

IEEE Industrial Electronics

JUNE 2022 – VOLUME 16 NUMBER 2

MAGAZINE

Information Technology in Industrial Electronics

Blockchain, Factory 5G, Wireless
Time-Sensitive Network and Edge Computing



*Dr. Bimal Bose contributes to
"My View" (page 65) on "Power Electronics –
My Life and Vision for the Future"*





Winners of the 8th Nagamori Awards Selected

Since motors appeared in the early 19th century, they have been used in all types of electrical appliances and are now an indispensable part of our daily lives. Today, a huge number of motors are used in a wide range of applications, and it is claimed motors account for more than 55% of the world's power consumption.

Therefore motor research is extremely important if we are to maintain our affluent lives while also perpetually conserving the global environment.

We created these Nagamori Awards to bring vitality to technological research of motors and related fields, such as generators and actuators, and also to support the researchers and development engineers who strive each day to fulfill their dreams.

S Nagamori



Nagamori Foundation decided six winners of the 8th Nagamori Awards, from whom the Grand Nagamori Award winner will be chosen on September 4th. The Grand Nagamori Award winner will receive 5 Million JPY and each Nagamori Award winner, 2 Million JPY.

The 8th Nagamori Awards Winners:

Huijun Gao

Professor and Director, Research Institute of Intelligent Control and Systems, Harbin Institute of Technology
For contributions to the advanced control for mechatronic systems

Yunwei Ryan Li

Professor and Acting Department Chair, Department of Electrical and Computer Engineering, University of Alberta
For contribution to the PWM, control and converter topology of medium voltage high power industrial drives

Burak Ozpineci

Section Head, Vehicle and Mobility Systems Research Section, Building and Transportation Science Division, Oak Ridge National Laboratory
Low cost, high efficiency, compact electric motor drives for more electrified transportation systems

Gianmario Pellegrino

Full Professor, Department of Energy "Galileo Ferraris", Politecnico di Torino
Synchronous and PM-synchronous reluctance motor drives - theory, design, and control methods

Maryam Saedifard

Associate Professor, School of Electrical and Computer Engineering, Georgia Institute of Technology
For contributions to highly-efficient, power-dense and fault-tolerant multilevel converter-based medium-voltage drives

Akio Yamamoto

Professor, Graduate School of Frontier Sciences, The University of Tokyo
Pioneering research and development on theoretical models and applied systems for electrostatic film actuators

Contact information: Nagamori Foundation
Address: 338 Kuzetonoshiro-cho, Minami-ku, Kyoto 601-8205, JAPAN
Tel: +81-75-935-7731 E-mail: n.awards@nidec.com
<https://www.nidec.com/en/nagamori-f/>

IEEE Industrial Electronics

JUNE 2022 — VOLUME 16 NUMBER 2

MAGAZINE

Features

- 4 Blockchain: What Does It Mean to Industrial Electronics? Technologies, Challenges, and Opportunities**
Xinghuo Yu, Changbing Tang, Peter Palensky, and Armando Walter Colombo
- 15 Dependency-Aware Tensor Scheduler for Industrial AI Applications**
Dymem—An Aggressive Data-Swapping Policy for Training Nonlinear Deep Neural Networks
Wei Rang, Donglin Yang, and Dazhao Cheng
- 24 Factory 5G**
A Review of Industry-Centric Features and Deployment Options
Amir Mahmood, Sarder Fakhru Abedin, Thilo Sauter, Mikael Gidlund, and Krister Landernäs
- 35 Clock Synchronization for Wireless Time-Sensitive Networking**
A March From Microsecond to Nanosecond
Óscar Seijo, Iñaki Val, Michele Luvisotto, and Zhibo Pang
- 44 Data-Driven Edge Computing**
A Fabric for Intelligent Building Energy Management Systems
Zhishu Shen, Jiong Jin, Tiegua Zhang, Atsushi Tagami, Teruo Higashino, and Qing-Long Han
- 53 System-on-Chip FPGA Devices for Complex Electrical Energy Systems Control**
Eric Monmasson, Mickaël Hilairt, Giovanni Spagnuolo, and Marcian N. Cirstea

Departments and Columns

- | | |
|-------------------------------------|---|
| 2 EDITOR'S COLUMN | 82 WOMEN IN IES NEWS |
| 3 MESSAGE FROM THE PRESIDENT | 84 STUDENTS AND YOUNG PROFESSIONALS NEWS |
| 65 MY VIEW | 90 BOOK NEWS |
| 73 HISTORICAL | 92 CALENDAR |
| 78 SOCIETY NEWS | |
| 79 CHAPTER NEWS | |

SCOPE—IEEE Industrial Electronics Magazine (IEM) publishes peer-reviewed articles that present emerging trends and practices in industrial electronics product research and development, key insights, and tutorial surveys in the field of interest to the membership of the IEEE Industrial Electronics Society (IEEE/IES). IEM is limited to the scope of the IES, which is given as theory and applications of electronics, controls, communications, instrumentation, and computational intelligence to industrial and manufacturing systems and processes.

IEEE Industrial Electronics Magazine (ISSN 1932-4529) (IEMAW) is published quarterly by The Institute of Electrical and Electronics Engineers, Inc. Headquarters: 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA +1 212 419 7900. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society, or its members. The magazine is a membership benefit of the IEEE Industrial Electronics Society, and subscriptions are included in Society fee. Replacement copies for members are available for US\$20 (one copy only). Nonmembers can purchase individual copies for US\$89. Nonmember subscription prices are available on request. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of the U.S. Copyright law for private use of patrons: 1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01970, USA; and 2) pre-1978 articles without fee. For other copying, reprint, or republication permission, write to: Copyrights and Permissions Department, IEEE Service Center, 445 Hoes Lane, Piscataway NJ 08854 U.S.A. Copyright © 2022 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals postage paid at New York, NY and at additional mailing offices. Postmaster: Send address changes to IEEE Industrial Electronics Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188 Printed in USA

Digital Object Identifier 10.1109/MIE.2022.3166358

EDITOR-IN-CHIEF

Prof. Eric Monmasson
CY Cergy Paris University, France
eric.monmasson@cyu.fr

PAST EDITOR-IN-CHIEF

Prof. Peter Palensky
TU Delft, The Netherlands

EDITORIAL BOARD

Prof. Kamal Al-Haddad
École de Technologie Supérieure, Canada

Prof. Seta Bogosyan
University of Alaska-Fairbanks, USA

Prof. Bimal K. Bose
University of Tennessee, USA

Prof. Chandan Chakraborty
Indian Institute of Technology, India

Dr. Michael W. Condry—*Industry Forum*, Intel, USA

Prof. Fernando da Silva—*Book News*
Technical University of Lisbon, Portugal

Prof. Jose Alfonso Antonino Daviu
Universitat Politècnica de Valencia, Spain

Dr. Giovanni De Carne
Karlsruhe Institute of Technology, Germany

Prof. Tomislav Dragicevic
Technical University of Denmark, Denmark

Prof. Hiroshi Fujimoto
University of Tokyo, Japan

Prof. Luis Gomes
New University of Lisbon, Portugal

Prof. Massimo Guarneri—*Historical*
University of Padua, Italy

Prof. Josep Guerrero
Aalborg University, Denmark

Prof. Qing-Long Han
Swinburne University of Technology, Australia

Gerhard P. Hancke
City University of Hong Kong, China

Dr. Victor K.L. Huang—*News in Industry Activities*,
Industry Forum, Better World and ZAP Motors, USA

Dr. Martin Gilje Jaatun
SINTEF, Norway

Dr. Marek Jasinski—*Students and Young Professionals News*
Warsaw University of Technology

Prof. Okyay Kaynak
Bogazici University, Turkey

Prof. Marian Kazmierkowski—*Book News*
Warsaw University of Technology, Poland

Prof. Marco Liserre
Christian-Albrechts University, Kiel, Germany

Prof. Lucia Lo Bello—*Women in IES News*
University of Catania, Italy

Prof. Oscar Lucia—*Society News*
University of Zaragoza, Spain

Prof. Antonio Luque Estepa
University of Sevilla, Spain

Prof. Mariusz Malinowski
Warsaw University of Technology, Poland

Prof. João F. Martins
NOVA University of Lisboa, Portugal

Dr. Dorin Neacsu
Technical University of Iasi, Romania

Prof. Kouhei Ohnishi
Keio University, Japan

Prof. Peter Palensky
TU Delft, The Netherlands

Prof. Marco Porta
University of Pavia, Italy

Prof. Juan José Rodríguez-Andina
University of Vigo, Spain

Prof. Yang Shi
University of Victoria, Canada

Prof. Kim-Fung Tsang—*Chapter News*
City University Hong Kong, China

Prof. Ligang Wu
Harbin Institute of Technology, China

Prof. Fuwen Yang
Griffith University, Australia

Prof. Dong Yue
Nanjing University of Posts and
Telecommunications, China

Dr. Richard Zurawski, Atut Technology, USA

IEEE PERIODICALS

MAGAZINES DEPARTMENT

Eric Charbonneau, *Journals Production Manager*
Patrick Kempf, *Manager, Journals Production*
Janet Dudar, *Senior Art Director*
Gail A. Schnitzer, *Associate Art Director*
Theresa L. Smith, *Production Coordinator*
Mark David, *Director, Business Development—
Media & Advertising*, +1 732 465 6473
Felicia Spagnoli, *Advertising Production Manager*
Peter M. Tuohy, *Production Director*
Kevin Lisankie, *Editorial Services Director*
Dawn M. Melley, *Senior Director, Publishing Operations*

ON THE COVER:

©SHUTTERSTOCK.COM/VOB STUDIO

IEEE prohibits discrimination, harassment, and bullying.
For more information, visit <http://www.ieee.org/ueh/aboutus/whatis/policies/p9-26.html>.

Digital Object Identifier 10.1109/MIE.2022.3166359



Advancing Technology for Humanity: More Than Words

Important transformations in the field of industrial electronics, whether in terms of paradigms, such as blockchain and data-driven systems, or industrial wireless networks and devices, are the subject of this issue. First, Xinghuo Yu et al., in “Blockchain: What Does It Mean to Industrial Electronics?” review basic concepts, key features and technologies, and future challenges and opportunities of blockchain, analyzing the technology’s future impact on major areas of industrial electronics. Then, in “Dependency-Aware Tensor Scheduler for Industrial AI Application,” Wei Rang et al. demonstrate why it is important, at a time when deep neural networks (DNNs) in artificial intelligence applications are spreading in the industrial community, to effectively schedule GPU memory for DNN training. To this end, innovative memory management for training nonlinear DNNs is proposed.

“Factory 5G: A Review of Industry-Centric Features and Deployment Options,” by Aamir Mahmood et al., exposes how 5G wireless communications is key to future smart manufacturing’s strive for digitization and agile operations. The topic of wireless industrial communication continues with “Clock Synchronization for Wireless Time-Sensitive Networking: A March From Microsecond to Nanosecond,” by Óscar Seijo et al., who

discuss the need for accurate clock synchronization in high-performance industrial wireless networks and technologies to achieve it.

The final two articles are “Data-Driven Edge Computing: A Fabric for Intelligent Building Energy Management Systems,” by Zhishu Shen et al., and “System-on-Chip FPGA Devices for Complex Electrical Energy Systems Control,” by Eric Monmasson et al. In the first one, the authors propose building energy management systems with a data-driven edge fabric for the future Building 4.0. In the second, the authors show how system-on-chip (SoC) field-programmable gate arrays (FPGAs) are good candidates for implementing future edge computing platforms to address complexity and storage challenges resulting from data-driven approaches. SoC FPGAs can make significant contributions to key developments in complex electrical energy systems.

In the editorials, we welcome Prof. Bimal K. Bose, a living legend in the field of power electronics, who shares his professional experiences and vision for the future. His article complements the March issue, which was dedicated to this topic.

I would also like to inform potential authors about two recent changes. First, as of 1 March, it is mandatory that authors’ primary email address be an institutional one. Noninstitutional email addresses can be used only for primary carbon copies. This measure has successfully been implemented in

IEEE Transactions on Industrial Electronics, and it aims to protect authors’ identity and integrity. Second, I am pleased to announce that *IEEE Industrial Electronics Magazine* will be integrated with IEEE DataPort, so authors will have an option to upload their research data sets when they submit articles. If an article is accepted, the data set will be automatically linked to *IEEE Xplore*.

Finally, let me echo the IEEE Industrial Electronics Society president’s message about the devastating situation in Ukraine and extend a prayer that the war ends soon. The values of peace, tolerance, and mutual aid that are at the heart of our international scientific community must prevail. A dramatic event like the Ukraine crisis makes us realize that the IEEE slogan “Advancing Technology for Humanity” is not a string of words but a noble objective. This humble magazine is animated by this vision and always will be.



by Mariusz Malinowski



Solidarity With Ukraine

I don't think any of us expected the beginning of the year to be so difficult. The pandemic has receded into the background, not because it suddenly ended but because of the even more significant threat of Russia's unjustified attack on Ukraine. This has destabilized the world as we know it and reminds us of the old days of global tension in international relations. As I write these words, I am 200 km from the Ukrainian border, on the other side of which a brutal war is taking place, causing suffering to millions of people, including our colleagues. Three weeks after the assault began, 2 million refugees, mainly women and children, have fled to Poland alone.

It is impossible to write in such circumstances about the journals,

THREE WEEKS AFTER THE ASSAULT BEGAN, 2 MILLION REFUGEES, MAINLY WOMEN AND CHILDREN, HAVE FLED TO POLAND ALONE.

conferences, and other activities of our Society, though they are essential. However, we are reminded that the IEEE Industrial Electronics Society's most precious asset

is its people, its members. Being together enables us to build ties that we harness to carry out joint scientific and industrial projects. Thanks to these relationships, we break through stereotypes and build understanding, getting to know different cultures and customs. Therefore, faced with the suffering

surrounding us, it is important not to pretend that nothing has happened. We cannot remain silent, and we cannot be passive. Our solidarity, sympathy for Ukraine, and material support are crucial these days.



IMAGE LICENSED BY INGRAM PUBLISHING

I think this is also a test for society. New technologies, mainly the Internet, have not killed the spirit of humanity and sensitivity we have in us. Rather, they are helping us to help, express our sympathy, and condemn what is wrong. I hope that by the time these words are published, this bad dream will be over, and peace will have come, which would allow us to rebuild strong relationships of cooperation without divisions and drives for conquest.



Digital Object Identifier 10.1109/MIE.2022.3166285

Date of current version: 24 June 2022

Technologies, Challenges, and Opportunities

XINGHUO YU, CHANGBING TANG,
PETER PALENSKY, and
ARMANDO WALTER COLOMBO



©SHUTTERSTOCK.COM/PHIVE

Blockchain: What Does It Mean to Industrial Electronics?

What Is Blockchain?

Imagine you want to send money to a friend overseas. Wouldn't it be good if you didn't have to pay hefty fees to the intermediaries, and your friend received the funds very quickly? Now imagine ordering parts to make a product in your manufacturing plant. Wouldn't it be great if you were able to verify where each part comes from and have access to a reliable certificate on its quality automatically? Also think about dealing with energy use

or selling off your excess solar energy as a prosumer. Wouldn't it be nice if you could purchase cheaper energy or sell it profitably at ease?

Blockchain can resolve these challenges. Blockchain is a distributed ledger of transaction and data management technology that enables distributed nodes to collaboratively affirm transaction provenance via a decentralized consensus mechanism. The interest in blockchain has been increasing exponentially in both industry and academia because of its potential to revolutionize modern industries and businesses [1], [2].

The Concept of Blockchain

The term *blockchain* was coined in the 2008 article "Bitcoin: A Peer-to-Peer Electronic Cash System," by Nakamoto [3]. In a narrow sense, it is a chained data structure storing data blocks sequentially and a nontamperable and unforgeable distributed ledger that is secured cryptographically. Broadly speaking, it can be considered a new distributed infrastructure and computing paradigm using chained block data structures to store and validate data, node consensus algorithms to generate and update data, cryptography to secure data transmission and access,

Digital Object Identifier 10.1109/MIE.2021.3066332
Date of current version: 12 April 2021



and smart contracts with automated scripts to program and manipulate data (Figure 1 illustrates how it works).

Currently, blockchain technology is regarded as a breakthrough that is changing the ways businesses and organizations operate [4]. Just like modern information technologies, such as big data, cloud computing, and the Internet of Things (IoT), it relies on existing technologies to deliver its promises.

The Journey of Blockchain

The development of blockchain technology has gone through three phases, namely, programmable currency, programmable finance, and programmable society, dubbed *Blockchain 1.0*, *2.0*, and *3.0*, respectively.

Soon after publishing [3], Nakamoto created software in 2009 to mine the foundation block, opening the era of Bitcoin. The initial interest in blockchain was in virtual currencies, i.e., for Blockchain 1.0, how much they were worth, how to mine, how to buy, and how to sell. A few years later, attention was placed on the technology itself, leading to a big step forward—Blockchain 2.0—marked by the publication of the “Ethereum Whitepaper” in 2013 [5].

Ethereum is a platform that offers a variety of modules allowing users to build applications. It works like building a house, where Ethereum provides building modules, such as the walls, roof, and floor, and customers need only to assemble the house using the modules. The core of Ethereum is the smart contract, which is an automated agent. However, Blockchain 2.0 could achieve only 70–80 transactions per second, which hindered its applications. Recent years have seen the emergence of Blockchain 3.0, which is a platform that is able to process the volumes of transactions necessary for mass adoption. It presents the future of blockchain: a decentralized Internet with data storage, smart contracts, cloud nodes, and open-chain networks, applicable to a wide range of fields, from finance to manufacturing, energy, logistics, medicine, and social networks. The journey of the blockchain developments is illustrated in Figure 2.

The Key Technologies of Blockchain

There are four key traditional technologies of blockchain: distributed storage, cryptography, consensus algorithms, and smart contracts (see Figure 3).

Distributed storage is used for data sharing and synchronization in a network composed of many distributed nodes in different physical addresses or organizations. Each participating node has complete data storage and is independent and peer-to-peer connected.

Blockchain relies on distributed storage to ensure reliability and security of the data, and increasing the number of participating nodes would enhance their improvements. On one hand, the technology generates block hard forks to achieve transaction rollback and avoid malicious tampering of data. On the other hand, it leads to a significant increase in storage.

Cryptography is used for addressing information security issues. Famous algorithms include hashing algorithms, encryption and decryption algorithms, digital certificates and signatures, and zero-knowledge proofs [6]. Hash algorithms generate header information for each unit (block) in the blockchain. The connection between the blocks is achieved by including the previous block header information in the next block header. Meanwhile, hash-based tree structures, such as the Merkle tree, are used to organize the specific transactions or states in the block and store the summary information (root hash) in the block header, making it extremely difficult to tamper with. The storage structure of blockchain is like a zipper: after each data item is stored independently, a chain is formed, and any node can be traced. In this process, the signature is determined by cryptography, and a zero-knowledge proof plays an increasingly important role in convincing a verifier that a certain assertion is deemed correct without

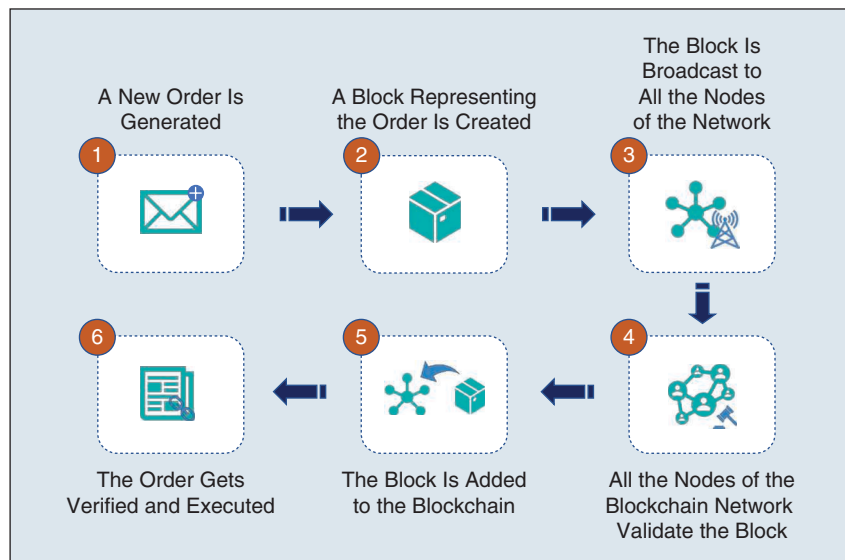


FIGURE 1 – How a blockchain works.

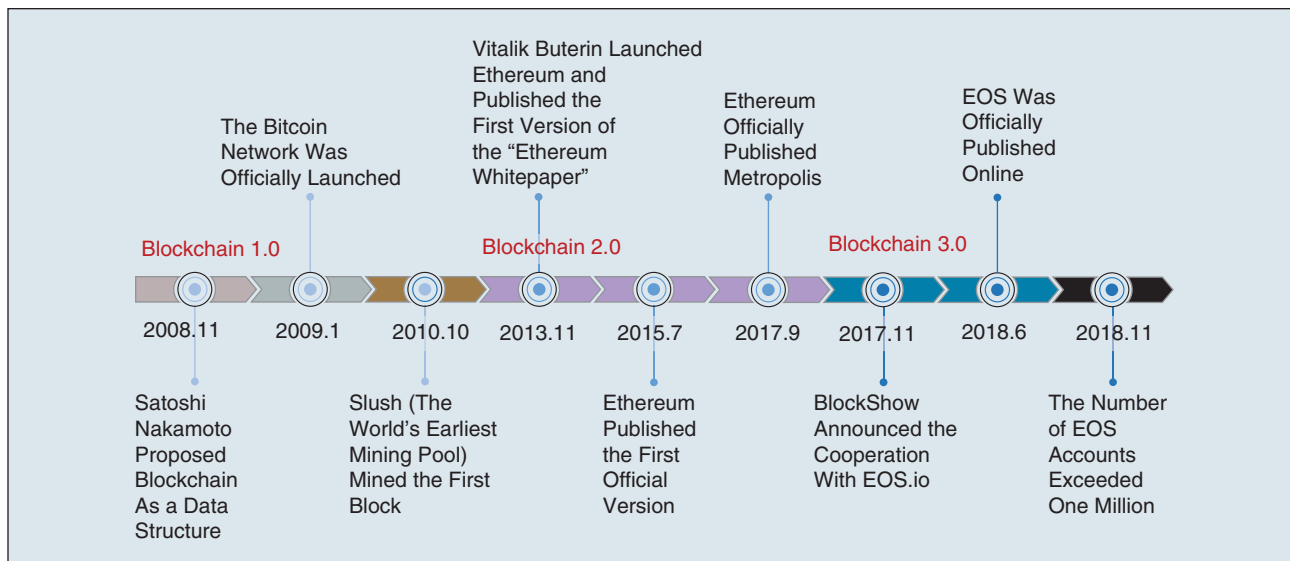


FIGURE 2 – The journey of blockchain. EOS: enterprise operating system.

providing any information to the verifier (e.g., Zcash [7] and zk-SNARKs [8]).

Consensus algorithms refer to how all nodes reach consensus to validate a record, which is used for both identification and tampering prevention to maintain decentralized multiparty mutual trust. In both public and private blockchains, all consensus algorithms achieve the same goal of determining which blocks are correct by checking how each block is added. Their differences lie in which blocks can be added on the chain at what rate, and what types of faults are allowed.

There are many different classifications for consensus algorithms [9]. According to the deployment mode, the blockchain consensus algorithms can be divided into public chain consensus,

alliance chain consensus, and private chain consensus algorithms. With regard to the fault-tolerant type, they can be classified as Byzantine fault tolerant (BFT) and non-BFT. Considering the degree of consistency, they can also be divided into strong consensus and weak consensus algorithms. In this article, we classify the consensus algorithms into four types, namely, BFT-based, Proof-of-Work (PoW)-based, Proof-of-Stake (PoS)-based, and mixed-type consensus algorithms.

BFT-based consensus algorithms are based on traditional distributed consistency-checking techniques; some examples are Paxos [10], Raft [11], Practical BFT [12], Stellar Consensus Protocol [13], Algorand [14], and Sleepy Consensus [15]. PoW-based consensus algorithms aim to

achieve capacity expansion of the blockchain (e.g., Bitcoin-Next Generation [16] and Elastico [17]) or improve the efficiency of the algorithm (e.g., Proof of Elapsed Time [18], Proof of Luck [19], Proof of Space [20], and Proof of Useful Work [21]). PoS-based consensus algorithms are used to solve the problem of “nothing at stake” [22], including Delegated Proof of State [23], Tendermint [24], Casper [25], and Proof of Unspent Transaction Output [26]. The mixed-type consensus algorithms draw lessons mainly from the consensus of PoW and PoS, including Proof of Stake Velocity [27], Proof of Burn [28], and Proof of Activity [29]. In short, all blockchain consensus algorithms focus primarily on three aspects: performance evaluation, adaptation and optimization, and consensus innovation under the new blockchain structure. For a comprehensive survey of various consensus algorithms, please refer to [30].

