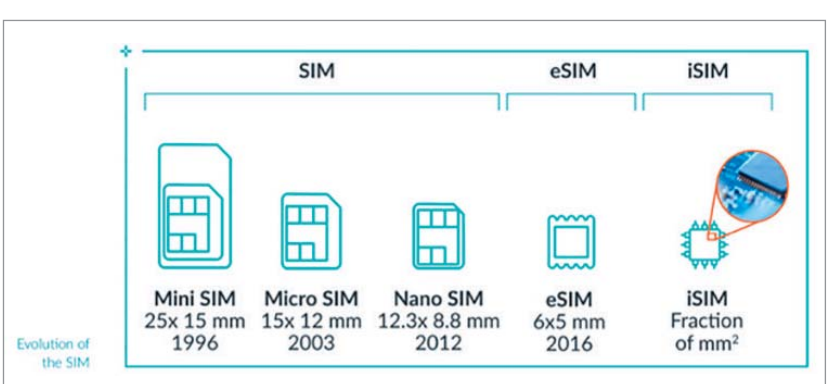
The GSMA standard IoT-SAFE (IoT SIM Applet For Secure End-2-End Communication) for iSIM supports the entire secure chip-to-cloud IoT infrastructure. It allows iSIMs to authenticate both connectivity and application credentials for any data being exchanged with any cloud over any cellular network by any device.

New Opportunities

eSIM and iSIM are solutions to existing problems, but they also open up new opportunities for the broader IoT ecosystem:

"In the IoT space, cost is everything and this is where iSIM stands out. Their significant cost difference enables cellular applications that weren't previously viable. It's a whole new IoT world that we're just starting to explore. And iSIM makes this new world possible."—Alex Sinclair, GSMA Chief Technology Officer

Device and module makers benefit from more streamlined processes, such as reducing SKUs that contain multiple SIMs for individual regions, freeing them to focus on value-add services in a number of use cases. For example, we're seeing logistics companies utilize cellular M2M devices featuring multiple device sensors powered by eSIM technology to provide live global tracking. Smart energy providers are turning to low-power wide-area networks (LPWANs) to deploy their eSIM-enabled smart energy devices globally.



The subscriber identity module (SIM) has progressed from a card to the embedded SIM (eSIM) and now the integrated SIM (iSIM).

> **"In the IoT space, cost is everything and this is where iSIM stands out. Their significant cost difference enables cellular applications that weren't previously viable. It's a whole new IoT world that we're just starting to explore. And iSIM makes this new world possible."**

In our cities, automotive OEMs are tapping into 5G to power their "cellular vehicle-to-everything" (C-V2X) service, serving as a means for cars to communicate with everything around them. Meanwhile, IoT service providers can expand beyond the constraints of Wi-Fi without the need for costly infrastructure by tapping into LTE-M and NB-IoT technology, exposing them to greater market opportunity.

The use cases are as broad as they are attractive. eSIM and iSIM are paving the way for centralization of agricultural operations, as smaller, less expensive tracking devices can be deployed in remote, rugged terrain. And global logistics providers can track more assets in real-time with low-power M2M modules that go for months between charges. The opportunities presented by cellular IoT, underpinned by 5G, are abundant. iSIM technology is reducing complexity and cost as well as providing the secure foundation that will allow the IoT to scale in the 5G wave.

WILLIAM WONG | Senior Content Director

AI on the Edge Comes in Compact Packages

The Jetson Nano and Xavier NX power these small, yet powerful and rugged systems.

The Jetson Nano and Jetson Xavier NX from NVIDIA share a common 260-pin SO-DIMM connector *(Fig. 1)*. They target compact, low-power applications that leverage artificial-intelligence/machine-learning (AI/ML) models.

1. NVIDIA's Jetson Nano, with its 260-pin SO-DIMM connector, is plug-compatible with Xavier NX. Both can handle AI/ML chores.

The Jetson Nano has a quad-core Arm Cortex-A57 with a 128-core Maxwell GPU delivering 472 GFLOPS of performance, while the Xavier NX delivers 21 TOPS via a six-core, 64-bit NVIDIA Carmel Arm v8.2 and a Volta GPU with 384 cores and 48 Tensor cores. The former can run with as little as 5 W; its heftier sibling uses 10 to 15 W.