A smart contract refers to a computing protocol for disseminating, verifying, and performing a contract negotiation or fulfillment of a contract in an informational manner. Its concept was originated by Szabo in 1994 [31]. As a kind of embedded programming, smart contracts can be built into any blockchain data, trading, and tangible or intangible assets to form a programmable control system. The key property of a smart contract is that it does not rely on third-party or centralized

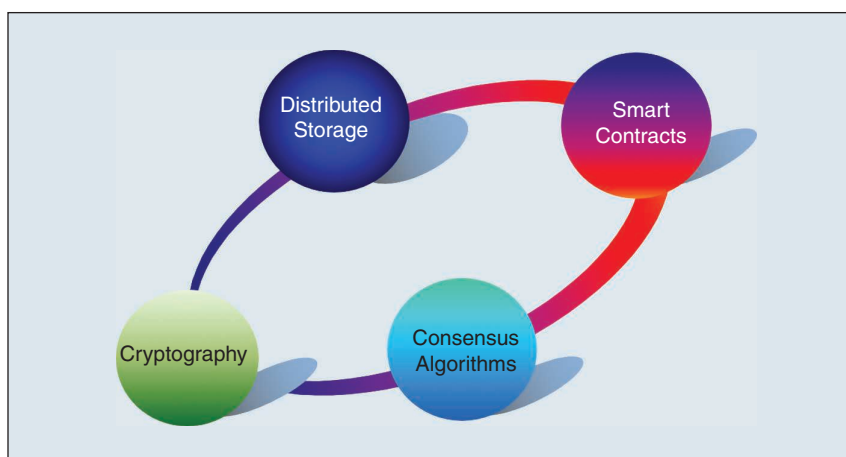


FIGURE 3 – The four key technologies of blockchain.

organization, which greatly reduces manual participation and cost with high efficiency and accuracy. It is noted that all smart contracts deployed on the blockchain public chain are visible and interactive, meaning that their vulnerabilities are made public.

A smart contract in blockchain is a set of codes automatically executed once an event triggers a clause in the contract. In the blockchain context, smart contracts are scripts stored on the blockchain, which are analogous to stored procedures in relational database management systems. According to the performance of the programming language or running environment, smart contracts can be divided into three types: script type, Turing-complete type, and verifiable-contract type [32]. Smart contracts have been successfully implemented on many blockchain systems, such as Ethereum [5] and Hyperledger Fabric [33]. Hyperledger Fabric has good flexibility, scalability, and versatility and supports various uncertain smart contracts and pluggable services. In short, the smart contract is implemented based on program code. Once deployed to the blockchain, it is not allowed to change, which eliminates the possibility of human intervention. However, there are still some limitations on the technology and implementation of smart contracts, especially the problems of stability and security. A comprehensive survey on this topic can be found in [34].

The Main Platforms of Blockchain

Blockchain platforms combining distributed storage, cryptography, consensus algorithms, and smart contracts together with network and data technologies are used for building blockchain-based systems. There are some quite generic platforms that can be used for different industrial domains, such as Ethereum and Hyperledger Fabric. Ethereum supports applications that use smart contracts, while Hyperledger Fabric provides good flexibility and versatility support for blockchain applications in domains such as finance, manufacturing, and logistics. Other platforms are more specialized and developed for

specific domains, such as Energy Web Foundation (EWF) [35] and Obelisk [36] for smart energy systems, Provenance [37] for logistics, Gem [38] for health care, and Genesis of Things [39] for 3D manufacturing. Generally, the selection of a blockchain platform is dependent on the needs of the users. For example, multiple collaborative diverse companies can use a platform like Ethereum to implement smart contract capabilities over their network, while a group of energy providers can use one platform like EWF that supports energy trade applications.

The Key Issues and Challenges in Blockchain

Blockchain has now become a huge technical field that is profoundly changing industry, economy, and society. However, there are many issues and challenges, as discussed in the following sections.

Technological Issues

The breakthrough construction of blockchain technology is limited by a famous theory: *the impossible triangle theory*; i.e., scalability, security, and decentralization cannot be achieved at the same time (see Figure 4). For example, Bitcoin is highly decentralized and secure, but its performance (its so-called *scalability*) is very low. Because of frequent network congestion, traders have to pay more in the transaction process. Therefore, one challenge is to address the impossible triangle problem to balance scalability, security, and decentralization.

Scalability refers to the ability to handle high volumes of business data. As usual, there is always a tradeoff among costs, security, and performance. To achieve scalability, we should consider the usage context and the performance metrics, such as validation latency, transaction throughput, energy costs, computation costs, storage costs, number of nodes, and so on. For example, the throughput of a blockchain is not scalable when the network size grows. Promising solutions to improve the scalability of blockchains include primarily sharding [40] and cross-chain [41] techniques. Sharding

technology is thought to be able to partition the network into different groups (shards) so that the compulsory duplication of communication, data storage, and computation overhead can be avoided for each participating node. These overheads must be incurred by all full nodes in traditional nonsharded blockchains. A cross-chain is a scheme that makes interconnection between blockchains possible. This interoperability is important for individuals and businesses as it helps them exchange values with minimal costs and risks.

Security is the most important issue for blockchain, involving software and hardware as well as protocols and messages required [42]. With the rapid development and wide application of blockchain, criminals may take advantage of the security loopholes to attack users, which exposes blockchain technology to many security threats and challenges. For example, in March 2014, some criminals used a distributed denial of service to attack the Bitcoin trading platform Mt. Gox, which resulted in 850,000 bitcoins stolen from the trading platform and more than US\$450 million lost [43]. In June 2016, the Decentralized Autonomous Organization (DAO), the largest crowdfunding project of blockchain at that time, was attacked and lost about US\$60 million [44].

We will now discuss the security of blockchain from the protocol layer, extension layer, and application layer perspectives. In the protocol layer, the security problems of blockchain include mainly encryption mechanism security (such as private key security), consensus mechanism security (such

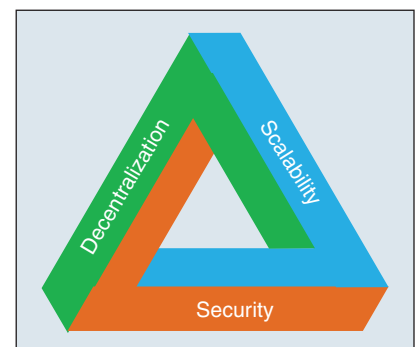


FIGURE 4 – The impossible triangle problem of blockchain.

as the double-spending attack, 51% attack, and coin age attack), and network communication security (such as the eclipse attack, routing attack, border gateway protocol attack, Sybil attack, and balance attack). In the extension layer, the security of blockchain is affected mainly by the vulnerability of the smart contract. Nikolić et al. classified the existing smart contract vulnerabilities as prodigal contracts, greedy contracts, suicidal contracts, and post-mortem contracts [45]. In the application layer, when a user interacts with the blockchain system, an attacker may obtain the user's physical identity or other additional information by means of data mining, which leads to the user's privacy disclosure. The main securities include identity privacy security and transaction privacy security. For a comprehensive survey of blockchain security, please refer to [46].

Decentralization is a key to roll out blockchain applications, which may also compromise blockchain security. Most existing technologies are still centralization oriented. As an example, the Enterprise Operation System (EOS) [47] uses 21 "super nodes" to block out nodes in a certain order, thereby avoiding accounting in a large number of nodes, which would otherwise significantly increase levels in the transaction processing system. However, it has been questioned whether the power is too centralized, which is not conducive to network security. At present, because of the emergence of the Application Specific Integrated Circuit 6 (ASIC6) machine, the PC nodes of ordinary users can hardly participate in the competition of accounting rights. Besides, more than 80% of the computing power is spent on a few mining pools, in which the owners of the mining pools have considerable disbursement power in the Bitcoin world.

Regulatory and Legal Issues

While many countries are actively supporting adoption of the blockchain technology, there are no comprehensive regulations and industry standards yet. Currently, regulations for blockchain are mainly in the finance sector for combating crimes such

as money laundering, extortion, and black-market transactions. For example, a total of US\$761 million in digital currency was stolen by hackers from digital currency exchanges around the world in the first six months of 2018, according to CipherTrace, a U.S. digital currency security company. In comparison, only US\$266 million was lost in 2017. China announced a ban on initial coin offering and shut down all domestic cryptocurrency exchanges in 2019 [48], leading to the challenge of using blockchain without digital currency. Furthermore, the technical rules themselves need to be regulated. The "distrusting" functions of blockchain cannot overcome the "dishonesty" problem of the technology setting itself, and the imbalance of rules wrapped in technology makes the regulation more difficult because of privacy concerns.

There are also significant legal issues in the context of docking and coordination within the existing legal systems. At present, there is no commonly accepted definition of a blockchain in legal systems or an agreement on which attributes are indispensable in each country. Furthermore, most current discussions on smart contracts are focused on how to implement programmable finance and replace intermediaries, ignoring the coordination and compatibility of smart contracts within existing legal systems, especially contracting laws. The ambiguity of semantic expressions and the variability of objective conditions require definitive legal interpretations, which are usually done by a credible third party (a law firm). But smart contracts completely depend on computer languages to stipulate authentication and execution among parties, begging the questions of whether the semantics of the contract terms can accurately express the intentions of the parties and whether the smart contracts can be legally recognized. Furthermore, during the execution of smart contracts, everything needs to comply with the preset code, regardless of the wishes of the parties. A mistake or change would require enormous effort to change the program code. The so-called *smart contract* is not so smart in this instance.

Other Challenges

Blockchain technology is still in its infancy, though it has broad appeal. Another challenge lies in its scalability when many participants are involved. Currently, the transaction chain is long, the centralization efficiency is low, the transparency is not transparent, and trust is lacking. These issues will have to be overcome for blockchain to become an important enabling technology in the emerging digital economy and society.

In terms of technology, the aspects of parallelization, consensus, cross-chain, and channel technologies are very important for the future. There has already been some good progress, including cryptographic security (such as zero-knowledge proofs [49] and ring signatures [50]), consensus mechanisms (such as verifiable random functions [51]), the infrastructure of blockchain (such as multichain, channel technology, and directed acyclic graphs), distributed file systems [such as InterPlanetary File System (IPFS) [52]], and identity management [such as decentralized identifiers (DIDs) [53] and self-sovereign identity (SSI) [54]], among others. For example, IPFS is a peer-to-peer distributed file system that seeks to connect all computing devices with the same system of files, which makes storing and sharing large files more efficient. IPFS provides a high-throughput, content-addressed block storage model with content-addressed hyperlinks. A DID is a new type of identifier that enables verifiable, decentralized digital identity. Compared to typical, federated identifiers, DIDs have been designed so that they may be decoupled from identity providers, centralized registries, and certificate authorities. SSI is a new type of identity management, in which identity and the valuable data generated belong to the users themselves. SSI allows users to manage their own information by themselves, independently of any organizations.

In terms of applications, the current blockchain is still in the 2.0 stage, namely "application + blockchain," which refers to the interactions between the traditional services and blockchain services. Blockchain 3.0 is emerging, in

which all business operations would run on blockchains based on smart contracts in a decentralized manner.

Blockchain for Industrial Electronics

The fast development of blockchain has had a far-reaching impact on many areas, including technological, social, and economic fields. The field of industrial electronics (IE) is no exception. IE tackles the challenges in intelligent and computer control systems, robotics, factory communications and automation, flexible manufacturing, data acquisition and signal processing, vision systems, and power electronics. Key thematic areas, such as power and energy systems, manufacturing systems, robotics and mechatronics, and so on, are being impacted by blockchain, as we now briefly describe.

Power and Energy Systems

The power and energy sector is much affected by blockchain, just as any other sector [1], though things are usually happening a bit more slowly. Power networks are considered to be cyber-physical systems [55] or, if prosumers and community/society are included

in the equation, cyber-physical-social systems (CPSSs). Blockchain technology, according to its promises, has a big future. Figure 5 shows how blockchain can be used in power-sharing applications. A prosumer first enters a contract as a user node through the blockchain network, where the seller's information is made available, while edge nodes equipped with certain computing and storage capabilities serve as miners to maintain the blockchain network. In each block generation cycle, the seller publishes its information of energy surplus to the network, and consumers then bid for the selling energy with successful bidder(s) chosen, and the amount of energy is then allowed for use. The transaction process is automatically completed by the smart contract, where the purchased energy flows from the seller to the buyer through the public grid, and the seller gets a payoff. Finally, the miners in the network package all of the transactions during this period. They then verify the transactions through consensus and generate new data blocks that are then added to the blockchain automatically as secured records.

However, the special features of power and energy CPSSs may mean

that various parts of the blockchain technology need to be made more flexible and less resource intensive as the general blockchain technology is not entirely designed for power and energy systems. For example, there would be stringent requirements of power and energy CPSSs to be dynamically responsive across the three layers of the cyber, physical, and social worlds and also to be robust against intermittent uncertainties, such as renewable energies and electric vehicles. The uptake of blockchain in power and energy CPSSs requires a strong willingness of the community and industry to make it work under the increasingly uncertain and insecure environments as well as in the economic considerations of return of investments to utilities.

For example, currently the need for a "real" (i.e., distributed) blockchain may not always be there since the resource to be managed by the blockchain (e.g., a distribution network) is owned and operated by one central entity, which could just offer a database with an application programming interface or a trusted third party or permissioned ledger [56].

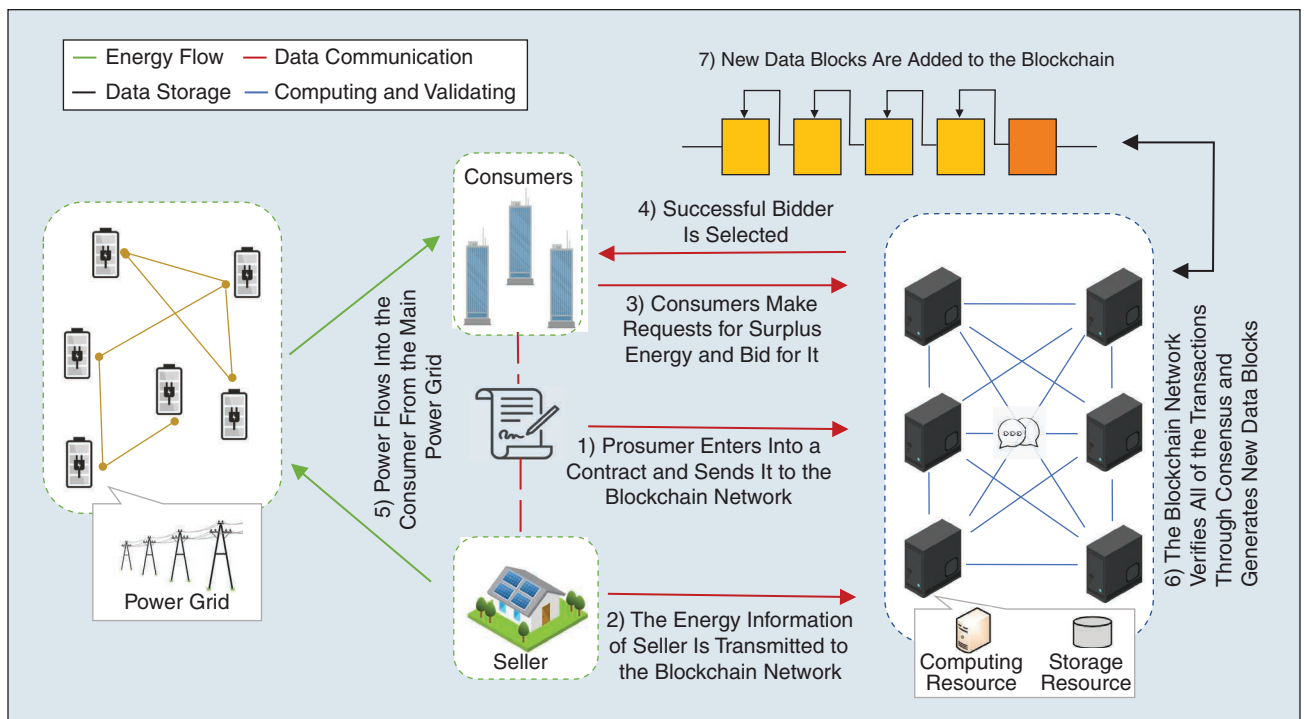


FIGURE 5 – A blockchain helps prosumers to match their needs.

A direct translation of a cryptocurrency into a crypto token for renewable energy amounts bears little complexity to distinguish between green and nongreen energies. This requires the consideration of more energy-oriented distributed storage, cryptography, and consensus algorithm techniques. For example, the Jouliette, a token based on blockchain implemented by a consortium around the Dutch distribution grid company Alliander, supports manual transactions, where customers can trade their Jouliettes, and also automated transactions, for the IoT to participate in this ecosystem. Distributed generation, such as photovoltaics, and intelligent loads, such as heat pumps, can organize themselves based on Jouliette transactions [57]. In China, a company called Energo in Shanghai is using blockchain to deal with trading clean and renewable energy [58], allowing producers to sell energy to consumers securely.

There have been many academic projects on blockchain for energy to improve distributed and local markets, manage distributed energy resources, and tokenize energy or access to energy, and so forth; see [59] and [60] for a list of such projects. Large-scale industrial rollouts of such ideas are, however, scarce. One most prominent example was given by the European transmission system operator (TSO) TenneT [61]. Germany's Sonnen and The Netherlands' Vandebron deliver flexibility services to TenneT to be used in balancing actions. The flexibility comes from Tesla's and household batteries, organized via blockchain, using IBM technology. Encouraged by that, a new and even larger initiative was just launched: the Equigy platform [62]. TenneT (Germany and The Netherlands), Swissgrid (Switzerland), and Terna (Italy) team up to develop a cross-border blockchain platform for energy flexibility operations. TSOs traditionally run their assets by contracting large generation units for a variety of services, such as frequency reserves. Since many of these large fossil-fuel-based units are phased out, TSOs need to acquire these services from other parties in the grid.

Replacing a few large generation units with many small renewable resources has many challenges, one of them being keeping enough flexible reserves for operations. Contracting thousands of resources in a transparent, easy, and flexible way is a perfect case for blockchain.

There are several technical challenges facing the adoption of blockchain in power and energy CPSSs [1]. The dynamical responsiveness of such systems requires the protocols and algorithms to be delay aware, security aware, and privacy aware as well as flexible enough to achieve tradeoffs in reaching consensus under the required latency and throughput. The blockchain network must be scalable as well. Another challenge is the resource constraints of the power and energy CPSSs, which make tamperproof data management difficult, especially considering the multiple types of data models. The security and timely processing of smart contracts are another challenge and may require some parallel processing mechanisms. These and many more activities ultimately lead to the development of standards [63]. While challenges such as transaction throughput can be addressed with the right blockchain design, other challenges, such as secure digital identities of embedded platforms, are equally important in power and energy systems but need to be solved in other ways. On top of that, the intrinsic challenges of a CPSS, such as matching market optima with physical feasibility, are still part of the application and are not "magically" solved by using a blockchain.

Manufacturing Systems in Industry 4.0

The manufacturing sector has witnessed rapid changes, driven by businesses and societies toward mass and extreme customization. New disruptive developments, such as software and hardware, cross-fertilization of concepts, and the integration of information, communication, and control technologies, in traditional industrial environments forge the core of current

networked industrial infrastructures. These include cyberrepresentation of physical assets through digitalization of information across the enterprise, the value stream, and process engineering lifecycle as well as the digital thread from suppliers to customers in the supply chain. The technological, economic, and social impacts are so enormous that the overall process is regarded as the fourth Industrial Revolution, namely, Industry 4.0 [64].

The emerging disruptive technologies are already creating an innovation ecosystem for many industries. They are establishing entirely new markets and platforms for future growth. They are also facilitating the creation of new functionalities based on collaboration of heterogeneous physical systems in the cyberspace able to be exposed and/or consumed as services in a network, enabling continuous improvement of the quality of life for the "citizens in a secure digital society" [65], [66].

In such an Industry 4.0-compliant setting, countless assets, people (humans), machines, and products as well as IT components and systems within the enterprise architecture are able to asynchronously communicate and cooperate directly with each other to perform a set of defined service-oriented business transactions. The production, logistics, and business processes among assets are intelligently networked for a common value creation process. Cooperation through "services" is to be flexibly negotiated and agreed on in the Industry 4.0-conforming communication-information-business network of digitized assets [67].

Central to these is the asset administration shell (AAS), in which blockchain can find its way into the Industry 4.0 context [68]. To help asynchronously interact and handle business transactions, the AAS enables direct communication and cooperation among components (service providers and service consumers) to perform a desired business [69]–[72]. Figure 6 shows an exemplary Industry 4.0-compliant infrastructure, representing three different business processes performed by four AASs,

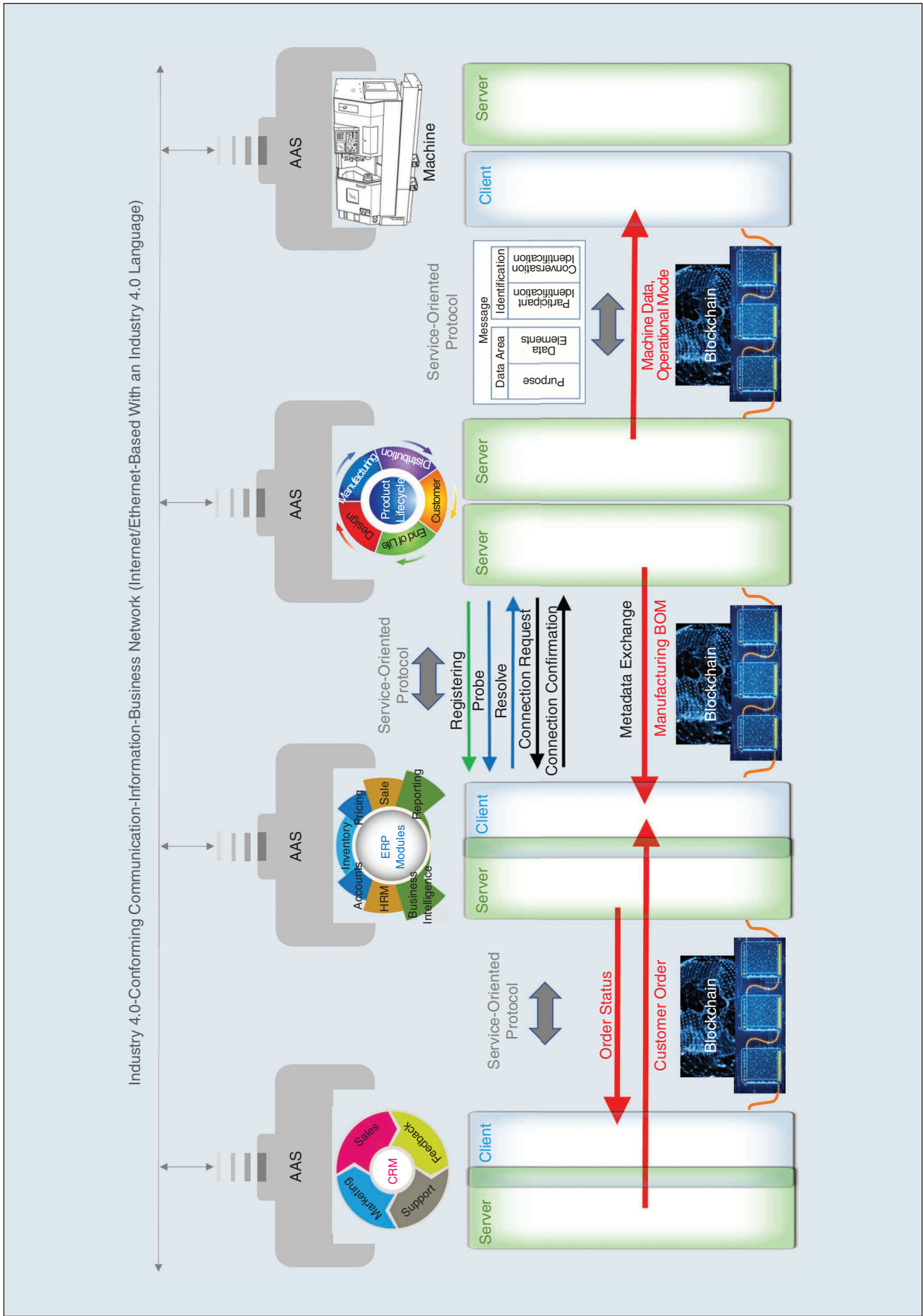


FIGURE 6 – Blockchain in Industry 4.0-compliant systems. CRM: customer relationship management; HRM: human resource management; ERP: enterprise resource planning; BOM bill of materials.

located at very different levels of an enterprise architecture with clearly different functionalities exposed as the Industrial Internet of Services (IIoS) [67]. Integrating blockchain technology within this solution provides reliability and the necessary trust among the AASs, allowing each of them to manage their own blocks and the blockchain-based service/business interaction protocol.

There have been numerous prototype implementations exploiting the features offered by blockchain in the industrial manufacturing sector with a focus on supply chain management. The benefits are enormous, including for example, reducing inventory costs and service times, automating trading and business negotiation processes, enhancing security and authentication, shortening production times, and monetizing ideas and capacities globally. Following the DIN SPEC 91345 (RAMI 4.0) [67] and considering the value stream and lifecycle dimension (International Electrotechnical Commission 62890) [79] as the basis for our example in Figure 6, the AAS-based digitalization aims to seamlessly manage all data, information, and knowledge generated throughout the asset lifecycle to achieve the desired business competitiveness.

The AAS-based approach allows smooth integration and sharing among the digitalized (cooperating) assets [68]. Major requirements, like interoperability, security, trust, and fundamental decentralization of decision-making processes, can easily be achieved by integrating the blockchain technology with the AAS. Essentially, this facilitates the realization of service-legal-agreements among digitalized assets with efficient consensus algorithms. Adequate open but secured information storage and customized blockchain information services, such as machine data or operational modes, can be shared between a digitalized product lifecycle management at the IT level and digitalized machines located at the operation technology levels of the enterprise. On one side, this AAS- and blockchain-based infrastructure not only can process the multisource and heterogeneous

services from the two named assets but can also broadcast the exposed services to the Industry 4.0- and blockchain-conforming network. On the other side, the AAS- and blockchain-based application between IIoS-based business partners allows both vertical as well as horizontal integration, including managed consensus, e.g., for co-design and cocreation of enterprise resource planning (ERP) applications as well as quick and accurate tracking and tracing of manufacturing orders with an AAS-based digitalized customer relationship management (CRM). With the successful development of the proposed solution, service-based interoperability and cooperation among digitalized stakeholders (assets) in the entire value stream and lifecycle are guaranteed.

The Mobility Open Blockchain Initiative (MOBI) and OriginTrail [73] are other examples of blockchain-based solutions. MOBI was founded by automakers such as Renault, Ford, General Motors, and BMW, aiming to “build a vehicle digital identity prototype or car passport that can track and secure a vehicle’s odometer and relevant data on distributed ledgers” [73]. OriginTrail aims to make supply chains more transparent by allowing interested parties to track an item’s origin and process in primary industries, such as vegetable producer Natureta and dairy producer Celeia. Another example is IBM and Maersk (a leading shipping company), who tested blockchain technology in logistics operations [74]. In China, Alibaba established supET [75], a platform for blockchain applications in the industrial Internet. Numerous new use cases are being reported in other industrial manufacturing sectors like Industry 4.0, the Industrial IoT, and so on. This confirms potentials and challenges and also provides an outlook for future research and innovation opportunities to further exploit the advantages of the blockchain technology.

The challenges for the adoption of blockchain in manufacturing systems in Industry 4.0 lie in its role to enhance process optimization (e.g., logistics optimization and product

lifecycle improvisation) and security and authentication (i.e., making parts tamperproof and cross-referencing them, providing identity management) [76]. While dynamical responsiveness is not required as much as it is for the power and energy systems, the complex and diverse nature of manufacturing systems would make scalability and flexibility the prominent issues. The enormous scale of IoT features in Industry 4.0 means there are huge amounts of critical and privacy-sensitive information that need to be protected from cyberattacks. However, because of limited resources, executing security functionalities is difficult to meet these security needs. This requires efficient consensus algorithms that can deal with the problems quickly in a distributed way. Identity management is another issue. The traditional methods of authentication, such as tokens or passwords, may not be useful. Finding a way to create trust among a big network of components/devices that is scalable and secure is a challenge, and this also applies to authorization, authentication, and integrity.

Robotics and Mechatronics and Other IE Areas

Blockchain has implications for many other key IE areas. For example, combining artificial intelligence (AI) with blockchain can improve efficiencies in swarm robotics or autonomous vehicles or can even help Bitcoin mining in a secure, flexible, and autonomous way, similar to that in power and energy systems. Swarm robotics is seen as an area benefiting from the combination of blockchain and AI. A team of autonomous robots work together in a “swarm” to perform tasks or operations; their collective behavior and interactive capabilities need to be robust and highly scalable. This can be enhanced by blockchain through advanced encryption techniques for optimal security for data across shared channels [77]. Blockchain also allows AI models and distributed large data sets to be shared, updated, and trained safely and securely, making wider adoption of AI possible [78].

Many systems and control issues can benefit from blockchain, especially in the multiagent system setting, where individual components cooperate to achieve a common goal quickly and securely in a distributed manner. However, the challenges facing the power and energy systems and manufacturing systems in Industry 4.0 are equally applicable, if not more so, to robotics and mechatronics and other IE areas. The dynamical responsiveness requirements would be more stringent, and resource-light and flexible blockchain platforms would be needed. The future of blockchain is very bright; however, the technological challenges involved in making it work are enormous.

Conclusion

In this article, we introduced the background and basic concepts of blockchain, its key features and technologies, as well as some future challenges and opportunities for blockchain in general. We specifically discussed the impact of blockchain on the future of major focal areas of the IEEE Industrial Electronics Society.

Biographies

Xinghuo Yu (x.yu@rmit.edu.au) earned his Ph.D. degree in control science and engineering from Southeast University, Nanjing, China, in 1988. He is an associate deputy vice chancellor, a distinguished professor, and a vice chancellor's professorial research fellow at the Royal Melbourne Institute of Technology, Melbourne, Victoria 3001, Australia. His research interests include control systems, complex and intelligent systems, smart grids, and energy systems. He has worked extensively in industrial information technologies. He is a Fellow of IEEE and a member of the IEEE Industrial Electronics Society.

Changbing Tang (tangcb@zjnu.edu.cn) earned his Ph.D. degree in electronic engineering from Fudan University, Shanghai, China, in 2014. He received his B.S. and M.S. degrees in mathematics and applied mathematics from Zhejiang Normal University at Jinhua, in 2004 and 2007. He is an associate professor in the Department

of Electronics Information and Engineering, Zhejiang Normal University, Jinhua, 321004, China. His research interests include game theory, blockchain and its applications, networks, and distributed optimization. He was an Academician Pairing Training Program for Young Talents of Zhejiang Province in 2019. He is a Member of IEEE.

Peter Palensky (palensky@ieee.org) earned his Ph.D. degree from the Vienna University of Technology, Austria, in 2001. He is currently a full professor for intelligent electric power grids at the Delft University of Technology (TU Delft) and the scientific director of TU Delft's PowerWeb Institute, Delft, 2628CD, The Netherlands. His research interests include the digital transformation of power systems. He is a Senior Member of IEEE and a member of the IEEE Industrial Electronics Society.

Armando Walter Colombo (aw.colombo@ieee.org) earned his Ph.D. degree in engineering from the University of Erlangen–Nuremberg, Germany, in 1998. He is a full professor in the Faculty of Engineering and director of the Institute for Industrial Informatics, Automation, and Robotics at the University of Applied Sciences Emden/Leer, Emden, D-26723, Germany. From 2001 to 2018, he was the director for Innovation Projects and Edison-Level-2 Group Senior Expert at Schneider Electric. His research interests include industrial-cyber-physical systems, Industry 4.0, the Internet-of-Things, and the Internet of Services. He is member of the IEEE Industrial Electronics Society (IES) Administrative Committee, chair of the IES Fellows Committee, IES representative to the IEEE Systems Council, and the co-editor-in-chief of *IEEE Open Journal of the Industrial Electronics Society*. He is a Fellow of IEEE and a member of the IEEE Industrial Electronics Society.

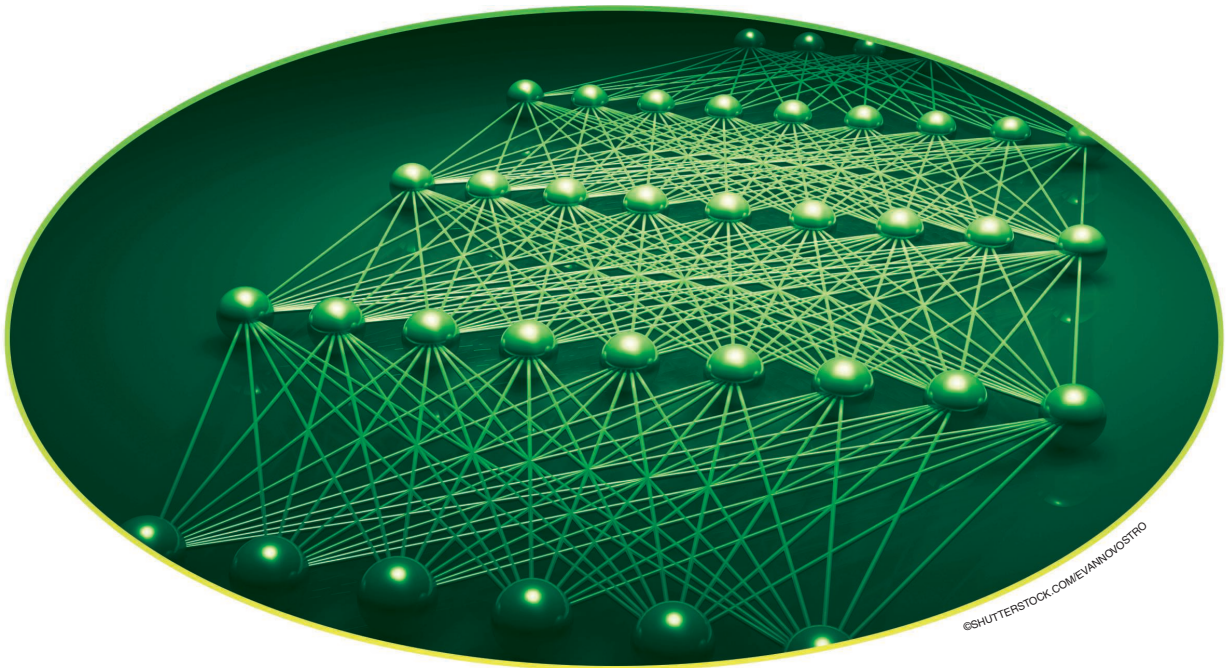
References

- [1] N. U. Hassan, C. Yuen, and D. Niyato, "Blockchain technologies for smart energy systems: Fundamentals, challenges, and solutions," *IEEE Ind. Electron. Mag.*, vol. 13, no. 4, pp. 106–118, Dec. 2019.
- [2] T. Aste, P. Tasca, and T. D. Matteo, "Blockchain technologies: The foreseeable impact on society and industry," *Computer*, vol. 50, no. 9, pp. 18–28, 2017. doi: 10.1109/MC.2017.3571064.

- [3] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Bitcoin, White Paper, 2008. <https://bitcoin.org/bitcoin.pdf> (accessed 15 June 2020).
- [4] P. K. Sharma, S. Singh, Y. S. Jeong, and J. H. Park, "DistBlockNet: A distributed blockchains-based secure SDN architecture for IoT networks," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 78–85, Sept. 2017. doi: 10.1109/MCOM.2017.1700041.
- [5] V. Buterin, "Ethereum: A next-generation smart contract and decentralized application platform," GitHub, Inc., San Francisco, 2013. <http://ethereum.org/ethereum.html> (accessed 15 June 2020).
- [6] L. Lu et al., "Pseudo trust: Zero-knowledge authentication in anonymous P2Ps," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 10, pp. 1325–1337, Oct. 2008.
- [7] D. Hopwood, S. Bowe, T. Hornby, and N. Wilcox, "Zcash protocol specification," Version 2020.1.2, ZeroCoin Electric Coin Co., Oakland, CA, Tech. Rep., Mar. 20, 2020.
- [8] E. Ben-Sasson, A. Chiesa, E. Tromer, and M. Virza, "Succinct non-interactive zero knowledge for a von neumann architecture," in *Proc. 2014 USENIX Security Symp.*, San Diego, CA, Aug. 20–22, 2014, pp. 781–796.
- [9] Y. Yuan, X. C. Ni, S. Zeng and F. Y. Wang, "Blockchain consensus algorithms: The state of the art and future trends," *Acta Automat. Sinica (in Chinese)*, vol. 44, no. 11, pp. 2011–2022, 2018.
- [10] L. Lamport, "The part-time parliament," *ACM Trans. Comput. Syst.*, vol. 16, no. 2, pp. 133–169, 1998. doi: 10.1145/279227.279229.
- [11] D. Ongaro and J. Ousterhout, "In search of an understandable consensus algorithm," in *Proc. USENIX Annu. Tech. Conf.*, Philadelphia, June 19–20, 2014, pp. 305–320.
- [12] M. Castro and B. Liskov, "Practical Byzantine fault tolerance," in *Proc. Operating Syst. Des. Implement.*, vol. 99, pp. 173–186, Feb. 1999.
- [13] D. Mazieres, "The stellar consensus protocol: A federated model for internet-level consensus." Stellar. <https://www.stellar.org/papers/stellar-consensus-protocol.pdf> (accessed Sept. 20, 2020).
- [14] Y. Gilad, R. Hemo, S. Micali, G. Vlachos, and N. Zeldovich, "Algorand: Scaling byzantine agreements for cryptocurrencies." IACR. <http://eprint.iacr.org/2017/454> (accessed Sept. 20, 2020).
- [15] R. Pass and E. Shi. "The sleepy model of consensus." IACR. <https://eprint.iacr.org/2016/918.pdf> (accessed Sept. 20, 2020).
- [16] I. Eyal, A. E. Gencer, E. G. Sirer, and R. V. Renesse, "Bitcoin-NG: A scalable blockchain protocol," in *Proc. 13th USENIX Conf. Netw. Syst. Des. Implementation*, 2016, pp. 45–59.
- [17] E. Kokoris-Kogias, P. Jovanovic, N. Gailly, I. Khoffi, L. Gasser, and B. Ford, "Enhancing bitcoin security and performance with strong consistency via collective signing," in *Proc. 25th USENIX Security Symp.*, 2016, pp. 279–296.
- [18] J. P. Buntinx, "What is proof of elapsed time?" Accessed Sept. 20, 2020. [Online]. Available: <https://www.investopedia.com/terms/p/proof-of-elapsed-time-cryptocurrency.asp>
- [19] M. Milutinovic, W. He, H. Wu, and M. Kanwal, "Proof of luck: An efficient blockchain consensus protocol." IACR. <https://eprint.iacr.org/2017/249.pdf> (accessed Sept. 20, 2020).
- [20] G. Ateniese, I. Bonacina, A. Faonio, and N. Galeasi, "Proofs of space: When space is of the essence," in *Proc. 9th Int. Conf. Security Cryptogr. Netw.*, 2014, pp. 538–557.
- [21] M. Ball, A. Rosen, M. Sabin, and P. N. Vasudevan, "Proofs of useful work." Allquantor. <https://allquantor.at/blockchain-bib/pdf/ball2017proofs.pdf> (accessed Sept. 20, 2020).
- [22] S. King and S. Nadal, "PPcoin: Peer-to-peer crypto-currency with proof-of-stake," Self-published paper, vol. 19, Aug. 2012.
- [23] D. Larimer, F. Schuh, and D. Larimer, "BitShares 2.0: Financial smart contract platform." BitShares. <https://www.weusecoins.com/assets/>

- pdf/library/Bitshares%20Financial%20Platform.pdf (accessed Sept. 20, 2020)
- [24] J. Kwon, "Tendermint: Consensus without mining." Tendermint. <https://tendermint.com/static/docs/ten-dermint.pdf> (accessed Sept. 20, 2020)
- [25] "Ethereum's Casper protocol explained in simple terms." Finder. <https://www.finder.com/ethereum-casper> (accessed Sept. 20, 2020)
- [26] A. Miller, A. Juels, E. Shi, B. Parno and J. Katz, "Permacoin: Repurposing Bitcoin work for long-term data preservation," in *Proc. IEEE Symp. Security Privacy*, 2014, vol. 1, pp. 475–490.
- [27] L. Ren, "Proof of stake velocity: Building the social currency of the digital age." <https://www.cryptoground.com/storage/files/1528454215-cannacoin.pdf> (accessed Sept. 20, 2020)
- [28] "Proof of burn." Bitcoin. https://en.bitcoin.it/wiki/Proof_of_burn (accessed Sept. 20, 2020)
- [29] I. Bentov, C. Lee, A. Mizrahi, and M. Rosenfeld, "Proof of activity: Extending Bitcoins proof of work via proof of stake." IACR. <http://eprint.iacr.org/2014/452> (accessed Sept. 20, 2020).
- [30] Y. Xiao, N. Zhang, W. Lou, and Y. T. Hou, "A survey of distributed consensus protocols for blockchain networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1432–1465, 2020. doi: 10.1109/COMST.2020.2969706.
- [31] N. Szabo, "Smart contracts." https://pipiwiki.com/wiki/Agoric_computing (accessed June 15, 2020).
- [32] T. T. A. Dinh, R. Liu, M. H. Zhang, G. Chen, B. C. Ooi, and J. Wang, "Untangling blockchain: A data processing view of blockchain systems," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1366–1385, 2018. doi: 10.1109/TKDE.2017.2781227.
- [33] Hyperledger Project. Accessed: June 15, 2020. [Online]. Available: <https://www.hyperledger.org/>
- [34] S. Wang, Y. Yuan, X. Wang, J. J. Li, R. Qin, and F. Y. Wang, "An overview of smart contract: Architecture, applications, and future trends," in *Proc. IEEE Intell. Veh. Symp. (IV)*, Changshu, China, June 26–30, 2018, pp. 108–113. doi: 10.1109/IVS.2018.8500488.
- [35] "The energyweb chain: Accelerating the energy transition with an open-source, decentralized blockchain platform." Energy Web Foundation, Zug, Switzerland. Accessed: Sept. 20, 2020. [Online]. Available: <https://energyweb.org/wp-content/uploads/2018/10/EWF-Paper-TheEnergyWebChain-v1-201810-FINAL.pdf>
- [36] "Initial coin offering." Solar Bankers, Prague, Czech Republic, White Paper. [Online]. Available: https://solarbankers.com/wp-content/uploads/2017/10/SB-White-Paper_version2.pdf (accessed Sept. 20, 2020).
- [37] J. Steiner and J. Baker, "Blockchain: The solution for transparency in product supply chains," Project Provenance Ltd., London, White Paper, 2015. [Online]. Available: <https://www.provenance.org/whitepaper> (accessed Sept. 20, 2020)
- [38] G. Prisco, "The blockchain for healthcare: Gem launches Gem Health Network with Philips Blockchain Lab." BitCoin Magazine. [Online]. Available: <https://bitcoinmagazine.com/articles/the-blockchain-for-healthcare-gem-launches-gem-health-network-with-philips-blockchain-lab-1461674938/> (accessed Sept. 20, 2020)
- [39] Genesis of Things Project. Accessed: Sept. 20, 2020. [Online]. Available: <http://www.genesisofthings.com/>
- [40] G. Yu, X. Wang, K. Yu, W. Ni, J. A. Zhang, and R. P. Liu, "Survey: Sharding in blockchains," *IEEE Access*, vol. 8, pp. 14,155–14,181, Jan. 2020. doi: 10.1109/ACCESS.2020.2965147.
- [41] A. Zamyatin et al., "SoK: Communication across distributed ledgers," Imperial College London, London, UK, IACR Cryptology ePrint Archive, 2019: 1128, Tech. Rep., 2019.
- [42] T. Salman, M. Zolanvari, A. Erbad, R. Jain, and M. Samaka, "Security services using blockchains: A state of the art survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 858–880, 2019. doi: 10.1109/COMST.2018.2863956.
- [43] A. Jake and S. Nathalie-Kyoko. "Behind the biggest Bitcoin heist in history: Inside the implosion of mt.gox." The Daily Beast. <https://www.thedailybeast.com/behind-the-biggest-bitcoin-heist-in-history-inside-the-implosion-of-mt-gox> (accessed Sept. 20, 2020).
- [44] V. Buterin. "Critical update re: Dao vulnerability." <https://blog.ethereum.org/2016/06/17/critical-update-re-dao-vulnerability/> (accessed Sept. 20, 2020)
- [45] I. Nikolić, A. Kolluri, I. Sergey, P. Saxena, and A. Hobor, "Finding the greedy, prodigal, and suicidal contracts at scale," in *Proc. 34th Annu. Comput. Security Appl. Conf.*, San Juan, PR, Dec. 2018, pp. 653–663.
- [46] D. Dasgupta, J. M. Shreiner, and K. D. Gupta, "A survey of blockchain from security perspective," *J. Banking Financial Technol.*, vol. 3, no. 1, pp. 1–17, 2019. doi: 10.1007/s42786-018-00002-6.
- [47] "EOSIO/eos: An open source smart contract platform," GitHub, San Francisco, Oct. 2, 2018. <https://github.com/pyun/eos> (accessed Mar. 28, 2021).
- [48] W. Song, Y. Li, and D. Yang, "Research on the application of blockchain in the energy power industry in China," *IOS J. Phys. Conf. Ser.*, vol. 1176, no. 4, p. 042079, 2019.
- [49] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," in *Proc. IEEE Symp. Security Privacy (SP)*, San Jose, CA, May 22–26, 2016, pp. 893–858.
- [50] S. F. Sun, M. H. Au, J. K. Liu, and T. H. Yuen, "RingCT 2.0: A compact accumulator-based (linkable ring signature) protocol for blockchain cryptocurrency Monero," in *Computer Security-ESORICS*. Cham: Springer-Verlag, 2017, vol. 10493, pp. 456–474.
- [51] W. T. Li, S. Andreina, J. M. Bohli, and G. Karame, "Securing proof-of-stake blockchain protocols," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Cham: Springer-Verlag, 2017, vol. 10436, pp. 297–315.
- [52] H. W. Huang, J. R. Lin, B. C. Zheng, Z. B. Zheng, and J. Bian, "When blockchain meets distributed file systems: An overview, challenges, and open issues," *IEEE Access*, vol. 8, pp. 50,574–50,586, Mar. 2020. doi: 10.1109/ACCESS.2020.2979881.
- [53] "Decentralized Identifiers (DIDs): Core architecture, data model, and representations." GitHub, San Francisco. <https://w3c.github.io/did-core/> (accessed Sept. 20, 2020).
- [54] "A gentle introduction to self-sovereign identity." Bits on Blocks. <https://bitsonblocks.net/2017/05/17/gentle-introduction-self-sovereign-identity/> (accessed Sept. 20, 2020)
- [55] X. Yu and Y. Xue, "Smart Grids: A cyber-physical systems perspective," *Proc. IEEE*, vol. 104, no. 5, pp. 1058–1070, 2016. doi: 10.1109/JPROC.2015.2503119.
- [56] M. E. Peck, "Do you need a blockchain?" *IEEE Spectrum*, vol. 54, no. 10, pp. 38–60, Oct. 2017. doi: 10.1109/MSPEC.2017.8048838.
- [57] ETIP SNET WG4. "Digitalization of the electricity system and customer participation," De Ceuvel, Amsterdam, The Netherlands, Tech. Position Paper WG4, Sept. 2018.
- [58] "Starting from the micro power grid, Energo tries to build a decentralized energy transaction system using blockchain." Energo Labs, Shanghai, China, http://www.8btc.com/energo-labs-blockchain_in_Chinese (accessed Jan. 25, 2018)
- [59] A. S. Musleh, G. Yao, and S. M. Mueen, "Blockchain applications in smart grid—review and frameworks," *IEEE Access*, vol. 7, pp. 86,746–86,757, June 2019. doi: 10.1109/ACCESS.2019.2920682.
- [60] S. Johanning and T. Bruckner, "Blockchain-based peer-to-peer energy trade: A critical review of disruptive potential," in *Proc. 16th Int. Conf. European Energy Market (EEM)*, Ljubljana, Slovenia, 2019, pp. 1–8. doi: 10.1109/EEM.2019.8916268.
- [61] K. Döppenbecker. "Undertaking energy transition interview with TenneT's Digital Transformation Lead René Kerkmeester." TenneT, 2019. https://www.tennet.eu/fileadmin/user_upload/ArtikelTenneT.pdf (accessed June 15, 2020)
- [62] "Equigy platform gives European consumers access to tomorrow's sustainable energy market," TenneT, European, TenneT Press Release 23. Apr. 2020. Accessed Mar. 28, 2021. [Online]. Available: <https://www.tennet.eu/#&panel1-1>
- [63] *Standard for Blockchain in Energy*, IEEE Standard P2418.5, 2018
- [64] A. W. Colombo, S. Karnouskos, O. Kaynak, Y. Shi, and S. Yin, "Industrial cyber-physical systems: A backbone of the fourth industrial revolution," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 6–16, 2017. doi: 10.1109/MIE.2017.2648857.
- [65] Federal Ministry of Education and Research (BMBF), Germany, "Innovations for the production, services and work of tomorrow (in German)," in *The New Hightech Strategy, Innovations for Germany*, 2014. [Online]. Available: <https://www.bmbf.de/de/innovationen-fuer-die-produktion-dienstleistung-und-arbeit-von-morgen-599.html>
- [66] ZVEI, Zentralverband Elektrotechnik- und Elektronikindustrie e.V., Safety and Security in Industry 4.0. <https://www.dke.de/resource/blob/1624282/f6372e8c85ee20491f6b7b967203ccbc/safety-security-im-bereich-industrie-4-0-prof-wegener-data.pdf> (accessed May 24, 2020)
- [67] *DIN SPEC 91335 RAM14.0*. [Online]. Available: <https://dx.doi.org/10.31030/2436156> (accessed June 15, 2020).
- [68] "Platform Industry 4.0 and ZVEI, details of the asset administration shell." https://www.platform-i40.de/P140/Redaktion/EN/Downloads/Publikation/Details-of-the-Asset-Administration-Shell-Part1.pdf?_blob=publicationFile&v=5 (accessed May 22, 2020).
- [69] "Blockchain Technology In Industry 4.0," BitDeal. <https://www.bitdeal.net/blockchain-in-industry-4-0> (accessed May 25, 2020)
- [70] R. Rosa Righi, A. M. Alberti, and M. Singh, Eds., *Blockchain Technology for Industry 4.0*. Springer Nature, Singapore Pte Ltd., 2020.
- [71] Blockchain – eine Technologie mit disruptivem Charakter (in German). https://www.vditz.de/fileadmin/media/bekanntmachungen/documents/vdi_publication_blockchain_RZ_web_neu.pdf (accessed May 20, 2020)
- [72] "Blockchain and Industry 4.0." Capgemini. <https://www.capgemini.com/au-en/wp-content/uploads/sites/9/2018/10/Blockchain-and-Industry-4.0.pdf> (accessed May 20, 2020)
- [73] "What is blockchain technology and how is it changing the manufacturing industry?" The Manufacturer. <https://www.themanufacturer.com/articles/blockchain-technology-changing-manufacturing-industry/> (accessed June 10, 2020)
- [74] Application of Blockchain in manufacturing industry." Blockchain Expert. <https://www.blockchainexpert.uk/blog/application-of-blockchain-in-manufacturing> (accessed June 10, 2020)
- [75] "Application of blockchain in industrial Internet by Ali Cloud." CQVIP. <http://www.cqvip.com/QK/80675A/201822/7000940533.html>, in Chinese (accessed June 12, 2020).
- [76] A. Mushtaq, I.U. Haq, "Implications of Blockchain in industry 4.0," in *Proc. Int. Conf. Eng. Emerg. Technol.*, Lahore, Pakistan, Feb. 21, 2019, pp. 2409–2493.
- [77] "How Blockchain and AI can help robotics technologies." Robotics Business Review. <https://www.roboticsbusinessreview.com/ai/how-blockchain-and-ai-can-help-robotics-technologies/> (accessed June 9, 2020).
- [78] R. Shroff, "When Blockchain meets Artificial Intelligence." Medium. <https://medium.com/swlh/when-blockchain-meets-artificial-intelligence-e448968d0482> (accessed June 12, 2020).
- [79] *IEC 62890:2020*, IEC Webstore. [Online]. Available: <https://webstore.iec.ch/publication/30583>





Dependency-Aware Tensor Scheduler for Industrial AI Applications

*Dymem—An Aggressive
Data-Swapping Policy for
Training Nonlinear Deep
Neural Networks*

WEI RANG,
DONGLIN YANG, and
DAZHAO CHENG

Digital Object Identifier 10.1109/MIE.2021.3084546
Date of current version: 13 August 2021

Artificial intelligence (AI) applications based on deep neural networks (DNNs) have been widely applied in industry, e.g., in natural language processing and computer vision, among other fields. Researchers and industry practitioners typically use GPUs to train complex, hundred-layer deep learning (DL) networks. However, as the networks become wider and deeper, the limited GPU memory becomes a significant bottleneck, restricting the size of the networks to be trained. In the training of DNN-based AI applications, the intermediate layer outputs are the major contributors to the memory footprint. Various data-swapping techniques, such as the offloading and prefetching of intermediate layer outputs, are proposed to overcome the GPU memory shortage by utilizing the CPU dynamic random-access memory (DRAM) as an external buffer for the GPU.