The modules are available separately, but they require a carrier board to make things work. Here we look at three complete solutions that can host either module. The specific feature sets can differ depending on the module used, so check the specs before ordering.

2. The JETBOX-Floyd from Diamond Systems features dual Ethernet and dual HDMI ports.

Diamond Systems' JETBOX-Floyd is the quintessential embedded platform that comes installed with Linux *(Fig. 2)*. It features dual Ethernet and dual HDMI ports. There are versions with a x4 PCIe M.2 socket capable of handling NVMe flash storage. There's also an SD card slot. The system is DIN-rail-mountable and has dual SMA antenna cutouts. A mini-PCIe socket can handle wireless options.

Axiomtek's AIE100-903-FL is a fanless and wireless edge AI solution *(Fig. 3)*. One of two Gigabit Ethernet ports supports Power over Ethernet (PoE) to handle a remote smart camera. It's a great solution for smart-city applications where streaming video is processed at the edge instead of the cloud.

3. The AIE100-903-FL, developed by Axiomtek, is well-suited for smart-city applications.

4. Adlink's DLAP-301-Nano includes an eight-port, Power-over-Ethernet switch in addition to an NVIDIA Jetson Nano.

Adlink's DLAP-301-Nano sports an eight-port, PoE-enabled switch *(Fig. 4)*, allowing the NVIDIA module to handle up to eight PoE IP cameras. A separate Gigabit Ethernet port is available for connecting to a LAN. The system has an HDMI display output as well as 8-bit digital inputs/outputs. The platform works well as a network video recorder (NVR) used in surveillance applications.

NVIDIA provides a tremendous amount of free, open-source software, starting with its CUDA platform that's included as part of the NVIDIA JetPack bundle. JetPack is compatible with the entire Jetson family. ED

IDEA!

Industrial Design
Engineering & Automation

A HYBRID EVENT TAKING PLACE LIVE AT THE HUNTINGTON CONVENTION CENTER IN CLEVELAND ON NOVEMBER 9-11, 2021 & VIRTUALLY ON DECEMBER 7-8, 2021

CONFERENCE SESSIONS & TOPICS

- Automation & IoT IDEAs
- Advanced Manufacturing IDEAs
- Power & Motion IDEAs
- Laser Assisted Processing IDEAs
- Machine Vision IDEAs

Featured During

M&T THE MANUFACTURING & TECHNOLOGY SHOW

STRATEGIES FOR EXCELLENCE
FROM DESIGN TO DELIVERY

REGISTER NOW & SAVE $200!

WWW.MFGTECHSHOW.COM

(Continued from page 21)

regulators. There's a limited selection, but they tend to be sufficient for most designs. The colors change to green as the various pieces are linked together.

I also added a USB 3.0 connector and an M.2 socket for an NVMe memory module with logic connectors to the SMARC module.

Well, Gumstix is now part of Altium and Geppetto is now Altium's Upverter. Adlink has partnered with Altium to support its SMARC modules. Significant advances have been made to Geppetto/Upverter since I first looked at it. For instance, PCI Express support was a major jump forward. It's possible to drop USB hubs and Ethernet switches onto the board, not just connectors. There are wireless and security options, and headers come in various forms.

The system doesn't solely verify connections—power requirements are considered as well. Trying to put too much of a load on a part will prompt for adding or swapping in a part that has more capacity.

The best way to see what the system can do is to try it. It's free—cost enters the picture if you want to buy a board.

A complete custom design using a tool like Altium Designer can come up with more advanced solutions than is possible with Upverter. However, the cost, verification, and time to market is significantly longer because there's another piece to the Upverter puzzle.

2. Developers can use Altium's Upverter online board designer to quickly create carrier boards.

Once you have a design, you also know how much it will cost to get fully functional boards. It has a one-time $1,999 startup charge and then a per-board cost for delivery of fully functional boards. Upverter also lists the cost of individual items. On top of that, the system generates a bill of materials (BOM) and documentation, including device tree definitions for Linux and other operating systems.