However, we find that the layer-by-layer asynchronous approach cannot be effectively applied to nonlinear DNNs because of the unbalanced overlap

We present a memory-efficient graph analysis to construct an execution order for nonlinear networks and propose a dynamic offloading/prefetching strategy to maximize the performance and usage of bandwidth.

between communication and computation. Based on our observations, we propose and design a novel aggressive data-swapping policy for training nonlinear DNNs. Instead of using the popular layer-by-layer strategy, Dymem adopts a more greedy asynchronous solution to maximize the DRAM bandwidth, balance memory usage, and improve performance. We implement and evaluate Dymem based on several linear and nonlinear networks. Compared with the other two representative approaches, Dymem improves the end-to-end throughput for nonlinear networks by up to 42%.

Introduction

DL has achieved great success in various domains, such as image classification [1], natural language processing [2], object detection [3], speech recognition [4], and so on. Obtaining accurate DL models is a computation-intensive process that requires large amounts of data and a substantial computing capacity. Previous studies have shown that the use of wider and deeper DNNs can significantly increase the model performance. In particular, various nonlinear neural network architectures [5], [6] have been proposed to further improve the quality of model training, especially for image recognition tasks.

However, the limited size of the GPU DRAM has been a major bottleneck for researchers in exploring deeper and wider DNNs for better generalization performance. For example, it is reported that a linear network, Visual Geometry Group (VGG)-16 [7], which is composed of 16 computation-intensive convolution layers, requests a total of 28 GB of memory usage for setting the batch size to 256 [8]. Another representative nonlinear DNN, Inception-v4 [6],

requests up to 45 GB of memory to keep the entire network on the GPU in training. Unfortunately, the largest GPU memory capacity offered by the commercial NVIDIA Volta architecture so far is 32 GB [9]. Apparently, the memory shortage of GPUs limits DL practitioners from deploying wider and deeper DNNs.

Many approaches have been introduced to reduce the GPU memory footprint of DNN training. However, these solutions have their limitations. For example, most prior works propose reducing the model size to lessen the memory footprint. However, this strategy either provides low memory footprint reduction or results in a loss in training accuracy [10], [11]. First, in DNN training, parameter weights account for only a small fraction of the total memory footprint; intermediate feature maps are the primary contributor to the significant increase in the memory footprint. Some recent efforts [12], [13] focus on reducing the intermediate result sizes with model layer fusion. These intermediate values should be stored/stashed in the forward pass so that they can be reused later in the backward pass. Additionally, approaches that apply lower precision computations for DNN training, mostly in the context of application-specified integrated circuits and field-programmable gate arrays, either do not target feature maps (and thus achieve low memory footprint reduction) or result in reduced training accuracy [14]. Memory compression [15], [16] and data encoding [17] are other approaches to reduce the GPU memory requirement for training via channel/filter pruning to reduce the intermediate result sizes; however, they introduce a high-performance overhead. State-of-the-art memory footprint reduction methods focus on

training swap data structures back and forth between the CPU and GPU memories [18], [19], but the existing swapping approaches are inefficient in reducing the memory footprint.

Inspired by the fact that DNN training follows a series of layer-wise computations, virtualized DNN (vDNN) [18] and SuperNeurons [19] propose to virtualize the memory usage of DNNs across both the GPU and CPU memories [20]. Considering that a GPU can process only one layer at any given time, it is not necessary to overprovision the memory allocation to accommodate the entire neural network on the GPU. vDNN and SuperNeurons release or move data structures, particularly the intermediate feature maps, between the CPU and GPU, by exploiting the interlayer dependencies and reuse patterns of DNNs.

However, those techniques are not well tuned to address the dependency and memory variations in nonlinear networks. First, the core idea of vDNN and SuperNeurons is to offload the data of one network layer when it is not required in the near future and can be released from GPU DRAM, saving space for other layers. The offloaded data are brought back to the GPU when needed in the backward pass. Offloading data from GPU DRAM can achieve optimal performance if the communication between the CPU and GPU can be well hidden by computation to utilize the bandwidth. However, we observe that offloading data structures from the GPU to the CPU or prefetching data back to the GPU from the CPU layer by layer brings significant inefficiency. For example, the transfer time can be longer or shorter than the forward computation time across layers, with the result that the communication can only be partially overlapped with the computation. Usually, the communication time is much longer than the computation time in pooling layers. By contrast, the computation time is usually much longer than the communication time in convolution layers, as illustrated in Figure 1. Specifically, for nonlinear blocks, where there are join or fork connections, more benefit can be

earned by aggressively advancing the computation in the forward pass or the prefetching operations in the backward pass.

In this article, we propose and design an aggressive data-swapping policy (called *Dymem*) for training nonlinear DNNs. Instead of using the popular layer-by-layer strategy, *Dymem* adopts a more greedy asynchronous solution to maximize the DRAM bandwidth, balance memory usage, and improve performance. In a nutshell, we make the following technical contributions. We empirically study the inefficiency of the memory-swapping policy in existing works, which motivates the need for the design of dependency-aware memory management for nonlinear networks. We present a memory-efficient graph analysis to construct an execution order for nonlinear networks and propose a dynamic offloading/prefetching strategy to maximize the performance and usage of bandwidth. We implement *Dymem* to evaluate several nonlinear DNNs and perform comprehensive evaluations with various depths. *Dymem* improves the end-to-end throughput for nonlinear networks by up to 42%, when compared to ResNet-50, running with batch size 100.

Motivation

Several memory-reduction techniques [18], [19] have been proposed to address the problem of limited GPU resident memory. They mainly focus on offloading selected layers to the preallocated pinned CPU memory and prefetching these data back to the GPU when required. Typically, in the forward pass, the input feature maps X from the preceding layer can be offloaded to the CPU memory if there is no more dependency, after which these data can be released from the GPU memory. As shown in Figure 2(a),

the runtime uses two independent processes to complete the computation and communication, which enables the CPU-to-GPU data transfers to overlap with the computation asynchronously. In the backward pass, for those offloaded layers, the runtime should bring the feature tensors X back to the GPU DRAM before the backward dependent layer starts its propagation. The prefetch operation for layer m can be overlapped with the computation of layer l in the backward pass, where $l > m$. Ideally, this design can maximize the performance

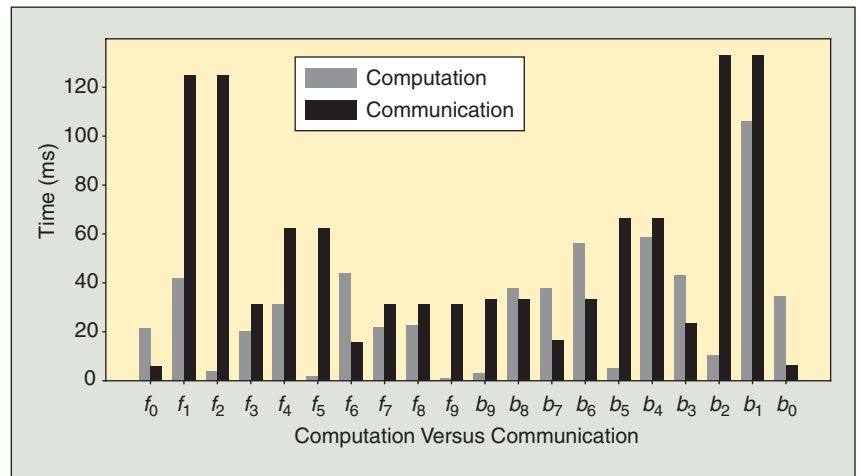


FIGURE 1 – Computation and communication times for different layers in the forward and backward passes.

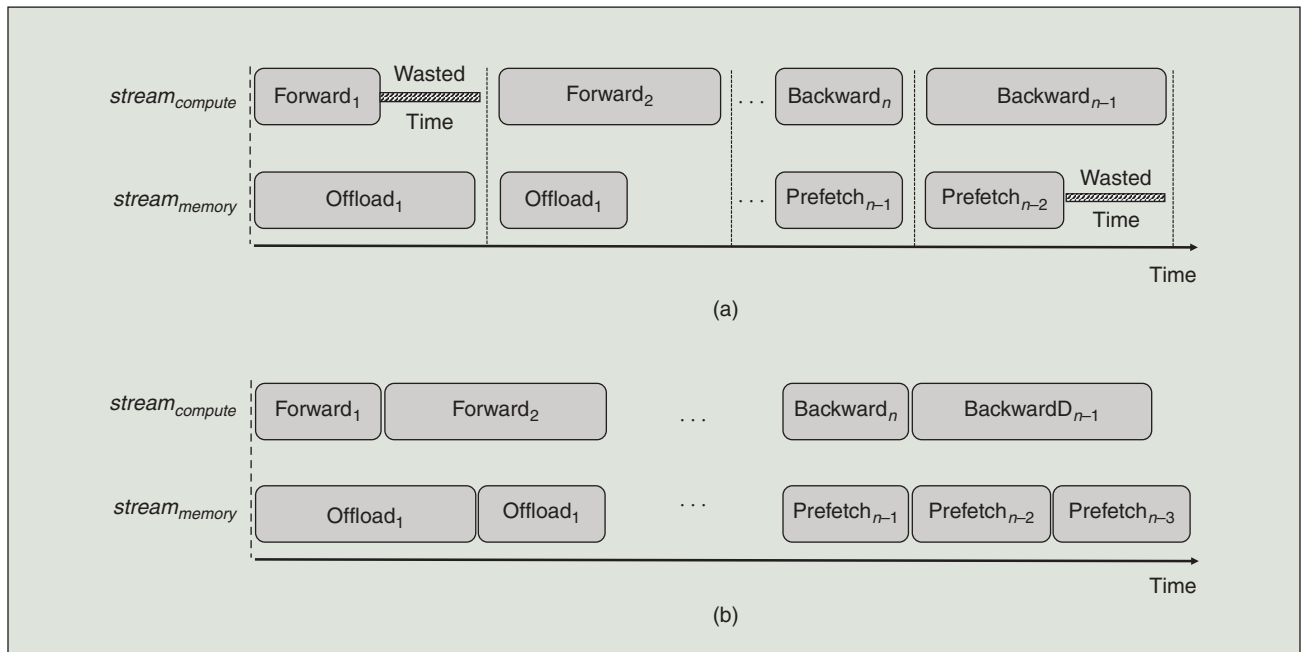


FIGURE 2 – Synchronization with and without a barrier. (a) Illustrates the default approach that uses two independent processes to complete the computation and communication; (b) shows a more aggressive synchronization approach without a barrier.

Dymem is a host-side runtime that interfaces with GPUs to dynamically move, allocate, and release data structures in terms of tensors.

by hiding the communication by the computation time [21]. To ensure the safety of the parallel streams, a synchronization is enforced at the end of each layer, which means the communication and computation streams cannot advance each other in both the backward and forward passes. However, this design largely depends on the communication/computation ratio. Indeed, it works well for a linear network, e.g., VGG-16 [7], which consists of 12 computation-intensive convolution layers. But this strategy can be extremely inefficient for nonlinear networks since the offload/prefetch operation can be well hidden as convolution layers that require a longer computation time than communication time, such as the case in Figure 2(a).

To demonstrate this, Figure 1 presents the communication and computation times for 10 layers in both the forward and backward passes of GoogLeNet [5]. The result shows that f_5 's computation time is much lower compared to the offloading time, while the next layer's forward computation time is higher than the communication time. If f_5 's input is decided to be offloaded, then it is not necessary to wait for the offloading of f_5 before starting the next layer's computation. Thus, as shown in Figure 2(b), a more aggressive synchronization approach without a barrier should be considered. Similarly, in the backward pass, when the layer b_7 is being propagated, the prefetching operation for layer b_5 can be initiated after the transfer of layer b_6 is finished. Secondly, the backward pass of each layer requires more memory space for gradients' input and output maps besides input and output feature maps compared with forward. Hence, in the forward pass, the peak memory requirement is not higher than that in the backward pass if this aggressive strategy is adopted to advance computation.

However, attention should be paid to the backward pass because aggressively prefetching data does not always bring benefits. These observations motivate us to propose a more efficient memory-scheduling strategy to balance memory cost and performance.

System Design and Implementation

The design objective of our dynamic memory manager, Dymem, is to automatically manage the memory usage of DNNs while minimizing the overhead and maximizing the reduction of the memory load. Dymem is a host-side runtime that interfaces with GPUs to dynamically move, allocate, and release data structures in terms of tensors [22]. Dymem is implemented based on the Compute Unified Device Architecture (CUDA) DNN (cuDNN) and can be deployed on any operating systems that support the cuDNN. Moreover, Dymem is orthogonal to any other parallel technologies; when it is applied on a cluster environment for better data parallel purposes, each node can use the memory-swapping policy we propose with Dymem. This issue is also discussed in Bai et al. [23].

In this section, we first discuss the limitations of data swapping in GPU memory management for nonlinear networks. Then we perform a graph analysis and construct a memory-efficient execution sequence. We finally design a dependency-aware tensor scheduler to handle the prefetch/offload operations and move tensors between the CPU and GPU to overcome the GPU memory shortage. Dymem releases or moves data structures, particularly the intermediate feature maps, between the CPU and GPU, by exploiting the interlayer dependencies and reuse patterns of the DNNs. Instead of using the popular layer-by-layer strategy, Dymem adopts a more greedy

asynchronous solution to maximize the DRAM bandwidth, balance memory usage, and improve performance.

The Execution Graph Construction

Given a nonlinear network, we need a memory-efficient approach to set up the execution order. Since the cuDNN [24] implements DL primitives at layer granularity, we use tensors as the basic scheduling unit. For basic networks, during the forward propagation, the results from $layer_{n-1}$ can be applied as the input for $layer_n$. The computation flow can be regarded as a sequential process. Only when the preceding layer is finished can it initiate the next layer's computation. This chain rule is similarly applied in the backward pass but in a reversed order. For networks with nonlinear blocks, there are nonlinearities, such as one-to-many (fork) and many-to-one (join) connections. A depth-first search (DFS) algorithm is used to decide the execution sequences for these nonlinear dependencies, as shown in Algorithm 1. Whenever there is a fork connection, the DFS algorithm is applied to explore all of the executable layers until it reaches the join connection in the nonlinear blocks, as shown in lines 7 and 8.

Figure 3(a) shows the schema for the Inception-A blocks in the Inception-v4 network. The detailed execution order obtained by DFS is demonstrated in Figure 3(b). In this example, the Inception block should be propagated in four branches in both the forward and backward passes. In the forward pass, $l_0 \rightarrow \{l_1, l_2, l_3, l_4\}$ represents the idea that output feature maps from layer l_0 should reside in the GPU memory until layers $l_1, l_2, l_3,$ and l_4 are executed due to their dependencies. Similarly, during the backward pass, in the branch $l_8 \rightarrow l_7 \rightarrow l_4$, when l_8 is being executed, layer l_4 should be prefetched from the CPU memory asynchronously based on DFS. The reason why DFS should be applied to construct the execution graph lies in two properties. First, DFS requires less memory space to reach the join connection node in the nonlinear blocks when exploring the traversal path. For example, the branch $l_4 \rightarrow l_7 \rightarrow l_8$ illustrates a simple

dependency in which the corresponding data for those memory-intensive convolution layers can be released sequentially from the GPU memory. Second, inside those nonlinear blocks, e.g., the residual block and Inception grid, most layers are computation-intensive convolution layers. The DFS algorithm can mostly serialize the sequences of convolution layers in each branch.

Memory Offload

After obtaining the execution graph, Dymem adaptively manages the offloading and release operations on the tensors to effectively improve the overlap ratio between communication and computation. As shown in Figure 2(b), we employ two separate *cudaStreams* to transfer tensors in/out of external memory asynchronously. *stream_{compute}* interfaces to *cuDNN* to handle all of the computations in the forward and backward passes. *stream_{memory}* is responsible for the tensor placement, movement, allocation, and deallocation.

During forward propagation, if *layer_n* is available for offloading, Dymem first allocates a pinned memory region in the host via *cudaMallocHost()*; then *stream_{memory}* can asynchronously swap feature maps from this layer via a nonblocking memory transfer. When the asynchronous offload is completed, the *cudaEvent* is registered

to record this event. Given that the input features for the convolution (CONV), pooling (POOL), and activation (ACTV) layers are read-only data structures, we can start the offload operation for them when they are being performed with forward propagation. As for *stream_{compute}*, *layer_n*'s computation can be started as soon as the *layer_{n-1}* computation is completed without waiting for the completion of the offload operation of *layer_{n-1}*. Non-linear blocks, e.g., residual blocks, can benefit from this strategy because the join operations do not necessarily wait for the completion of the tensor transfer from 1×1 CONV and 3×3 CONV, which is illustrated in Table 1.

stream_{compute} guarantees the completion of computation for *layer_n* by using the *cudaStreamSynchronize()* application programming interface. When both of these events for *layer_n* are finished, a shared queue is used to record this tensor. The release of the tensors chosen for offloading from the GPU is done when there is no dependency for these layers in the shared queue. An individual thread is launched to release *layer_n* from the GPU memory. At the end of the forward propagation, we synchronize *stream_{compute}* and *stream_{memory}* to make sure that *stream_{memory}* has offloaded its feature maps. This safely ensures that all layers chosen to be offloaded are offloaded

ALGORITHM 1 – EXECUTION FLOW FOR NONLINEAR BLOCKS.

```

1: function flowConstruct(int layerId)
2:   if layerID == Null then
3:     return;
4:   end if
5:   refcnt++;
6:   if layerId.refcnt < prevLayer.refcnt then
7:     return;
8:   end if
9:   execFlow.push(layerId)
10:  L = layerId → get-next();
11:  for l ∈ L do
12:    flowConstruct(l)
13:  end for
14: end function

```

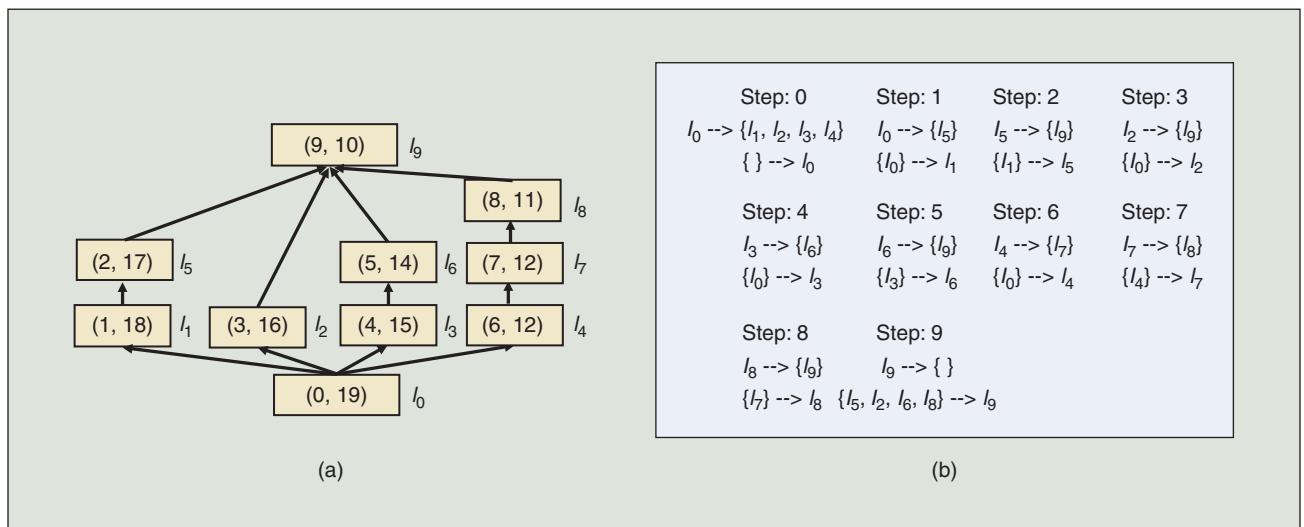


FIGURE 3 – The execution order for the Inception-v4 network in the forward pass. Here l_i represents the i th layer and $l_0 \rightarrow \{l_1, l_2, l_3, l_4\}$ means that layers l_1, l_2, l_3 , and l_4 have a dependency on layer l_0 . (a) A schema of the Inception-v4 network. (b) The forward pass of Inception-v4.

We apply approaches proposed by vDNN and SuperNeurons as the baselines for performance comparison.

from GPU memory before the start of backward propagation, maximizing the memory saving and improving the performance greedily. However, there is an exception in that the execution for the next layer has to be blocked if the available memory is not enough; there is a wait for the release of completed layers. In general, memory space is traded for performance in the forward pass.

Memory Prefetch

In the backward pass, prefetching the offloaded input feature maps back to the GPU can be overlapped with the computation of backward propagation using `cudaMemcpyAsync()` as

well. After an asynchronous transfer for $layer_n$ is completed, a `cudaEvent` is registered in the `stream_memory`, after which the computation can be started for this layer. The `stream_compute` is synchronized with the offload event to guarantee that the computation can be safely launched with the available input feature maps. Similar to the forward pass, we only synchronize `stream_compute` and `stream_memory` at the end of backward propagation before the next iteration. Instead of simply launching the prefetch operations in the reverse order, we have to consider the execution order and prefetch latency when searching for the optimal candidate layer.

Another problem is that if the prefetched $layer_m$ is too far away from the overlapped $layer_n$, the memory-saving benefit is reduced because the prefetched data will be reused immediately, wasting the GPU memory. Jointly considering the memory-saving and prefetch latency, we propose an efficient searching algorithm to decide the layer to be prefetched, as

presented in Algorithm 2. Whenever there is a nonlinear block, we decide on the preceding layers based on DFS, which is similar to the procedure in the forward pass. After obtaining the layer, we restrict that no more than two convolution layers can reside in the GPU, as is illustrated in line 11. This is done because the convolution layers are computation intensive. Prefetching these layers too early will underutilize the GPU resources. As long as it is not a convolution layer and not available yet in the GPU memory, a layer can be chosen as the candidate, as shown in line 14, because other layers require shorter computation times compared with convolution layers. This feature can gain performance improvement because the prefetch latency can be well hidden by the computation time.

Evaluation

Our experimental evaluation is performed on a GeForce GTX TITAN X with 12-GB GPU memory. The machine has a 3.4-GHz Intel i7-3770 CPU (four cores) and a 32-GB CPU memory. The GPU communicates with the CPU via a peripheral component interconnect (PCIe) switch, which has a 16-GB/s data transfer bandwidth. The machine is installed with Ubuntu 16.04, CUDA 9.0, cuDNN 7.0, and g++ 5.4.0.

The Baselines

We apply approaches proposed by vDNN [18] and SuperNeurons [19] as the baselines for performance comparison. Regarding memory management, vDNN uses the default NVIDIA CNMeM [25] library to allocate/deallocate tensors, while SuperNeurons adopts a fast heap-based GPU memory pool utility. The core concept of SuperNeurons is to divide the preallocated pool into an allocated list and an empty list. For these two techniques, we implement the best-fit algorithm as the memory management policy.

The execution order for the nonlinear network is not detailed in vDNN. So we adopt the same construction, DFS, for vDNN for comparison. We can only release tensors from the GPU memory when there is no further

TABLE 1 – THE COMPUTATION AND COMMUNICATION TIMES (IN MILLISECONDS).

LAYER	1 × 1 CONV	3 × 3 CONV	JOIN
Computation	25	76	3
Communication	66	68	×

ALGORITHM 2 – SEARCHING THE CANDIDATE LAYER.

```

1: function searchPrefetchLayer(int layerId)
2:   n = 0;
3:   if layerId → type == CONV then
4:     n++;
5:   end if
6:   next = flowConstruct(layerId).pop();
7:   while id do
8:     if next → type == CONV && n < 2 then
9:       pf.push(id); n++;
10:      next → pf = True;
11:    else if next → of && !(next → pf) && next → type != CONV then
12:      next → pf = True;
13:      pf.push(next);
14:    end if
15:    next = flowConstruct(next).pop();
16:  end while
17: end function

```

reference in the forward or backward pass. As for the tensor scheduling policy, we implement the dynamic policy mentioned in the vDNN article [18], which automatically decides which offloading layers are employed to balance the trainability and performance of a DNN at runtime.

For SuperNeurons, we only implement the liveness analysis and unified tensor management components because recomputation for specific layers is not considered in our work. SuperNeurons supports cost-aware recomputation, which in turn makes it capable of running the benchmark with a larger batch size. We disable SuperNeurons' recomputation feature because cost awareness is not our main concern; instead, we focus on memory-swapping performance, so that SuperNeurons cannot run the benchmark with a larger batch size. As shown in Figure 4, we use Modified SuperNeurons to denote the SuperNeurons' configurations. By contrast, Dymem focuses on improving the performance of nonlinear neural networks and adopts an aggressive

memory-swapping policy to maximize the DRAM bandwidth and balance memory usage. When doing memory swapping, SuperNeurons focuses only on the convolutional layer, while Dymem considers multiple types of layers, such as convolutional and pooling layers. Furthermore, we propose a memory-efficient graph analysis to construct an execution order for nonlinear networks and introduce a dynamic offloading/prefetching strategy to maximize the performance and usage of bandwidth.

The DNN Benchmarks

First, we perform the evaluation compared with vDNN and Modified SuperNeurons on the linear networks VGG-16 and AlexNet. We use the same training configurations as in the published articles [7], [26]. We further perform the evaluation against vDNN and Modified SuperNeurons on two representative nonlinear networks, ResNet [27] and Inception-v4 [5]. Specifically, we implement the basic residual block, which has two 3×3 convolutional layers with the same

number of output channels. Each convolution layer is followed by a batch normalization layer and a rectified linear unit activation function. For all of the previous benchmarks, we use the image data set Canadian Institute for Advanced Research, 10 classes [28].

An End-to-End Throughput Evaluation

Figure 4 presents the end-to-end training throughput comparison of Dymem to vDNN and Modified SuperNeurons. The training throughput is measured by the number of processed images per second. We vary the batch sizes for different DNNs and compare the corresponding throughputs. For the linear networks VGG-16 and AlexNet, there is not much performance improvement over vDNN and SuperNeurons because these networks are composed of simple and sequential layers. For example, VGG-16 consists of 16 convolution layers, which are computation intensive. The computation time is always longer than the transfer time. The propagation computation dominates the total delay. As a result,

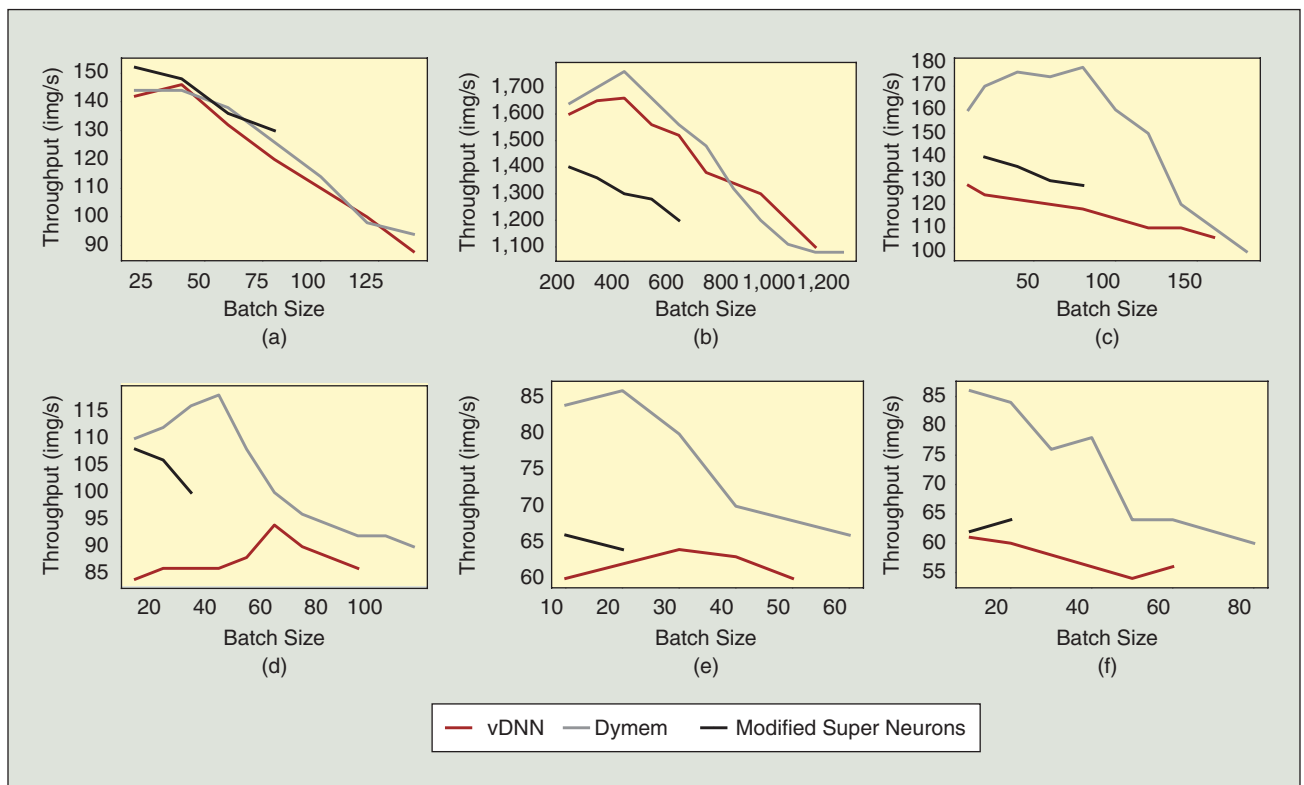


FIGURE 4 – An end-to-end evaluation on the throughput of different DNN models under vDNN, Dymem, and Modified SuperNeurons. (a) VGG-16, (b) AlexNet, (c) ResNet-50, (d) ResNet-101, (e) ResNet-152, and (f) Inception-v4.

Compared with state-of-the-art approaches, our proposed solution can improve the end-to-end training throughput for ResNet-50 by up to 42%.

there is not much performance benefit achieved by removing the layer-by-layer synchronization barriers. For ResNet-152 mentioned in the original SuperNeurons paper, the benchmark can support a batch size up to 80 [26], while the batch size of Modified SuperNeurons is only 20 in Figure 4(e). We can explain this phenomenon by the reason that the original SuperNeurons supports cost-aware recomputation, which brings more free GPU memory space to hold more data and increase training throughput. By contrast, we disable the recomputation feature of SuperNeurons in our experiments so that the maximum batch size of ResNet-152 is lower.

However, Dymem obviously outperforms SuperNeurons in training throughput. In some cases, for linear networks, we can see that Modified SuperNeurons perform better than Dymem and vDNN because Modified SuperNeurons only offloads convolution layers, avoiding the communication overhead. However, for both linear and nonlinear networks, when the batch sizes are increased,

Modified SuperNeurons cannot train these networks because of the limited memory availability. With the growth of the batch size, the memory demand of each layer also increases. In some cases, a GPU can only hold one layer with a huge batch size and may even fail to hold a complete layer because of the limited memory space. For nonlinear networks, the results consistently demonstrate the leading throughput on ResNet-50, ResNet-101, ResNet-152, and Inception-v4. The largest throughput improvement comes from ResNet-50, running with batch size 100, which achieves up to 42% compared with vDNN. The performance largely results from the improved communication/computation ratio. This occurs because Dymem can better utilize the overlap of communication and computation among layers. We also observe that the throughput has slowly deteriorated with increasing batch size. This happens because the GPU memory can only accommodate fewer network layers with wider networks, resulting in a decreased communication/computation ratio. Less

layer overlapping requires growing communications in more frequent tensor swapping between the CPU and GPU. Then, the runtime has to constantly offload the current layer before proceeding to the next one.

The Efficiency of Dependency-Aware Swapping

Figure 5 plots the breakdown of the normalized execution time of two representative nonlinear networks, Inception-v4 and ResNet-32. These two networks are training on Dymem and vDNN with the memory-optimal configuration to avoid the impact from the speedup of convolution. Specifically, the time is decomposed into the overlapped time, the nonoverlapped communication time, and the nonoverlapped computation time. In this experiment, the baseline only uses one stream, which restricts the computation and offload/prefetch in both the forward and backward passes to be executed sequentially. We also configure the memory-optimal algorithm for these three experiments to avoid the impact of the dynamics in the convolution layers. As shown in the figure, the overlapped time in the baseline is zero since the communication and the computation are performed sequentially. The layer-by-layer strategy adopted by vDNN can overlap the communication with the computation to some extent by 18% and 12% for Inception-v4 and ResNet, respectively. The overlapped time in Dymem is longer than that in the vDNN, showing that a more aggressive batching strategy is more effective in terms of performance. As a result, compared with the baseline, Dymem can achieve up to a nearly 46% reduction in the execution time.

Conclusion and Future Work

With DNNs in AI applications going wider and deeper in the industrial community, there is a need to effectively schedule GPU memory for DNN training to overcome an insufficient capacity. In this article, we focus on memory management for the training of nonlinear DNNs. We propose the runtime to adopt a layer-wise

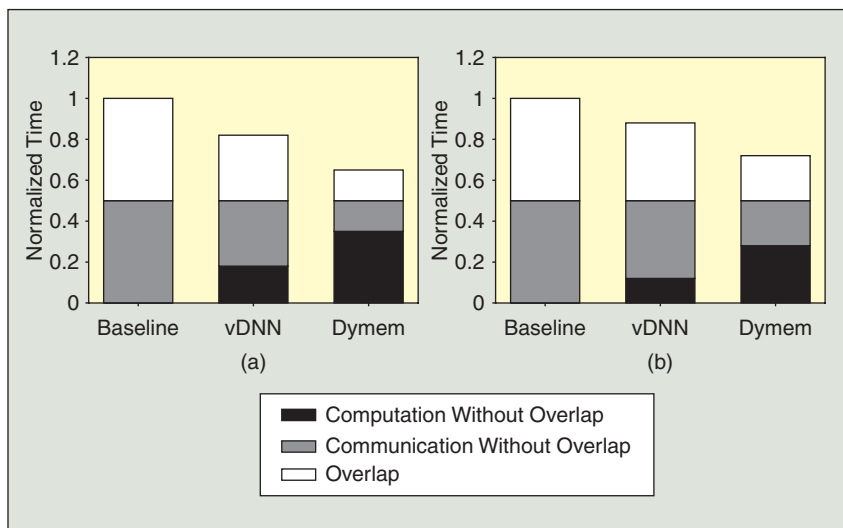


FIGURE 5 – The execution time decomposed into the overlapped time, the nonoverlapped communication time, and the nonoverlapped computation time in two networks. (a) Inception-v4; (b) ResNet-32.

graph analysis and dependency-aware offloading/prefetching strategy to improve the throughput of DNN training. Compared with state-of-the-art approaches, our proposed solution can improve the end-to-end training throughput for ResNet-50 by up to 42%. The experiments also show that Dymem can achieve better scalability for nonlinear DNNs with various network depths. Currently, the proposed solution supports GPU memory optimization only for neural networks with a static dataflow graph and a fixed shape of the input, i.e., a DNN. In the future, we are going to extend our work to support 1) dynamic neural networks, whose data samples have variable shapes and for which the computation graph topology depends on input or parameter values, e.g., recurrent neural networks and long short-term memory networks and 2) some new hardware technologies, such as PCIe 4.0, double data rate five, and nonvolatile memory.

Acknowledgments

This work is supported by NSF grants CCF-1908843 and CNS-2008265. Dazhao Cheng is the corresponding author.

Biographies

Wei Rang (wrang@uncc.edu) earned his B.S. degree in computer science from Shandong Normal University, China, in 2013. He earned his M.S. degree in computer science from Southern Illinois University Carbondale in 2017. Currently, he is a Ph.D. candidate in computer science at the University of North Carolina at Charlotte, Charlotte, North Carolina, 28223, USA. His research interests focus mainly on cloud computing and parallel computing.

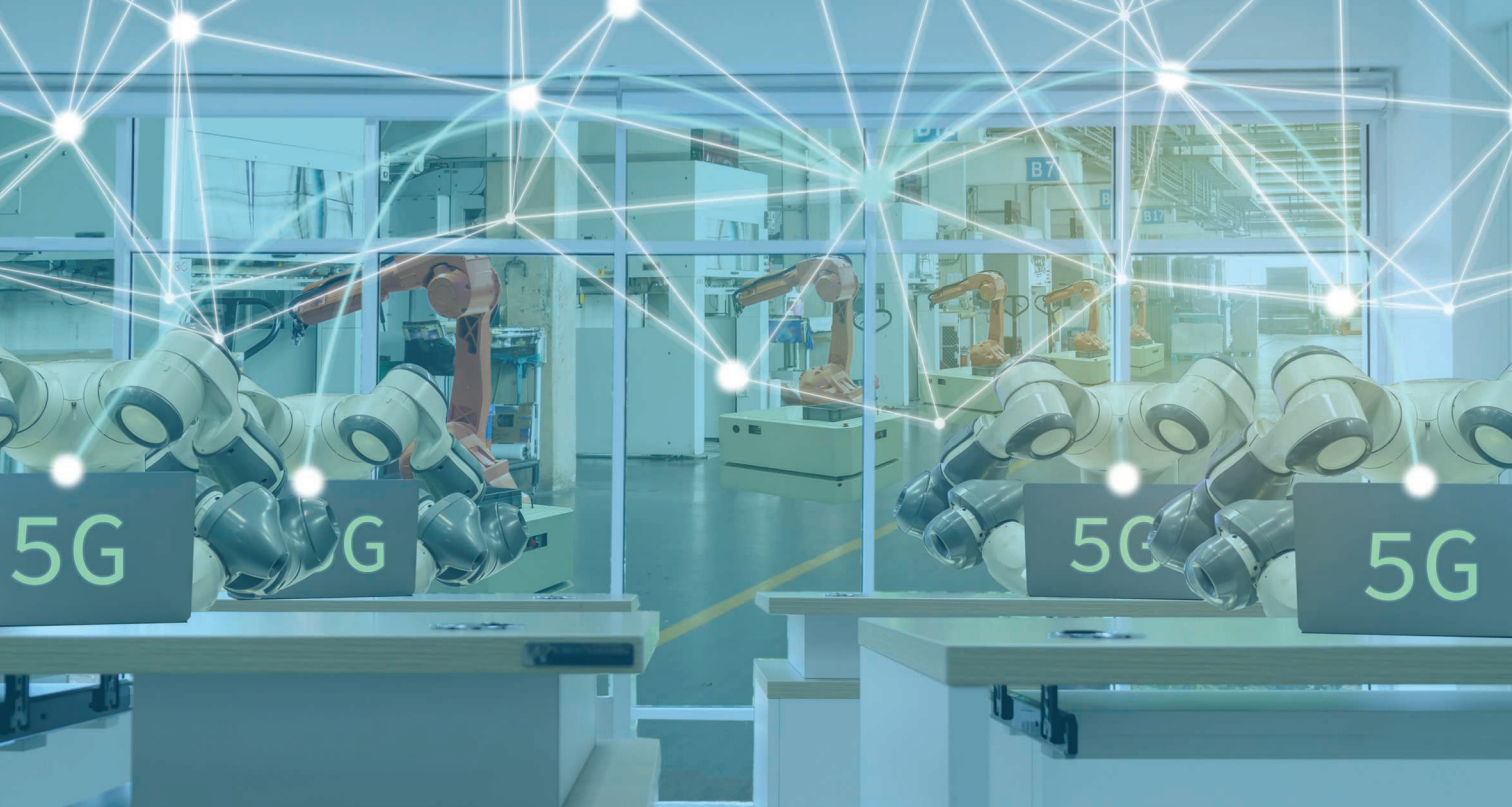
Donglin Yang (dyang33@uncc.edu) earned his B.S. degree in electrical engineering from Sun Yat-sen University, China, in 2016. Currently, he is working toward a Ph.D. degree in computer science at the University of North Carolina at Charlotte, Charlotte, North Carolina, 28223, USA. His work focuses on big data analytics platforms and machine learning computing systems.

The experiments also show that Dymem can achieve better scalability for nonlinear DNNs with various network depths.

Dazhao Cheng (dazhao.cheng@uncc.edu) earned his Ph.D. degree from the University of Colorado, Colorado Springs, in 2016. He earned his B.S. and M.S. degrees in electronic engineering from the Hefei University of Technology, China, in 2006 and the University of Science and Technology of China in 2009, respectively. He is currently an assistant professor in the Department of Computer Science at the University of North Carolina at Charlotte, Charlotte, North Carolina, 28223, USA. His research interests include big data and cloud computing. He is a Senior Member of IEEE.

References

- [1] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
- [2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learning Res.*, vol. 26, pp. 2553–2561, 2013.
- [3] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," *Adv. Neural Inf. Process. Syst.*, vol. 26, pp. 2553–2561, 2013.
- [4] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–87, 2012. doi: 10.1109/MSP.2012.2205597.
- [5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
- [6] S. Bianco, R. Cadene, L. Celona, and P. Napolitano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64,270–64,277, Oct. 2018. doi: 10.1109/ACCESS.2018.2877890.
- [7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556.
- [8] S. Bahrampour, N. Ramakrishnan, L. Schott, and M. Shah, "Comparative study of Caffe, Neon, Theano, and Torch for deep learning," 2016, arXiv:1511.06435v3.
- [9] "NVIDIA v100 Tensor Core GPU," NVIDIA, Santa Clara, CA, 2020. [Online]. Available: <https://www.nvidia.com/en-us/data-center/v100/>
- [10] S. Han et al., "EIE: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 243–254, 2016.
- [11] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, arXiv:1510.00149.
- [12] W. Niu, J. Guan, Y. Wang, G. Agrawal, and B. Ren, "DNNfusion: Accelerating deep neural networks execution with advanced operator fusion," in *Proc. 42nd ACM SIGPLAN Inf. Conf. Program. Lang. Des. Implementation*, 2021, pp. 883–898.
- [13] W. Niu, P. Zhao, Z. Zhan, X. Lin, Y. Wang, and B. Ren, "Towards real-time DNN inference on mobile platforms with model pruning and compiler optimization," 2020, arXiv:2004.11250.
- [14] P. Judd, J. Albericio, T. Hetherington, T. M. Aamodt, N. E. Jerger, and A. Moshovos, "Proteus: Exploiting numerical precision variability in deep neural networks," in *Proc. ACM Int. Conf. Supercomput. (ICS)*, 2016, pp. 1–12.
- [15] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," 2014, arXiv:1412.6115.
- [16] J. Tian et al., "waveSZ: A hardware-algorithm co-design of efficient lossy compression for scientific data," in *Proc. 25th ACM SIGPLAN Symp. Principles Pract. Parallel Programming*, 2020, pp. 74–88.
- [17] A. Jain, A. Phanishayee, J. Mars, L. Tang, and G. Pekhimenko, "Gist: Efficient data encoding for deep neural network training," in *Proc. ACM/IEEE 45th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2018, pp. 776–789.
- [18] M. Rhu, N. Gimelshein, J. Clemons, A. Zulfikar, and S. W. Keckler, "VDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchit. (MICRO)*, 2016, pp. 1–13.
- [19] L. Wang et al., "Superneurons: Dynamic GPU memory management for training deep neural networks," in *Proc. 23rd ACM SIGPLAN Symp. Princ. Pract. Parallel Program. (PPoPP)*, 2018, pp. 41–53.
- [20] D. Yang, W. Rang, D. Cheng, Y. Wang, J. Tian, and D. Tao, "Elastic executor provisioning for iterative workloads on Apache spark," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2019, pp. 413–422. doi: 10.1109/BigData47090.2019.9006021.
- [21] D. Yang, D. Cheng, W. Rang, and Y. Wang, "Joint optimization of mapreduce scheduling and network policy in hierarchical data centers," *IEEE Trans. Cloud Comput.*, 2019, early access, Dec. 23, 2019.
- [22] Tensor <https://en.wikipedia.org/wiki/Tensor>.
- [23] Z. Bai, Z. Zhang, Y. Zhu, and X. Jin, "Pipeswitch: Fast pipelined context switching for deep learning applications," in *Proc. 14th USENIX Symp. Operat. Syst. Des. Implementation (OSDI)*, 2020, pp. 499–514.
- [24] S. Chetlur et al., "cuDNN: Efficient primitives for deep learning," 2014, arXiv:1410.0759.
- [25] Nvidia cnmem, NVIDIA Corp., Santa Clara, CA, 2020. Accessed: Aug. 4, 2021. [Online]. Available: <https://github.com/NVIDIA/cnmem>
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
- [28] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, 2009. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220>



©SHUTTERSTOCK.COM/MARVELOUS STUDIO

Factory 5G

A Review of Industry-Centric Features and Deployment Options

AAMIR MAHMOOD,
SARDER FAKHRUL ABEDIN,
THILO SAUTER, MIKAEL GIDLUND,
and KRISTER LANDERNÄS

Digital Object Identifier 10.1109/MIE.2022.3149209
Date of current version: 23 February 2022

Fine-grained and wide-scale connectivity is a precondition to fully digitalize the manufacturing industry. Driven by this need, new technologies such as time-sensitive networking (TSN), 5G wireless networks, and industrial Internet-of-things (IIoT) are being applied to industrial communication networks to reach the desired connectivity spectrum. With TSN emerging as a wired networking solution, converging IT and operational technology (OT) data streams, 5G is upscaling its access and core networks to function as an independent or a transparent TSN carrier in demanding OT use-cases. In this article, we discuss the drivers for future industrial wireless systems and review the role of 5G and its industrial-centric evolution towards meeting the strict performance standards of factories. We also elaborate

on the 5G deployment options, including frequency spectrum allocation and private networks, to help the factory owners discern various dimensions of solution space and concerns to integrate 5G in industrial networks.

The envisioned Industry 4.0 defines the transformative progression of industrial manufacturing systems toward future smart factories, offering flexibility and efficiency [1], [2]. The smart factories of the future will be characterized by [3], [4] 1) holistic management in the entire value chain through the horizontal and vertical integration of OT and enterprise IT domains, 2) modular and customizable production lines, and 3) support for mobility use cases, including mobile robots, control panels, and automated guided vehicles (AGVs). However, such progression requires the seamless connectivity of people, machines,

and computing resources in manufacturing processes.

Connectivity is fundamental to collecting and utilizing data for bridging the divide between physical and digital worlds and developing new revenue streams and cost savings. The challenge is to satisfy diverse connectivity demands for reliability, low latency, and capacity. Industrial OT networks include diverse solutions, such as Ethernet-based networks (e.g., the Process Field Net and Ethernet for Control Automation Technology) and field buses (e.g., the Process Field Bus) [5]. Meanwhile, the IEEE 802.1 TSN task group is developing a set of standards to extend the best-effort Ethernet networking model to provide deterministic streaming services [6]. Concurrently, the TSN profile (defining features, options, and requirements) for industrial automation is being defined by International Electrotechnical Commission (IEC)/IEEE 60802. These initiatives will provide enhanced industrial Ethernet solutions for enterprise-wide connectivity, support the capabilities of industrial Ethernet variants, and increase hardware homogeneity on the factory floor [7], [8]. However, the full-scale connectivity of the entire value chain requires a combination of wired and wireless solutions to support massive sensing, mobility, and customizable

Connectivity is fundamental to collecting and utilizing data for bridging the divide between physical and digital worlds and developing new revenue streams and cost savings.

production lines. Nevertheless, wireless systems remain unpopular because of reliability and performance concerns around industrial-grade deterministic communication [9]. Thus, current wireless networks exist on a limited scale, penetrating factories only to process automation and for other noncritical applications [10].

Over the years, cellular systems have been optimized to provide extended coverage and data rates for the mobile Internet. However, connectivity demands in vertical industries, such as factory and building automation, energy, and transportation, are substantially different. To this end, 5G networks are upscaling in terms of capability and flexibility, while the network design stretches beyond consumer-oriented mobile broadband services by introducing features tailored for the IoT, IIoT, and connected cyberphysical systems (CPSs) [11], [12]. The IIoT describes a communication network of industrial objects (machines, devices, and processes)

for reliably exchanging monitoring/control data, while CPSs use the IIoT to provide a consistent (synchronized and interactive) digital description of the objects [13].