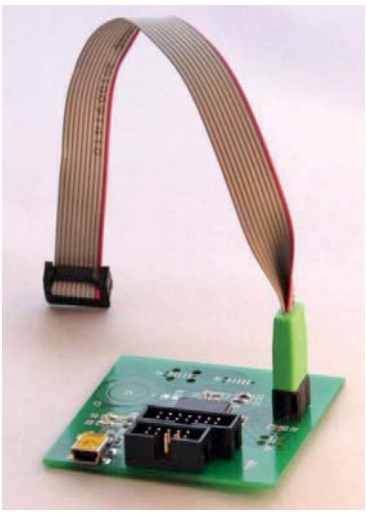
Also, SMARC is only one of many module form factors supported by Upverter. It's even possible to design boards for platforms like Raspberry Pi.

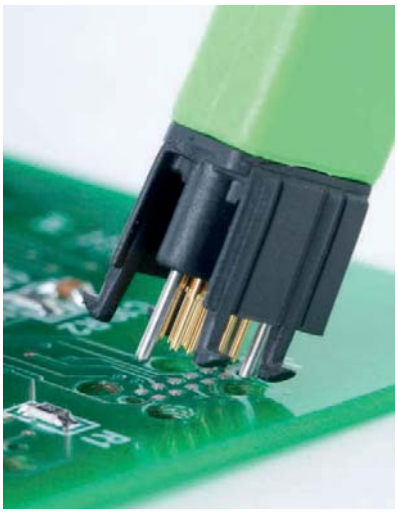
Time to market is one reason for using modules like Adlink's SMARC solutions. Getting a carrier board is a necessary requirement. Many designs will warrant more sophisticated designs than can be provided by Upverter, but it will be hard to beat the cost and turnaround time of Upverter if it can meet your needs.

Now if they could just come up with a better name. ed

Ad Index

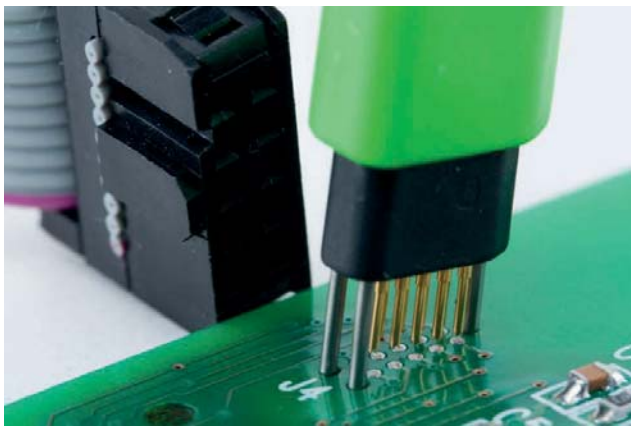
Tag-Connect

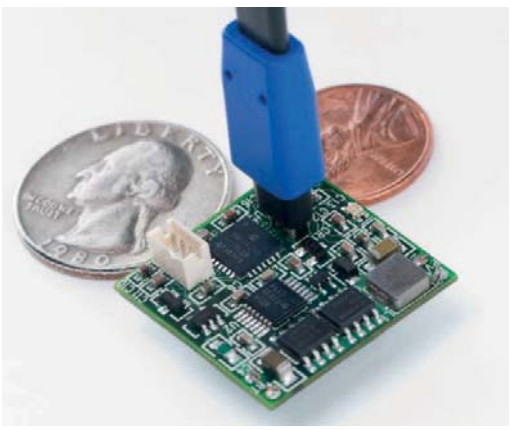




JTAG Connector Plugs Directly into PCB!!

No Header! No Brainer!





Our patented range of Plug-of-Nails™ spring-pin cables plug directly into a tiny footprint of pads and locating holes in your PCB, eliminating the need for a mating header. Spring-Pins tested to over 150,000 cycles!

Save Cost & Space on Every PCB!!

Solutions for: PIC . dsPIC . ARM Cortex . Atmel . JTAG . Altera . Xilinx . TI DSP . MSP430 . SPI / IIC . FTDI UART & More

Tag-Connector footprints as small as 0.02 sq. inch (0.13 sq cm)

www.Tag-Connect.com

ANALOG DEVICES

AHEAD OF WHAT'S POSSIBLE™

ADAQ23875

16-Bit, 15 MSPS, µModule® Data Acquisition Solution

- Includes front-end signal conditioning functions in a 9 mm × 9 mm BGA package.
- Low gain and offset error drift due to *i*Passives® signal scaling resistors.
- Reduces signal chain calibration, maximizes channel-to-channel matching performance.

Buy Now at mou.sr/ADAQ23875

MOUSER ELECTRONICS