In fact, 5G introduces three main connectivity services: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultrareliable low-latency communication (URLLC). The apparent industrial use cases of these services are shown in Figure 1, yet by combining these services, 5G can support extreme variations of IIoT applications. Although 5G becomes a unified connectivity fabric, its seamless introduction to the factory floor raises several questions and challenges. For instance, what connectivity gaps can 5G fill? How is it integrated with industrial networks, and can it satisfy time-sensitive communication (TSC) targets while keeping critical data local? Consequently, businesses and mobile network operators (MNOs) are seeking to understand the 5G

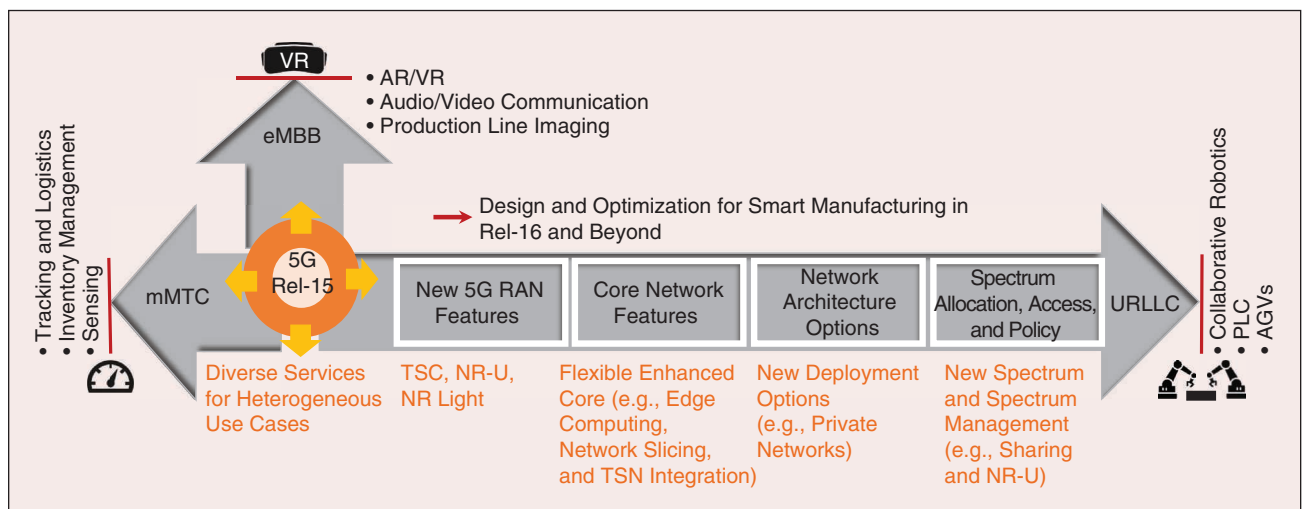


FIGURE 1 – The 5G design and optimization for diverse industrial applications. 3rd Generation Partnership Project Release 15 (Rel-15) defined eMBB, mMTC, and URLLC services for 5G New Radio (NR). Rel-16 focused on IIoT-related enhancements, including time-sensitive communication (TSC), 5G-TSN integration, localization services, private networks, network slicing, and NR on unlicensed bands (NR-U) for URLLC; Rel-17 and beyond will cover further improvements in private networks, convergence with industrial networks, network automation, and NR Light for new use cases [14]. VR: virtual reality; AR: augmented reality; RAN: radio access network; PLC: power line communication.

architecture's design and optimization (the extending ripple in Figure 1) to expand their business models beyond mobile broadband [15], [16].

In this direction, private (nonpublic) 5G networking models [17] and associated spectrum licensing options need to be carefully explored to realize dedicated, customizable, and cost-effective services for industrial use cases [18]. While reflecting on such concerns and challenges, this article presents a technical review of 5G's industry-centric features and deployment options in the ongoing effort to develop factory 5G networks, i.e., private 5G networks for industrial use. The term *factory 5G* is specifically used to refer to a private 5G network deployed for the process/manufacturing industry.

Drivers for Enhanced Industrial Wireless Communications

Through full mobility support and the replacement of cables, wireless automation has the potential to transform industrial production systems. However, until recently, the wireless communication sector lacked a clear understanding of the integration scenarios and dependable communication requirements

of industrial use cases. Studies (e.g., [19] and [20]) by the 3rd Generation Partnership Project (3GPP) and close collaboration between wireless communication and industrial sectors have led to the industry-specific design of time-sensitive, ultrareliable, and massive connectivity features in 5G networks. These features can serve a wide range of use cases on the factory floor, independently or jointly (by augmenting wired industrial solutions), as shown in Figure 2. Particular needs for enhanced wireless systems arise from the following:

- *Dynamic network customization:* Scaling capacity, speed, and control according to changing manufacturing demands require rapidly reconfigurable and modular production lines, termed *swarm* or *matrix production* [4], [21]. However, due to costly and complex cable management, production lines are mostly statically configured. By leveraging wireless connectivity, the desired flexibility can be achieved; however, radio resource allocation and network management become challenging.
- *Mobility and collaboration:* Support for mobile objects (AGVs, robots, and control panels) and their col-

laboration in flexible/modular manufacturing are vital to automate recurring, labor-intensive, and costly tasks. Mobility and collaboration-oriented use cases, referred to as the *critical IoT* in Figure 2, require robust wireless connectivity to guarantee fail-safe operation [5].

- *Data-intensive use cases:* To enable remote inspection, monitoring, and surveillance, human-machine interface (HMI) applications for real-time situational awareness are needed, including control panels, massive diagnostic data uploads, augmented reality (AR)/virtual reality (VR), and high-definition video/audio [19]. These use cases are data intensive and radio resource demanding; thus, they are classified as the broadband IoT in Figure 2.

Table 1 lists industrial use cases and their key performance indicators (KPIs) based on 3GPP studies [19], [20]. Meanwhile, the most relevant 5G use cases across various industries and their influence with respect to generated traffic volume on overall network architecture are projected in [22]. These studies underscore the need for a versatile wireless solution to satisfy diverse connectivity

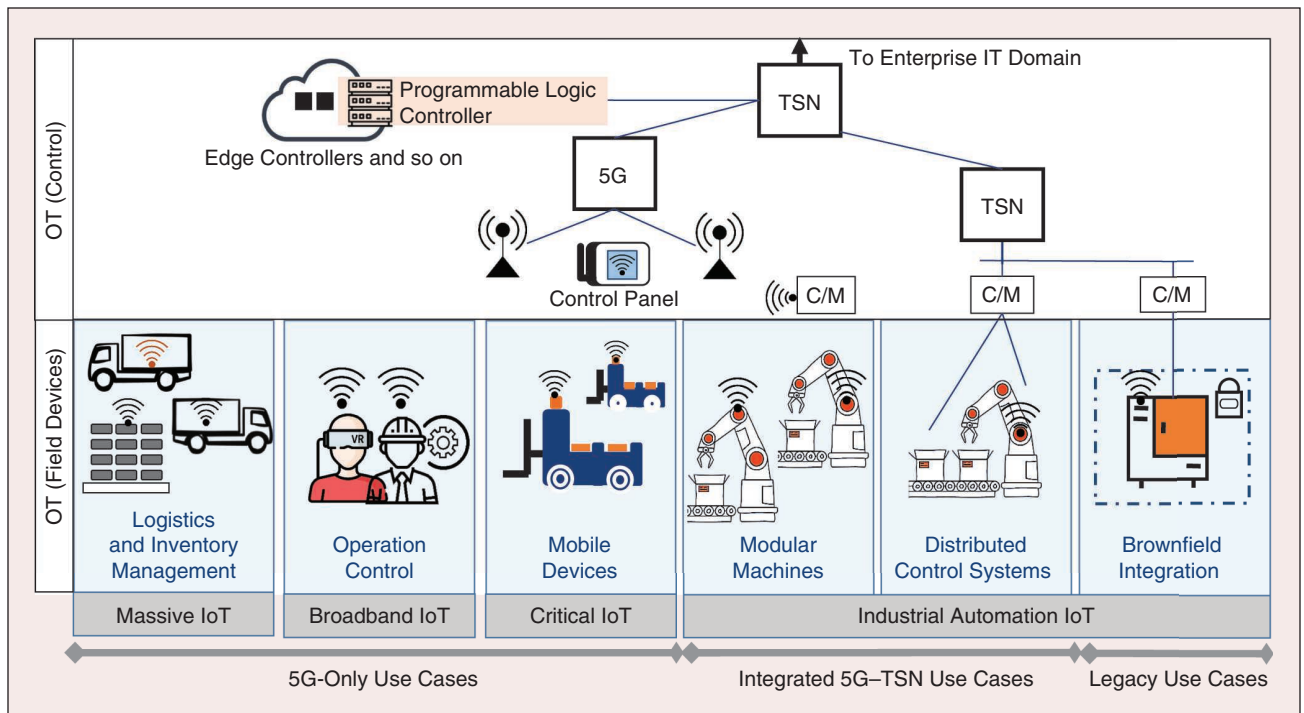


FIGURE 2 – Factory 5G use cases in a converged industrial network, which can broadly be classified into various IoT segments: massive IoT for wide-scale tracking and massive sensing; broadband IoT for massive monitoring, imaging, and AR/VR; critical IoT for collaborative robotics and AGVs; and industrial automation IoT for modular/distributed control with/without wired-wireless convergence. C/M: controller/master.

requirements, ranging from time sensitive to data intensive, and massive communications with scalability and mobility, as depicted in Figure 2. The wireless standards explicitly designed for industrial communication (e.g., International Society of Automation 100.11a and the Wireless Highway Addressable Remote Transducer Protocol) are limited in scope; using the low-rate IEEE 802.15.4 standard, they offer limited rates, coverage, and scalability. Conversely, 5G's flexible, service-oriented design and targeted evolution are expected to provide unified industrial connectivity based on eMBB, mMTC, and URLLC services, each enabling coverage and capacity, massive access, and TSC.

On the downside, compared to the ad hoc and unlicensed spectrum-based operation of IEEE 802.15.4, a factory 5G deployment requires technical and business evaluations to demonstrate its technoeconomic feasibility. The technical challenges that persist today are 1) the feasibility of satisfying industrial KPIs and convergence with industrial networks and 2) the control over business-critical data and network/service customization. The business viability of factory 5G cannot be predicted this early, and vendors and operators are struggling to define a clear road map for the technology's rollout (e.g., based on early stage testbeds [4]). Besides,

The business viability of factory 5G cannot be predicted this early, and vendors and operators are struggling to define a clear road map for the technology's rollout.

the figures for business value are more hyped than proved to determine concrete returns on investment (ROIs).

Nevertheless, to mitigate technological concerns, 5G is unfolding New Radio (NR)-based radio access network (RAN) and core network (CN) designs consistent with industrial-grade quality-of-service (QoS) targets as well as end-to-end (E2E) latency of < 1 ms and reliability on the order of $1-10^{-6}$ [20]. These stringent E2E performance targets require the joint dimensioning of the RAN and CN, while two broad directions are as follows:

- industry-centric enhancements in radio and CN features for TSC and integration with wired Ethernet (e.g., TSN)
- offering various network/service deployment options with private 5G to meet industrial requirements for 1) dedicated coverage and guaranteed QoS; 2) keeping business-critical data local, secure, and private; and 3) flexible spectrum use options.

5G System Features and Enhancements

A 5G system (5GS) consists of user equipment (UE), the RAN, and the CN. The CN includes the control plane (CP) for signaling, user plane (UP) for data connectivity, and provisioning/management layer. The division between the CP and UP provides flexible network deployment/operation, where only the Ethernet/Internet Protocol connection is visible to a 5GS user (e.g., an industrial network).

5G RAN Design for TSC

To support critical data flows, 5G NR introduces several URLLC features broadly grouped into low latency, ultrareliability, scheduling, and over-the-air time synchronization [14]. For low-latency communications, NR provides adjustable resource granularity and minislots in a radio frame. The former enables shorter time slots, while the latter facilitates prioritized transmissions at other-than-slot boundaries.

TABLE 1 – THE FACTORY-OF-THE-FUTURE APPLICATIONS AND USE CASES, WITH PERFORMANCE REQUIREMENTS (BASED ON [19] AND [20]).

APPLICATION AREA	USE CASE	KEY PERFORMANCE INDICATORS				
		RELIABILITY	LATENCY	DATA RATE	PAYLOAD	DEVICES
Factory automation	Motion control	99.9999%	0.5–2 ms	1–5 Mb/s	20–50 B	20–100
	Control to control		10–50 ms	–	1 kB	Five to 10
	Mobile robotics (cooperative)		1–50 ms	–	40–250 B	100
Process automation	Closed-loop process control	99.9999%	≤ 10 ms	–	20 B	–
	Process monitoring	99.99%	50–100 ms	0.5–2 Mb/s	–	100–1,000
	Condition monitoring (safety)	99.9%	5–10 ms	0.1–0.5 Mb/s	–	$> 1,000$
	Condition monitoring (interval/event based)	99.9%	50 ms–1s	0.1–0.5 Mb/s	–	$> 1,000$
HMI and production IT	Mobile control panels with safety control	99.9999%	4–12 ms	–	40–250 B	Two to four
	AR/VR	99.9%	< 10 ms	5–25 Mb/s	–	10–20
Logistics and warehousing	Mobile robotics (video operations)	99.9999%	10–100 ms	–	15 k–250 kB	100
	Mobile robotics (standard operations)		40–500 ms	–	40–250 B	100
Monitoring and maintenance	Massive wireless sensor networks	Noncritical, massive devices, and energy aware				

Moreover, NR avoids handshake-based transmission grants by semipersistent and configured grant scheduling schemes to enable periodic transmissions on preconfigured resources. Also, to provide prioritized access, NR introduces uplink/downlink preemption by which a URLLC transmission can preempt an ongoing noncritical one. Besides, NR supports hybrid automatic repeat request (HARQ)-based retransmissions within lower latency bounds by fast packet processing. NR meets URLLC reliability targets via ultrarobust modulation and coding schemes (MCSs) for data and control channels. To avoid the extra degree of uncertainty in control channels, NR supports grant-free transmissions. In addition, various diversity schemes, such as multiantenna techniques, dual connectivity, multiple carriers, and packet duplication, are part of the URLLC reliability toolbox [5].

Time synchronization is an essential element of 5GS and deeply embedded in radio/CN entities through a telecom profile of the precision time protocol (PTP). 3GPP Release 16 (Rel-16) extends time synchronization support to the UE level for time-sensitive applications, using a reference time indication from the base station (BS) and timing advance (TA)-based adjustment of the propagation delay (PD) at the devices. However, it is hard to meet the tight synchronization budget of $\leq 1 \mu\text{s}$ in a 5GS due to BS time alignment errors, limited TA granularity, and other timing and multipath propagation errors [23]. Therefore, further study and analysis are needed for enhanced PD compensation techniques with a high-resolution TA, round-trip time-based PD estimation with dedicated signaling, and other innovative directions [24]. Since synchronization complexity is accentuated in converged wired/wireless networks, readers may refer to [25] and [26] for synchronization techniques/analysis in such scenarios.

Note that all these RAN features cover the nontrivial design aspects for TSC. Further, the design and assessment of these features in dynamic (with respect to multipath fading, blockage, and interference) industrial wireless channels require statistical

learning and analysis frameworks [27]. This is because statistical channel models, often developed based on measurements [5], cannot capture environment-specific conditions, correlations of events, and occurrences of performance-disrupting, rare events.

5G CN Design for TSN Integration

The 5G CN provides various mechanisms for greenfield industrial deployments and to support interworking with Ethernet-TSN networks, such as the following:

- 1) Packet duplication is achieved by establishing redundant UP paths within the RAN, core, and transport network [28]. This feature is similar to TSN's frame replication and elimination for reliability (FRER) features defined in the IEEE 802.1CB-2017 standard or parallel redundancy protocol specified in the IEC 62439 standard for the industrial Ethernet. FRER protects against packet failures and latency violations by transmitting every data packet concurrently across two independent paths.
- 2) Native support for Ethernet services and procedures enables integration with industrial Ethernet networks.
- 3) Network slicing, which can span all 5GSs and even multiple operators, enables creating virtual networks across a common physical network, each capable of providing a negotiated QoS to a private customer. Network slicing techniques can be exploited to isolate TSN-related control/data traffic in 5G-TSN integration scenarios and to realize factory 5G on a public network (see the "Factory 5G Deployment Options" section).
- 4) An edge computing environment brings network functions (NFs) close to a wireless edge to improve privacy and service performance for real-time, control-sensitive use cases. Edge computing can also be applied to the Rel-17 feature of time-sensitive device-to-device communication in TSN via a UP function (UPF).

These features constitute the 5G integration path with TSN to establish

combined wired-wireless solutions for full-scale industrial connectivity. 5G-TSN integration can enable new use cases and migrate existing ones requiring deterministic communications between end devices and remote controllers. Similar to the 5G URLLC toolbox, TSN provides an open set of standards for guaranteed latency, ultrareliability, time synchronization, and resource management, with Figure 3 illustrating the 5G URLLC and TSN toolboxes of standards.

3GPP Rel-16 introduced TSN translator (TT) functions to interface a 5GS to an external TSN network as a (set of) virtual TSN-capable bridge(s) [23]. TTs expose the necessary 5GS CP and UP interfaces for handling TSN's bridge configuration, flow control, and time synchronization while masking 5GS internal functions. In particular, with TT functionality at the CP, the 5G TSN application function (AF) provides the management application to interact with the TSN controller [central network controller (CNC)] according to IEEE 802.1Qcc to report capabilities and receive configurations from CNC via a network management protocol [29], [30]. Here, the 5G AF is an example of NF virtualization (NFV) for allocating/configuring 5G network resources for TSN streams. Meanwhile, 5GS provide virtual bridges per a UPF on the network side by using a network-side (NW) TT as a gateway to the TSN, with a device-side (DS) TT acting as a port for a 5G wireless device. For synchronization, 5GS use an internal clock for adjusting the residence time of TSN's synchronization messages in the 5G system (i.e., between an NW TT and a DS TT) according to IEEE 802.1AS. Further details on 5G-TSN integration and research challenges are available in [18], [29], and [30].

Radio Spectrum Options for Factory 5G

Spectrum policy—how spectrum is licensed—is critical to 5G solutions and services. Spectrum allocation is essential for growing data rate demands, and it is critical for industrial applications with stringent QoS requirements. A key question for industries concerns

which spectrum is suitable for factory 5G. Currently, licensing authorities are at different stages of 5G spectrum assignment. The NR spectrum for 5G is being allocated in low-, mid-, and high-band frequency ranges, providing different performance and coverage characteristics (see Figure 4). Of these, the mid-band 3.3–4.2-GHz spectrum is deemed ideal for 5G signaling, giving an optimal tradeoff among bandwidth, distance coverage, and building penetration.

The harmonized use of this spectrum range for 5G services is under consideration worldwide, while in the European Union, the 3.4–3.8-GHz band has already been harmonized. Meanwhile, in the United States, the 3.55–3.7-GHz band, called the *Citizens Broadband Radio Service*, enables 5G network deployments through shared access via a three-tiered mechanism, including incumbent, priority access, and general access licenses.

Dedicated Spectrum for Industries

Traditionally, spectrum policy follows a market-based approach—using auctions to ensure efficient spectrum allocation for new use cases. However, spectrum reservation for localized use is an unprecedented approach, promoting industrial policy. The idea of dedicated spectrum allocation for industries was initially promoted by Germany, citing that 1) many factories are located in rural areas where MNOs will never utilize

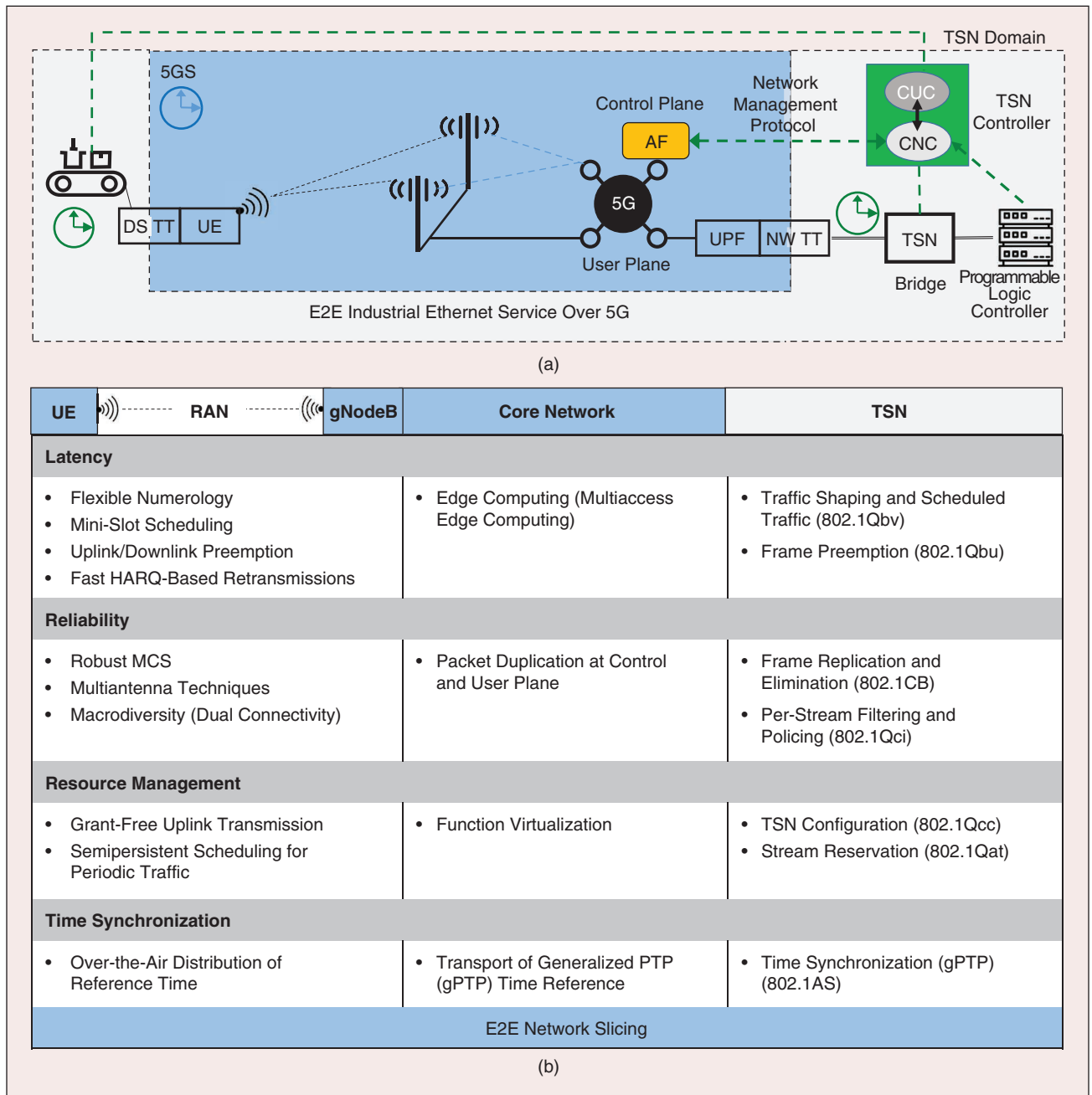


FIGURE 3 – An example of TSN over 5G by mapping 5G and TSN features. (a) The time-aware 5GS as a virtual TSN bridge. (b) Feature mapping between 5GS (RAN and Core) and TSN with respect to latency, reliability, resource management, and time synchronization for E2E industrial services. DS: device side; TT: TSN translator; AF: application function; CUC: centralized user configuration; CNC: central network controller; NW: network side.

the spectrum, 2) only dedicated spectrum will enable a leading role for 5G applications, and 3) it will allow industries to independently operate private 5G and protect trade secrets. At present, Germany has reserved 100 MHz in the 3.7–3.8-GHz band for local licenses, while the United Kingdom has made the 3.8–4.2-GHz band available through shared/coordinated spectrum access for local licenses. Meanwhile, the Netherlands and Sweden have proposals to allocate at least 80 MHz in local licenses in the midband. A comprehensive snapshot of spectrum allocations for industries is available in [31].

Alternative Spectrum Access Approaches

Although clean and dedicated licensed spectrum is preferred for industrial 5G, there could be several spectrum use alternatives for industries without spectrum reservation, e.g., through spectrum subleasing from MNOs, (licensed) shared access, and unlicensed spectrum. Another alternative is to use licensed spectrum under an agreement with a mobile operator. There is precedence for this, with some private

LTE networks typically using the model for large, high-value clients. Usually, facilities with a demanding QoS prefer licensed spectrum to avoid any risk of communication failures. This holds in the 5G era; however, emerging technologies [e.g., spatial diversity and artificial intelligence (AI)-enabled spectrum sharing and interference mapping] are empowering the construction of robust wireless networks even across license-free and shared bands [32], [33]. Also, NR on unlicensed bands (NR-U) is expected to be enhanced for industrial use cases in Rel-17 ... and beyond [14]. Nevertheless, while considering industrial use case requirements, a portfolio of spectrum usage options can be developed, as demonstrated in Figure 5.

Factory 5G Deployment Options

Private networking is another industry-centric 5G feature to support factory automation with desired flexibility. 5G private networks give complete control over various critical network aspects, including capacity and coverage, isolated use of network resources, and on-demand customization. The

3GPP defines such networks as *non-public networks (NPNs)*. Private 5G offers flexible deployment options using physical and virtual elements of a public cellular network, where private and public data can be segregated to maintain security/privacy [17].

NPN Functional Perspective

From a functional standpoint, the 3GPP defines two types of NPNs: stand alone (that is, isolated [17]) and public network integrated (PNI), while the applicable model depends on the use case and its requirements (see Figure 6).

Isolated NPN

An isolated NPN has all functions (control and data planes) located inside the factory, while it can use any of the previously discussed spectrum options. Although dedicated 5G may result in high cost and management complexity, it gives scalability and security gains.

PNI NPN

Depending on how 5G functional elements are distributed across a deployment area, this option provides different integration levels with the public network.

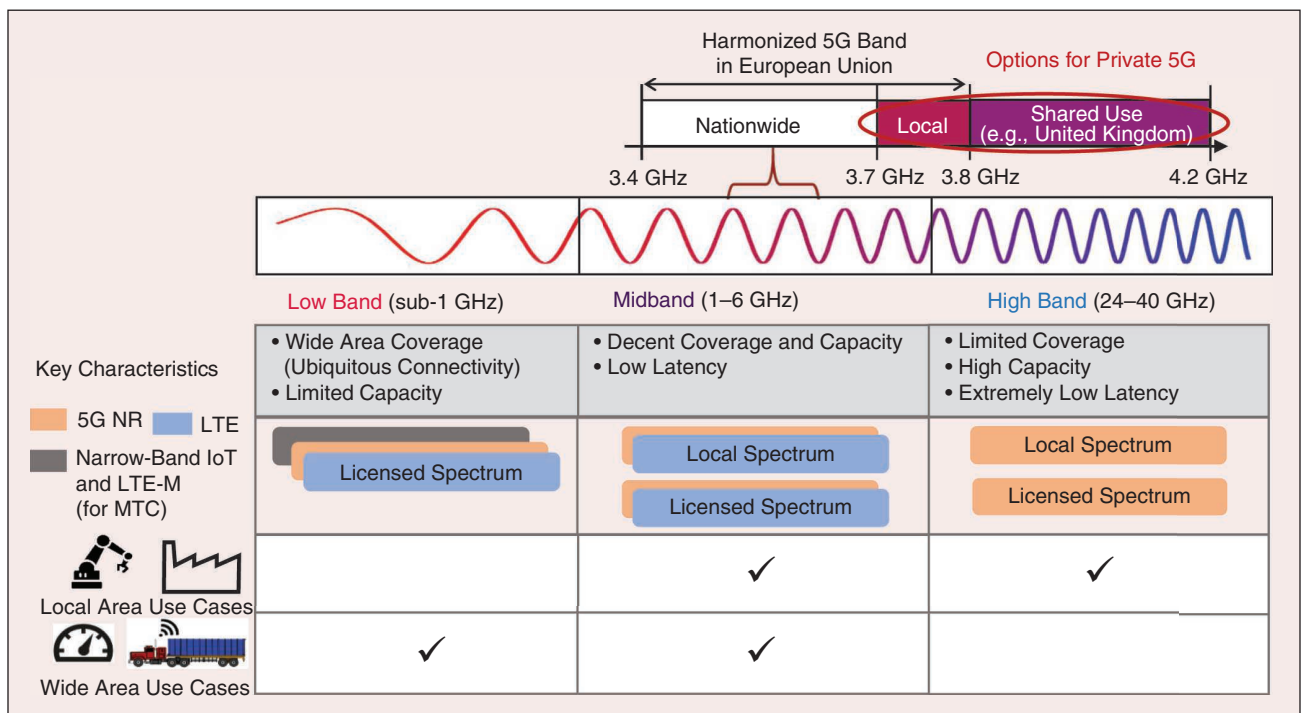


FIGURE 4 – The 5G spectrum allocation in different bands, with key characteristics and band-specific use cases: low bands for wide area tracking and sensing (i.e., massive IoT); high bands for local area, data-intensive, and time-sensitive applications (i.e., broadband/critical/industrial automation IoT); and midbands for functions requiring decent indoor–outdoor coverage and capacity. LTE-M: LTE for machines; MTC: machine-type communications.

Consequently, a nonservice provider (e.g., an OT company) will need to cooperate with an MNO for NF sharing, network management, service

continuity, and spectrum access. The main options are as follows:

- *NPN hosted by public network*: In this model, the public network dedicates

user and CP resources to an enterprise using network slicing or access point name functionality. A network slice consists of a dedicated or shared

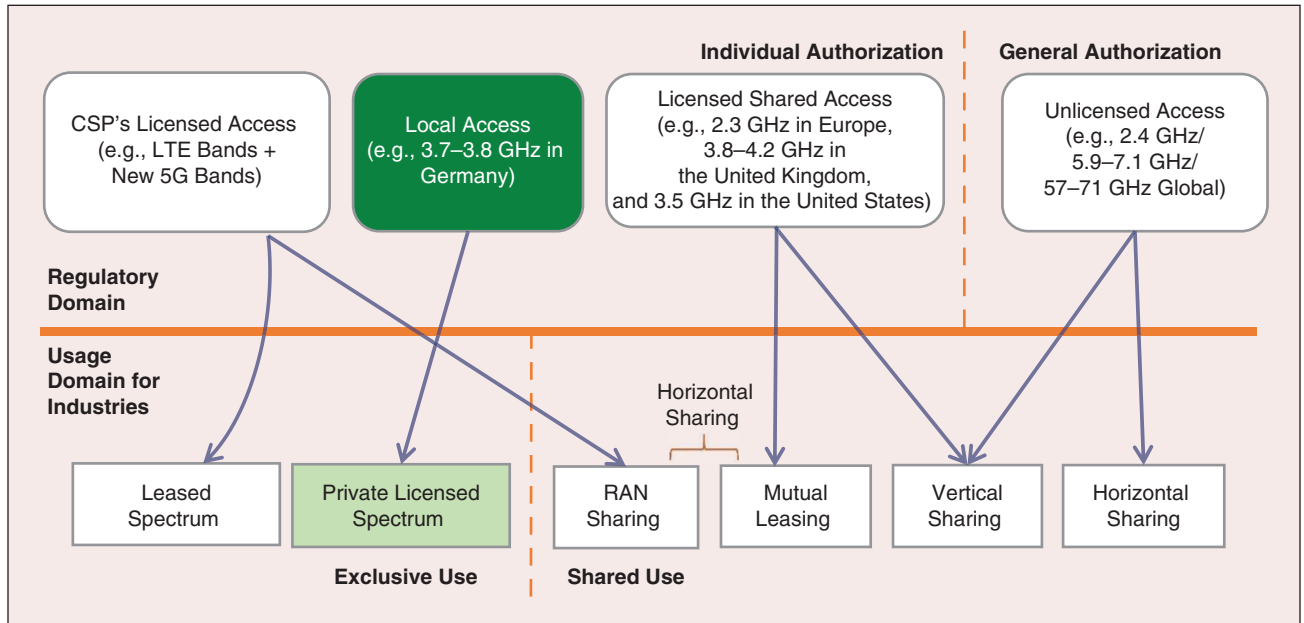


FIGURE 5 – The spectrum use options for vertical industries, based on spectrum regularity models. CSP: communication service provider.

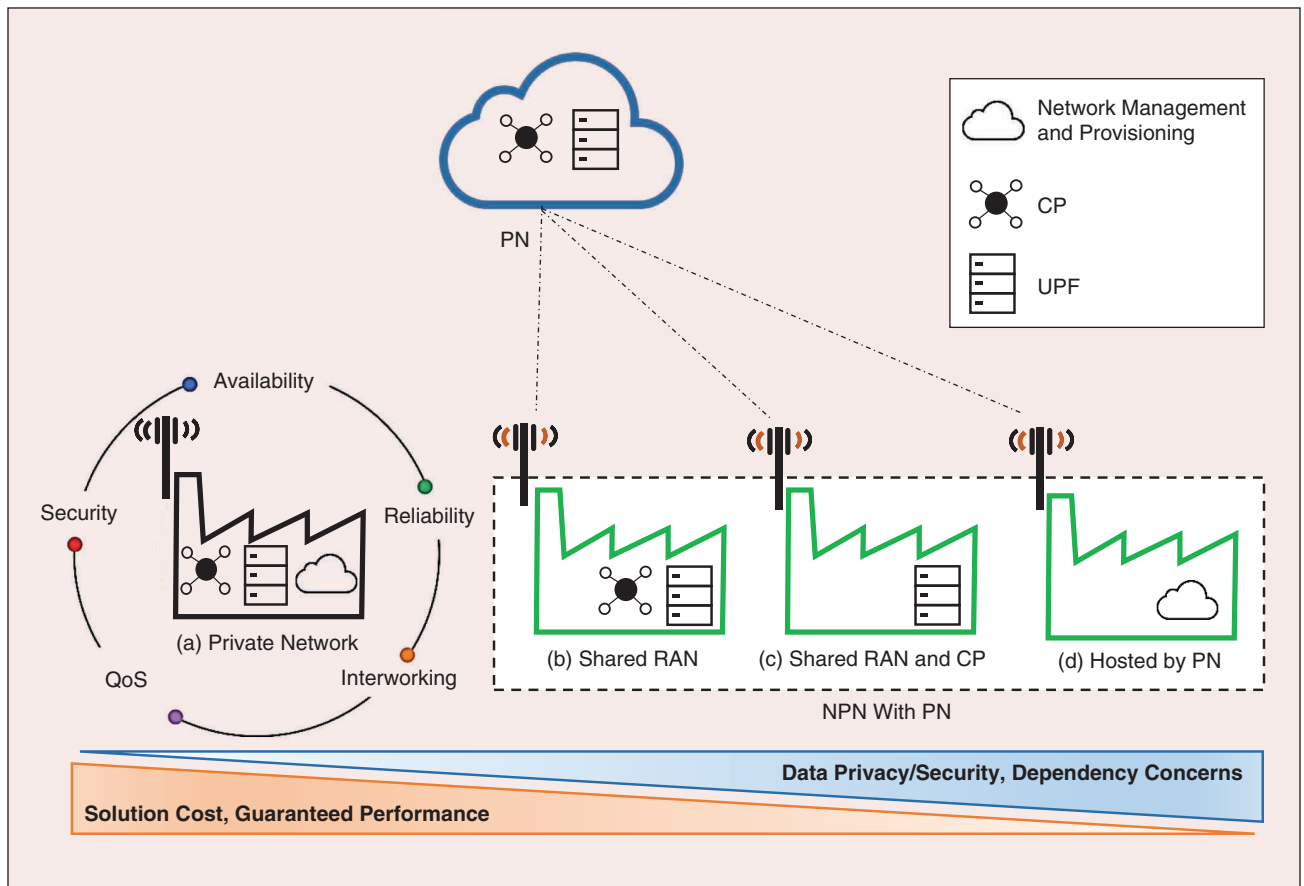


FIGURE 6 – The deployment options for 5G NPNs. From (a)–(d), the solution cost and performance guarantees reduce, while data privacy, security, and dependency concerns increase (based on [38, Tab. 5]).

subset of network resources (e.g., processing power, storage, and bandwidth). It can concurrently support heterogeneous service classes, e.g., an eMBB slice supporting data- and time-sensitive applications. Also, edge computing, as part of a network slice, can enable NFs as software instances for industrial use cases [34]–[36]. Using this model, enterprises can get private network benefits without upfront installation and operation cost/complexity.

- *Shared RAN*: An enterprise shares only the RAN with the public network, enabled by 3GPP RAN-sharing specifications [37]. Since the 5G users and CPs remain local, the data/signaling are local at the factory premises.
- *Shared RAN and CP*: This option shares the RAN and CP with the public network. The local UP can provide access to application layer data through multiaccess edge computing, e.g., for real-time control communications, and enable other value-added services, such as data analytics, location-based services, and information caching. Note that CP placement becomes critical when extensive control and signaling are involved.

NPN Operational Perspective

Concerning operation models, NPNs can be placed along three dimensions: management, ownership, and spectrum policy. Examples can include the following:

- *MNO-provided service*: An MNO owns/manages private 5G using licensed spectrum.
- *Self-managed service*: An OT company exclusively owns/manages a dedicated NPN using leased, local, or unlicensed spectrum.
- *Hybrid solution*: An OT provider manages its private 5G and partners with an MNO for licensed spectrum, or an MNO provides 5G using an OT company's private spectrum. Management/monitoring tasks in private 5G define how an OT company can manage (create, scale, and configure) and monitor a deployment model's NFs. Usually, OT providers prefer to have

complete control to manage and adapt NFs and resources and troubleshoot an NPN's behavior according to industrial processes' functional and performance needs. Meanwhile, the real-time monitoring of critical traffic's QoS and service availability requires employing machine learning/AI-based predictive management solutions [39]. In this respect, 5G softwarization efforts (e.g., OpenAir-Interface, OpenRAN, software-defined networking/NFV [18], [34], [35]) will accelerate private 5G cost-effective realization and customization.

Security and Privacy in Factory 5G

Apart from encryption and data integrity, the privacy and security concerns/solutions of private 5G deployment span service resilience against failures and radio jamming, service isolation, and interoperability with existing industrial security functions. In any deployment model, 5G's service-based architecture provides resilience through the redundancy/duplication of NFs and resources, while radio jamming attacks can be avoided by combining techniques, such as spectrum monitoring, dynamic spectrum access, and beamforming [32], [40].

Traditionally, OT networks are physically isolated by perimeter protection and access control mechanisms [41]. In this respect, isolated NPNs can provide the highest security via physically isolated network elements; however, PNI NPNs need logical isolation/access authorization and additional E2E encryption and integrity protection mechanisms. For isolated NPNs, 5G supports operator-controlled alternative device authentication methods [e.g., a key-generating Extensible Authentication Protocol (EAP)–Transport Layer Security protocol] [42]. In contrast, for PNI NPNs, using 5G authentication and key management (AKA) and EAP–AKA authentication methods is mandatory, together with universal subscriber identity models for public network authentication. When an NPN is deployed as a slice, slice-specific authentication can be optionally performed after primary device–network verification [42].

To avoid the overhead and extra delay of additional integrity protection

measures and E2E encryption, other solutions, e.g., physical layer security and intrusion detection mechanisms, must be investigated. Besides intrusions, network jamming attacks are potential cyberthreats for which federated and reinforcement learning-based collaborative attack detection/defense mechanisms are emerging [43].

Reflections and Conclusion

Connectivity is key to future smart manufacturing's strive for digitization and agile factory operations. With 5G wireless communications, this vision is becoming more plausible via mobility support, augmenting workers, collaborative robotics, edge computing and control, and machine vision. Industrial applications entailing these functions have diverse and stringent KPIs for which emerging 5G networking capabilities and services are creating an opportunity to realize wireless industrial control. However, as a factory 5G solution, the technology has to satisfy the KPIs and deal with the integration, management, and operation complexities of OT networks. In this article, we discussed industrial use cases driving 5G design and optimization and reviewed industry-centric network features, deployment models, and spectrum options.

3GPP Rel-16 and Rel-17+ continuously study and specify enhancements to handle industrial use cases. Even so, it is too early to say when industry-friendly deployments will exist due to the gap between product development and industrial acceptance and needs. Technical feasibility and solutions for industrial KPIs constitute one side of the coin. The other side is multifaceted, including the following:

- 1) 5G URLLC features address non-trivial issues, but their fail-safe design and validation for dynamic industrial wireless channels require new statistical frameworks and data-driven strategies.
- 2) To create trust, validation and evaluation in real deployments are needed.
- 3) Cross-industry communication, standardization, and regulations must be more inclusive.

For example, related to point 2, recent measurement results reported in

[44] with a Rel-15-based private 5G network indicate the need to reduce the upper percentile of packet latencies and improve the consistency of packet delays in overall 5G systems. Readers are encouraged to refer to [44] and the references therein for further insights. Therefore, in demonstrating factory 5G's potentials, emphasizing future research and development imperatives, and removing the silos between telecom and automation industries, all these aspects are crucial to address. It is critical to discover/develop anchor use cases showing 5G's effectiveness in performance, safety, and ROI. Promising anchor use cases for which factory 5G efforts are notable include network robots/AGVs, vision and augmented reality, an intelligent edge for closed-loop control, and massive wireless sensor networks. For instance, for matrix and additive manufacturing concepts, which require reliability, timeliness, and synchronization for connected mobile robots/AGVs coursing on condition-based dynamic routes instead of predefined ones, the 5G RAN and core enhancements are fitting for low-latency access to edge computing resources.

The remaining challenges include network optimization (coverage and scheduling) under operational dynamics. Additionally, 5G broadband can support vision- and augmentation-related uses cases for virtualization and remote maintenance, respectively, for a limited set of devices. However, in massive connectivity for critical production processes and interactive extended reality for collaborating field agents as well as haptic control, maintaining minimum peak data rates with low latency remains open to further enhancements. Meanwhile, integrated communication, sensing, and localization is still an issue for context-aware human-machine interactivity. For integration with Ethernet/TSN networks, 5G already provides E2E support for scheduled flows, synchronization, and reliability. Still, further work is needed to translate TSN resource allocation and time synchronization within the converged 5G-TSN system in challenging industrial radio-frequency environments.

Spectrum use options and different network deployment models will be decisive in factory 5G's success. In composite deployment models, inputs from all stakeholders, including spectrum authorities, MNOs, and OT/IT technology providers, must be considered while discovering innovative service and business models. For alternative spectrum policies/use options, a cost-benefit analysis is necessary while considering the following:

- Dedicated spectrum is expensive, and ROIs will depend on anchor use cases,
- An MNO-owned spectrum portfolio can provide network scalability and harmonized indoor/outdoor coverage.
- NR-U with URLLC enhancements might be more sustainable for small/medium enterprises joining the digitalization drive.

The isolated private 5G deployment model can enable dedicated support for critical data-intensive and time-sensitive use cases while providing necessary privacy and security features. Meanwhile, PNI private 5G deployment options provide cost-effective shared infrastructure for less stringent applications, e.g., process automation and monitoring/maintenance. However, solution providers need to optimize operation and management complexity and provide autonomous solutions to network operation, optimization, on-site and remote functional splits, new service deployments, and QoS assurances for multiple scenarios and services.

Above all, private 5G wireless networking brings many security concerns to the factory floor, especially when sharing resources with a public network. In this respect, the 3GPP offers solutions to increase the resilience of private networks by duplicating NFs/resources and through service/slice authorization and isolation. To defend against radio jamming attacks, solutions can be devised by combining 5G's dynamic spectrum monitoring and allocation and beamforming techniques. Meanwhile, to reduce the additional communication overhead of security measures, new

techniques, such as physical layer security and federated learning, are promising but require realistic suitability analysis for real-time control and analytics.

Biographies

Aamir Mahmood (aamir.mahmood@miun.se) earned his D.Sc. degree in communications engineering from the School of Electrical Engineering, Aalto University, Finland, in 2014. He is an assistant professor in the Department of Information Systems and Technology, Mid Sweden University, Sundsvall, 851 70, Sweden. His research interests include wireless networks for dependable communication, focusing on intelligent solutions for radio-frequency coexistence, radio resource management, and time synchronization. He is a member of the IEEE Industrial Electronics Society and a Senior Member of IEEE.

Sarder Fakhrul Abedin (sarder.abedin@miun.se) earned his Ph.D. degree in computer engineering from Kyung Hee University, Seoul, South Korea, in 2020. He is an assistant professor in the Department of Information Systems and Technology, Mid Sweden University, Sundsvall, 851 70, Sweden. His research interests include Internet of Things network management, fog computing, machine learning, industrial 5G, and wireless networking. He is a Member of IEEE.

Thilo Sauter (thilo.sauter@tuwien.ac.at) earned his Ph.D. degree in electrical engineering from TU Wien, Vienna, 1040, Austria, in 1999, where he is currently a professor of automation technology. His research interests include smart sensors and automation networks, with a focus on real-time, security, interconnection, and integration issues. He is a senior Administrative Committee member of the IEEE Industrial Electronics Society and a Fellow of IEEE.

Mikael Gidlund (mikael.gidlund@miun.se) earned his Ph.D. degree in electrical engineering from Mid Sweden University, Sundsvall, 851 70, Sweden, in 2005, where he has been a professor of computer engineering since 2015. His research interests include wireless communication and

networks, wireless sensor networks, access protocols, and security. He is an associate editor of *IEEE Transactions on Industrial Informatics*, a member of the IEEE Industrial Electronics Society, and a Senior Member of IEEE.

Krister Landernäs (krister.landernas@se.abb.com) earned his Ph.D. degree in electrical engineering at Mälardalen University, Västerås, Sweden, in 2006. He is currently the technical coordinator of the Horizon 2020 Information and Communication Technologies 2019 project 5G for Smart Manufacturing, ABB Corporate Research, Västerås, 721 71, Sweden. His research interests include industrial wireless communication.

References

- [1] J. Jasperneite, T. Sauter, and M. Wollschlaeger, "Why we need automation models: Handling complexity in industry 4.0 and the Internet of Things," *IEEE Ind. Electron. Mag.*, vol. 14, no. 1, pp. 29–40, 2020, doi: 10.1109/MIE.2019.2947119.
- [2] P. K. Malik *et al.*, "Industrial Internet of things and its applications in industry 4.0: State of the art," *Comput. Commun.*, vol. 166, pp. 125–139, Jan. 2021, doi: 10.1016/j.comcom.2020.11.016.
- [3] D. Ginhör, R. Guillaume, N. Nayak, and J. V. Hoyningen-Huene, "Time-sensitive networking for industrial control networks," in *Wireless Networks and Industrial IoT*, N. H. Mahmood, N. Marchenko, M. Gidlund, and P. Popovski, Eds. Cham: Springer International Publishing, 2021, pp. 39–54.
- [4] I. Rodriguez *et al.*, "5G swarm production: Advanced industrial manufacturing concepts enabled by wireless automation," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 48–54, 2021, doi: 10.1109/MCOM.001.2000560.
- [5] D. Cavalcanti *et al.*, "Extending accurate time distribution and timeliness capabilities over the air to enable future wireless industrial automation systems," *Proc. IEEE*, vol. 107, no. 6, pp. 1132–1152, 2019, doi: 10.1109/JPROC.2019.2903414.
- [6] *Time-Sensitive Networking (TSN) Task Group*, IEEE Standard 802.1. Accessed: Jan. 7, 2022. [Online]. Available: <https://1.ieee802.org/tsn/>
- [7] A. Neumann *et al.*, "5G into Profinet integration as a use case for network slicing," in *Proc. 2019 24th IEEE Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, pp. 1293–1296, doi: 10.1109/ETFA.2019.8869041.
- [8] M. Schüngel, S. Dietrich, D. Ginhör, S.-P. Chen, and M. Kuhn, "Analysis of time synchronization for converged wired and wireless networks," in *Proc. 2020 25th IEEE Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, vol. 1, pp. 198–205, doi: 10.1109/ETFA46521.2020.9212068.
- [9] S. Vitturi, C. Zunino, and T. Sauter, "Industrial communication systems and their future challenges: Next-generation Ethernet, IIoT, and 5G," *Proc. IEEE*, vol. 107, no. 6, pp. 944–961, 2019, doi: 10.1109/JPROC.2019.2913443.
- [10] V. K. L. Huang, Z. Pang, C.-J. A. Chen, and K. F. Tsang, "New trends in the practical deployment of industrial wireless: From noncritical to critical use cases," *IEEE Ind. Electron. Mag.*, vol. 12, no. 2, pp. 50–58, 2018, doi: 10.1109/MIE.2018.2825480.
- [11] A. Mahmood *et al.*, "Industrial IoT in 5G-and-beyond networks: Vision, architecture, and design trends," *IEEE Trans. Ind. Informat.*, early access, 2021, doi: 10.1109/TII.2021.3115697.
- [12] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial Internet of Things: Challenges, opportunities, and directions," *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4724–4734, 2018, doi: 10.1109/TII.2018.2852491.
- [13] A. W. Colombo, S. Karnouskos, O. Kaynak, Y. Shi, and S. Yin, "Industrial cyberphysical systems: A backbone of the fourth industrial revolution," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 6–16, 2017, doi: 10.1109/MIE.2017.2648857.
- [14] T.-K. Le, U. Salim, and F. Kaltenberger, "An overview of physical layer design for ultra-reliable low-latency communications in 3GPP releases 15, 16, and 17," *IEEE Access*, vol. 9, pp. 433–444, 2021, doi: 10.1109/ACCESS.2020.3046773.
- [15] M. Attaran, "The impact of 5G on the evolution of intelligent automation and industry digitization," *J. Ambient Intell. Humanized Comput.*, pp. 1–17, Feb. 2021, doi: 10.1007/s12652-020-02521-x.
- [16] M. Gundall *et al.*, "Introduction of a 5G-enabled architecture for the realization of Industry 4.0 use cases," *IEEE Access*, vol. 9, pp. 25,508–25,521, Feb. 2021, doi: 10.1109/ACCESS.2021.3057675.
- [17] "5G non-public networks for industrial scenarios," 5G ACIA, Jul. 2019. <https://www.gsma.com/iot/resources/5g-non-public-networks-for-industrial-scenarios/>
- [18] A. Ajiz, "Private 5G: The future of industrial wireless," *IEEE Ind. Electron. Mag.*, vol. 14, no. 4, pp. 136–145, 2020, doi: 10.1109/MIE.2020.3004975.
- [19] "Study on communication for automation in vertical domains (Release 16)," 3GPP, Sophia Antipolis, France, 3GPP TR 22.804, Jul. 2020. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/22_series/22.804/
- [20] "Service requirements for cyber-physical control applications in vertical domains," 3GPP, Sophia Antipolis, France, 3GPP TS 22.104, Sep. 2020. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/22_series/22.104/
- [21] M. Seitz, F. Gehlhoff, L. A. Cruz Salazar, A. Fay, and B. Vogel-Heuser, "Automation platform independent multi-agent system for robust networks of production resources in industry 4.0," *J. Intell. Manuf.*, vol. 32, pp. 2023–2041, Apr. 2021, doi: 10.1007/s10845-021-01759-2.
- [22] T. Hoeschele, C. Dietzel, D. Kopp, F. H. Fitzek, and M. Reisslein, "Importance of Internet Exchange Point (IXP) infrastructure for 5G: Estimating the impact of 5G use cases," *Telecommun. Policy*, vol. 45, no. 3, pp. 102,091, 2021, doi: 10.1016/j.telpol.2020.102091.
- [23] A. Mahmood, M. I. Ashraf, M. Gidlund, J. Torsner, and J. Sachs, "Time synchronization in 5G wireless edge: Requirements and solutions for critical-MTC," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 45–51, 2019, doi: 10.1109/MCOM.001.1900379.
- [24] O. Seijo Gomez, I. Val, M. Luvisotto, and Z. Pang, "Clock synchronization for wireless time-sensitive networking: A march from microsecond to nanosecond," *IEEE Ind. Electron. Mag.*, early access, 2021, doi: 10.1109/MIE.2021.3078071.
- [25] H. Shi, A. Ajiz, and N. Jiang, "Evaluating the performance of over-the-air time synchronization for 5G and TSN integration," in *Proc. 2021 IEEE Int. Black Sea Conf. Commun. Netw. (BlackSeaCom)*, pp. 1–6, doi: 10.1109/BlackSeaCom52164.2021.9527833.
- [26] M. Schungel, S. Dietrich, D. Ginhör, S.-P. Chen, and M. Kuhn, "Heterogeneous synchronization in converged wired and wireless time-sensitive networks," in *Proc. 7th IEEE Int. Conf. Factory Commun. Syst. (WFCS)*, 2021, pp. 67–74, doi: 10.1109/WFCS46889.2021.9483592.
- [27] M. Angjelichinoski, K. F. Trillingsgaard, and P. Popovski, "A statistical learning approach to ultra-reliable low latency communication," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5153–5166, 2019, doi: 10.1109/TCOMM.2019.2907241.
- [28] A. Ajiz, "Packet duplication in dual connectivity enabled 5G wireless networks: Overview and challenges," *IEEE Commun. Standards Mag.*, vol. 3, no. 3, pp. 20–28, 2019, doi: 10.1109/MCOMSTD.001.1700065.
- [29] "System architecture for the 5G system (Release 17)," 3GPP, Sophia Antipolis, France, 3GPP TS 23.501, Mar. 2021. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.501/
- [30] "Procedures for the 5G system (5GS) (Release 17)," 3GPP, Sophia Antipolis, France, 3GPP TS 23.502, Mar. 2021. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.502/
- [31] M. Norin, R. Högman, M. Buchmayer, G. Lemme, F. Pedersen, and A. Zaidi, "5G spectrum for local industrial networks," Ericsson, Stockholm, Sweden, White Paper, Jun. 2020. [Online]. Available: <https://www.ericsson.com/en/reports-and-papers/white-papers/5g-spectrum-for-local-industrial-networks>
- [32] P. Yang, L. Kong, and G. Chen, "Spectrum sharing for 5G/6G URLLC: Research frontiers and standards," *IEEE Commun. Standard Mag.*, vol. 5, no. 2, pp. 120–125, 2021, doi: 10.1109/MCOMSTD.001.2000054.
- [33] S. Grimaldi, A. Mahmood, S. A. Hassan, M. Gidlund, and G. P. Hancke, "Autonomous interference mapping for industrial Internet of things networks over unlicensed bands: Identifying cross-technology interference," *IEEE Ind. Electron. Mag.*, vol. 15, no. 1, pp. 67–78, 2021, doi: 10.1109/MIE.2020.3007568.
- [34] A. A. Barakat, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Comput. Netw.*, vol. 167, p. 106,984, Feb. 2020, doi: 10.1016/j.comnet.2019.106984.
- [35] S. Wijethilaka and M. Liyanage, "Survey on network slicing for Internet of Things realization in 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 957–994, 2021, doi: 10.1109/COMST.2021.3067807.
- [36] S. F. Abedin, M. S. Munir, N. H. Tran, Z. Han, and C. S. Hong, "Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5994–6006, 2021, doi: 10.1109/TITS.2020.3039617.
- [37] "Network sharing: Architecture and functional description," 3GPP, Sophia Antipolis, France, 3GPP TS 23.251, Jul. 2020. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.251/
- [38] "Security aspects of 5G for industrial networks," 5G ACIA, Frankfurt am Main, Germany, White Paper, Feb. 2021. [Online]. Available: <https://5g-acia.org/whitepapers/security-aspects-of-5g-for-industrial-networks/>
- [39] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 2020, doi: 10.1109/COMST.2020.2965856.
- [40] S. Grimaldi, L. Martenvormfelde, A. Mahmood, and M. Gidlund, "Onboard spectral analysis for low-complexity IoT devices," *IEEE Access*, vol. 8, pp. 43,027–43,045, Mar. 2020, doi: 10.1109/ACCESS.2020.2977842.
- [41] C. Schwaiger and T. Sauter, "Security strategies for field area networks," in *Proc. IEEE 2002 28th Annu. Conf. Ind. Electron. Soc.*, vol. 4, pp. 2915–2920, doi: 10.1109/IECON.2002.1182859.
- [42] "Security architecture and procedures for 5G system (Release 17)," 3GPP, Sophia Antipolis, France, 3GPP TS 23.251, 3GPP TS 33.501, Mar. 2021. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/33_series/33.501/
- [43] N. I. Mowla, N. H. Tran, I. Doh, and K. Chae, "AFRL: Adaptive federated reinforcement learning for intelligent jamming defense in FANET," *J. Commun. Netw.*, vol. 22, no. 3, pp. 244–258, 2020, doi: 10.1109/JCN.2020.000015.
- [44] J. Rischke, P. Sossalla, S. Itting, F. H. P. Fitzek, and M. Reisslein, "5G campus networks: A first measurement study," *IEEE Access*, vol. 9, pp. 121,786–121,803, Aug. 2021, doi: 10.1109/ACCESS.2021.3108423.



*A March From Microsecond
to Nanosecond*

ÓSCAR SEIJO, IÑAKI VAL,
MICHELE LUVISOTTO,
and ZHIBO PANG

Clock Synchronization for Wireless Time- Sensitive Networking

Industrial control systems in the era of Industry 4.0 present significant challenges from a communication perspective, that is, low latency, ultrahigh reliability, and accurate synchronization. Time-sensitive networking (TSN) has emerged as the main solver of these challenges. As a research trend, TSN and wireless technologies are expected to converge in the wireless TSN para-

digm. This convergence starts with the adoption of accurate clock synchronization over wireless systems. In this article, we review the existing wireless clock-synchronization approaches and their attainable performances, and we discuss their feasibility to enable wireless TSN. We conclude that the existing clock-synchronization techniques are enough to enable wireless TSN, although significant implementation efforts are required to incorporate accurate clock synchronization over wireless systems.

The Vision of Wireless TSN

As a promising direction of wireless communications for industrial automation, a vision of wireless TSN was brought up as early as 2018 [1]. Our vision of wireless TSN is to achieve a wireless solution that directly supports the TSN functionalities and that provides the equivalent performance of the wired TSN by wireless communications, thus acquiring all the advantages inherent to wireless, such as better scalability, lower commissioning costs, and free movement

Digital Object Identifier 10.1109/MIE.2021.3078071
Date of current version: 28 May 2021

Our vision of Wireless TSN is to achieve a wireless solution that directly supports the TSN functionalities and that provides the equivalent performance of the Wired TSN by wireless communications.

of the wireless nodes. The envisioned wireless network presents a point-to-multipoint architecture (i.e., Wi-Fi/5G like) where a wireless TSN master is, on the one hand, wiredly connected to a wired TSN network, and on the other hand extends TSN to multiple stations in the wireless domain. Two challenges in terms of clock synchronization can be highlighted to achieve this vision. First, the wireless network must accurately synchronize the local clock of all the nodes in the wireless network to the same common time using wireless signals (referred to as *Wi-Sync*) [2]. Second, the bridge among the domains must ensure the seamless extension of the synchronization from wired to wireless TSN.

In this article, we first summarize the requirements of Wi-Sync for wireless TSN, then we point out the fundamental limits that lie in attainable synchronization. The latest progresses in the state of the art are presented based on the technologies involved and the levels of performance, ranging from microseconds to subnanosecond. Based on the validated feasibilities, we appeal for more research and implementation efforts on Wi-Sync and also on the integration of Wi-Sync and wired TSN clock synchronization. Finally, we briefly outline some other critical features required to enable wireless TSN, which are outside the scope of the article, such as securing the wireless communication, wireless time-aware scheduling, and time-aware scheduling, and orchestration among wired/wireless domains. Compared to the previous works that analyze 5G [3], [4] and Wi-Fi Wi-Sync capabilities [5], this work presents the wireless TSN Wi-Sync requirements, a detailed overview of the current state of research, and the existing

techniques used to enable accurate Wi-Sync over Wi-Fi and 5G and its integration with TSN clock synchronization.

Clock-Synchronization Requirements for Wireless TSN

From the user's point of view, a strategic requirement of Wi-Sync for wireless TSN is to make the solution self-contained, i.e., the Wi-Sync functionalities shall be realized by the wireless TSN itself, without dependency on external infrastructure. Consequently, common solutions relying on global navigation satellite systems (GNSS) are outside the scope of this research. A self-contained wireless TSN solution is desirable because it has fewer per-device costs and complexities, it has better deployment flexibility (the devices can be deployed indoor or outdoor), and it is not vulnerable to GNSS spoofing [6]. In addition, another strategic requirement is the integration of the synchronization between wired and wireless TSN to achieve a large-scale internetwork clock synchronization.

The required Wi-Sync performance for wireless TSN is mainly derived from three sources: the requirements inherited from wired TSN, the requirements of the specific wireless technology, and the requirements for wireless localization.

- *Baseline requirements inherited from wired TSN:* Communication networks used at the industrial field level have long been offering mechanisms to synchronize the local clock of network nodes to a master clock, such as the distributed clock mechanism in EtherCAT [7]. With the advent of TSN, such mechanisms are being embedded in the Ethernet standard itself [8]. The required level of clock-synchronization accuracy of TSN varies across

applications. The IEC/IEEE 60802 *TSN Profile for Industrial Automation* sets it between 100 ns (e.g., for motion control) and 1 μ s (e.g., for process control) [9], and these values are in line with many example applications from the scientific literature [10] for applications that involve both movable and static nodes. These requirements are independent of the adopted wireless/wired technology. For instance, these requirements are considered by the 5G-ACIA and the 3rd Generation Partnership Project in the integration of 5G and TSN [11].

- *Wireless communications protocol-specific requirements:* Clock synchronization is also a precondition to allow for coordinated access to the shared wireless channel in a time-division multiple access fashion. The required accuracy, in this case, is determined by the required length of time slot or minimal transmission interval. The rule of thumb is that the synchronization error should not exceed 1/10 (baseline) or 1/20 (desirable) of the slot length. The synchronization requirement has been considered in several mobile standards, such as LTE and 5G, and in 802.11ax to allow for coordinated access to the medium. For instance, a few-microsecond accuracy is required if the slot length is in the 100- μ s level in the real-time Wi-Fi [12], 1 μ s is required in 5G [4], and 100-ns level accuracy is required if the slot length is a few microseconds in the wireless high-performance (WirelessHP) [10] or wireless-SHARP [13].
- *Requirements for wireless localization:* Many industrial applications require accurate asset localization and tracking. As an industrial trend, wireless communication systems are expected to converge with indoor localization systems, typically based on measuring the time difference of arrival (TDoA) [14]. The accuracy of TDoA strongly depends on the synchronization performance, e.g., a localization accuracy of 1 m requires a synchronization accuracy of 3.3 ns for

a rough estimation. These applications introduce the most stringent requirement for Wi-Sync, obviously.

Key Factors of Accuracy in Wi-Sync

Wi-Sync accuracy resides on three processes: the protocol used to exchange synchronization messages between one master and multiple slaves (messaging), the egress and ingress time of the messages (timestamps) used to compute the Wi-Sync error, and the Wi-Sync error calculation and local time adjustment (tuning). The accuracy of a certain Wi-Sync solution is determined by the techniques applied in these processes.

Messaging Protocols

Several messaging protocols are being considered to enable Wi-Sync in wireless TSN. The feasibility of each of them mainly depends on the coverage range of the wireless network, the wireless technology, and the Wi-Sync performance required by the network and applications. In this section, we summarize the messaging protocols that are currently being considered to

As an industrial trend, wireless communication systems are expected to converge with indoor localization systems, typically based on measuring the time difference of arrival.

enable Wi-Sync for wireless TSN over Wi-Fi and 5G.

Precision Time Protocol (PTP) has been defined by the 802.1AS standard for TSN time synchronization over Ethernet TSN thanks to its simplicity and performance in the 10-ns range over adequate Ethernet hardware (HW). Due to its popularity, PTP is also considered by several researchers as a natural Wi-Sync solution over Wi-Fi [5]. The protocol is based on a three-frame exchange, which enables channel-delay compensation (see “PTP Scheme” in Figure 1).

The 802.11 timing measurement (TM) and fine timing measurement (FTM) schemes [15] exploit the inherent structure of Wi-Fi transmissions based on data frame+Acknowledgment to perform the clock synchronization

(see “FTM Scheme” in Figure 1). The FTM is an evolution of TM that includes better timestamping resolution granularity (from 10 to 0.1 ns) and some extra functionalities, e.g., synchronization bursts, where several synchronization frames are transmitted in a row, enhancing the Wi-Sync precision. The FTM and PTP schemes have a similar performance range as both are based on a similar frame exchange. The TM and FTM exchanges have been adopted by the 802.11AS standard (the 2012 and 2020 revisions, respectively [2]) for TSN clock synchronization over Wi-Fi.

The 5G New Radio (NR) synchronization scheme exploits the current existing radio synchronization mechanisms of 5G to enable clock synchronization (see “5G NR Sync Scheme” in

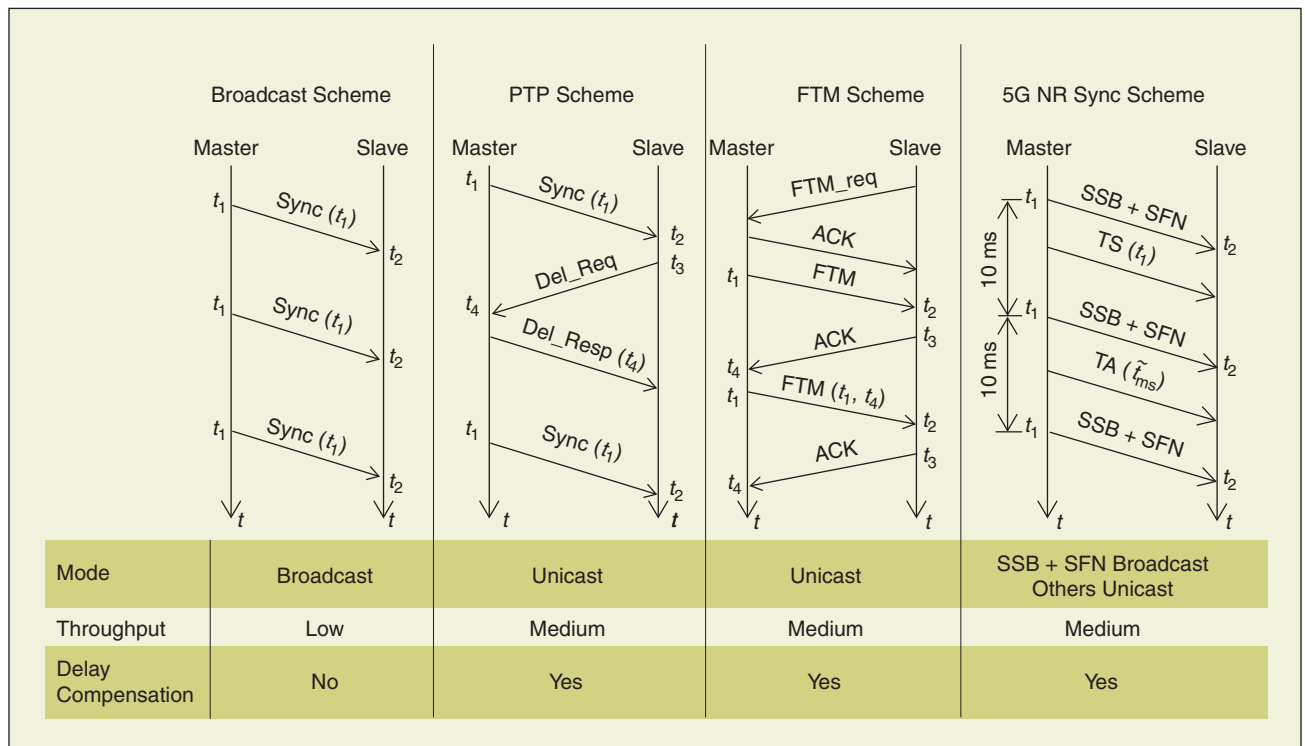


FIGURE 1 – The messaging protocols developed for Wi-Sync. FTM: fine timing measurement; ACK: acknowledgment; SSBs: synchronization signal blocks; SFN: system frame number; TS: timestamp.

The SFN is an identifier that univocally identifies the transmission of an SSB.

Figure 1) [4]. A 5G base station (BS) periodically transmits synchronization signal blocks (SSBs), which are used to synchronize the user equipment (UE) devices within the 5G frame. Then the 5G BS can transmit a time signaling, which includes the egress timestamp of an SSB with a specific system frame number (SFN) that was sent. The SFN is an identifier that univocally identifies the transmission of an SSB. In addition, the 5G BS can also compute and send to the UEs the channel delay [(\hat{t}_{ms})] as in Figure 1 “5G NR Sync Scheme” timing advance (TA) [(\hat{t}_{ms})] through the timing advance mechanism. With this information, the UEs can perform an accurate clock synchronization to the BS. Besides, the TSN clock synchronization over 5G has been already defined [4].

After the 5G network is clock synchronized, the 5G system acts from a TSN perspective as a transparent bridge. Essentially, the TSN devices connected to the 5G network UE and BS ends transmit PTP frames through the user data plane of the 5G network. The 5G network calculates the total residence time of the PTP frames in the 5G network and appends the residence time to the PTP frames (i.e., a transparent clock). Therefore, the TSN end devices are able to synchronize their local clock without noticing that a 5G network is in the middle of the TSN network routing the PTP

frames between the master clock and the slaves.

Finally, even though the presented schemes are robust and require little bandwidth, a simplified messaging scheme based on the broadcasting of sync frames is commonly used in wireless [5] (see “Broadcast Scheme” in Figure 1). The broadcast scheme has a clear advantage: As it is based on the periodic broadcasting of synchronization frames, its throughput is independent of the number of devices connected to the network. On the downside, this scheme cannot compensate for the channel delay.

The feasibility of the broadcast scheme depends on the coverage range of the wireless network and the required Wi-Sync. Theoretically, the broadcast scheme can be used when the required Wi-Sync equals the distance between the master and the farthest slave divided by the speed of light. However, wireless impairments (e.g., the multipath and fading) can introduce imprecisions in the Wi-Sync. As a rule of thumb, the broadcast scheme can be recommended in networks with a maximum Wi-Sync equal to the double of the distance between the wireless TSN master and the farthest wireless TSN slave divided by the speed of light. For instance, the broadcast scheme is feasible in a network with a maximum Wi-Sync error of 500 ns and a deployment range below 75 m.

Timestamping and Tuning

Depending on the layer of the communication stack where the signal or data processing happens and the way of implementing the clock and calculation, the approaches for timestamping and tuning are organized into four classes (A, B, C, and D), which are summarized in Table 1. Each class results in largely different levels of performance even using the same messaging protocol and the same wireless system. Classes A and B use a pure software (SW) approach, where timestamps, the clock, and calculations are performed in SW in a PTP daemon [see Figure 2(a)], and classes C and D replace the SW clock with a HW one, although the Wi-Sync calculations are still performed in SW [see Figure 2(b)].

Class A: SW in the Application Layer

In class A, timestamps are taken when a clock-synchronization message leaves or arrives at the application layer. Class A suffers from significant jitter, which comes from various sources, including random delays of the application and network stack depending on the network and processor load, random medium access time, and wireless propagation phenomena. Class A is out of the pursued requirements, providing performances on the order of milliseconds [16] (see Table 1).

Class B: SW in the Device-Driver Layer

Class B performs timestamps in the interruptions of the network interface driver—very close to the HW—although it keeps the SW tuning. These timestamps are affected only by the jitter of the interruption routine used

TABLE 1 – A COMPARISON OF THE APPROACHES OF TIMESTAMPING AND TUNING FOR WI-SYNC.

CLASS	LAYER OF PROCESSING	TUNING	PROS	CONS	PRECISION
A: SW in application layer	APP	SW	Commercial off-the-shelf HW, simple	Low performance	1 ms
B: SW in media access control (MAC) layer	MAC	SW	Commercial off-the-shelf HW	Affected by network and processor loads	0.5–10 μ s
C: HW in the PHY layer, baseband	PHY baseband	HW	Stable, high performance	Custom HW	20–100 ns
D: HW in the PHY layer, baseband-radio frequency (RF)	PHY baseband and RF	HW	Stable, highest performance	Custom HW and complex implementation	<5 ns

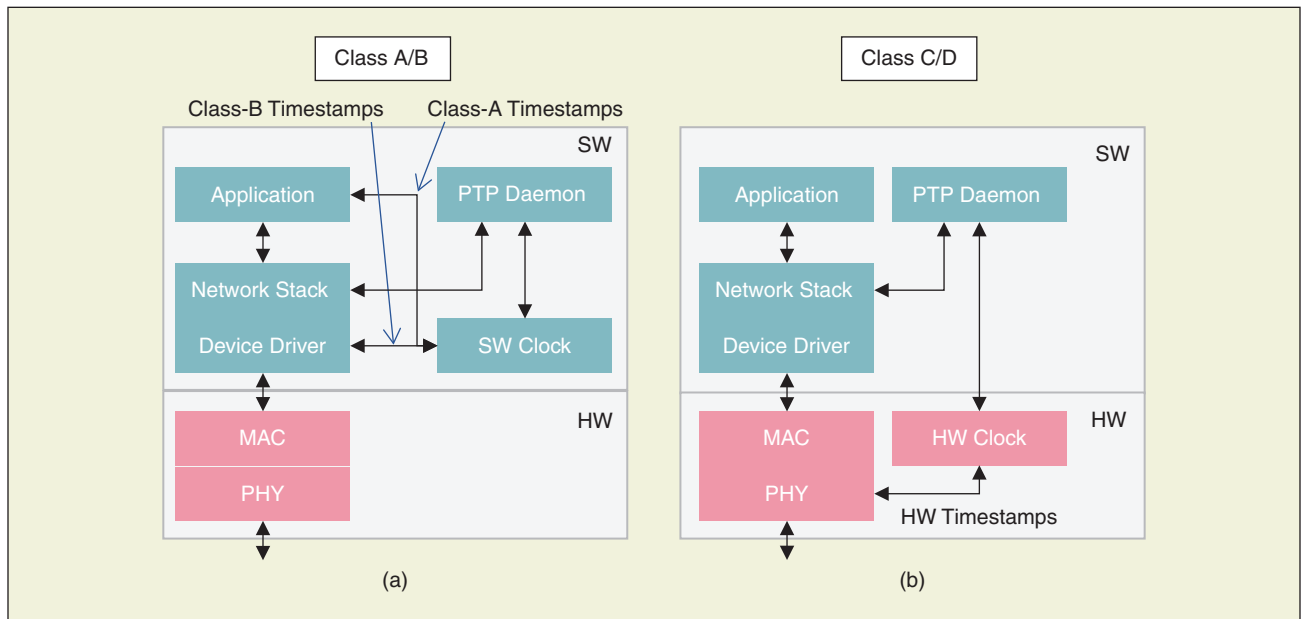


FIGURE 2 – The network stack for (a) devices with class-A/B timestamping and tuning and (b) devices with class-C/D timestamping and tuning. PHY: physical layer; MAC: media access control.

to take timestamps. To provide an adequate performance, class-B timestamps require deep knowledge of the wireless system, processor architecture, operating system, and driver itself. Class B provides a performance range on the order of 1–10 μs [17].

Class C: HW in the Physical Layer, Baseband

In class C, timestamps are directly taken at the exact moment of departure and detection of the frames from/to the baseband processor [physical (PHY) layer] of the network interface. The PHY timestamps avoid the uncertainty of the communication stack, although they do not consider the radio-frequency (RF) properties of the wireless system (e.g., the bandwidth and radio impairments) and the wireless propagation phenomena. Consequently, class-C precision is bounded by the sampling resolution of the baseband processor and by the wireless propagation impairments (the communication noise, multipath, and wireless channel asymmetries). Regarding the tuning technique, class C uses a tunable HW clock that can be adjusted in time and frequency. The class-C performance range, combined with PTP over Wi-Fi, is on the order of tens of nanoseconds [18].

Class D: HW in the PHY Baseband RF Layer

Finally, class D includes the RF chain and the wireless propagation phenomena in the timestamping model. The precision of class-D timestamps

is constrained only by the physical properties of the propagation environment and the RF chain impairments, and its precision can range from a few nanoseconds to subnanoseconds. Class-D research is mainly

TABLE 2 – THE WIRELESS CLOCK-SYNCHRONIZATION APPROACHES.

REFERENCE	YEAR	VALIDATION	WIRELESS SYSTEM	MESSAGING PROTOCOL	CLASS OF TIMESTAMPING AND TUNING	ACCURACY
[16]	2009	Experiment	802.11	NTP	A	1 ms
[17]	2018	Experiment	802.15.4	Custom	B	2–5 μs
[20]	2014	Experiment	802.11	PTP	B	0.5–3 μs
[18]	2018	Experiment	802.11	PTP	C	30–50 ns
[13]	2020	Experiment	w-SHARP, WirelessHP	Broadcast scheme	C	30–50 ns
[21]	2017	Experiment	Ultrawideband radio (LR-WPAN)	PTP	C	1–5 ns
[4]	2020	Technology analysis	5G	5G NR sync scheme + PTP	C	~1 μs
[19]	2012	Experiment	802.11b	PTP	D	<1 ns over specific channels
[22], [23]	2013, 2014	Experiment	802.15.4	PTP	D	<1 ns over specific channels
[24]	2020	Simulations, experiments	802.11	PTP	D	<1 ns

WPAN: wireless personal area network; NTP: network time protocol.

In PI filtering, clock-synchronization error samples are introduced into a PI filter, which reduces the noise variance and estimates the local frequency drift.

focused on localization applications. As an example, the work in [19] presents a class-D timestamping and tuning technique optimized for wireless communications for low-multipath conditions that provides subnanosecond Wi-Sync performance.

The tuning process also involves a noise-reduction technique that enhances Wi-Sync precision. The noise-reduction techniques are common to every class of timestamping and tuning. Two techniques are commonly adopted in standard PTP implementations: proportional integral (PI) filtering and linear regression. In PI filtering, clock-synchronization error samples are introduced into a PI filter, which reduces the noise variance and estimates the local frequency drift. In linear regression, a set of clock-synchronization error samples are used to estimate the start and slope of a line, which represents the clock-synchronization error and local frequency drift. In general, both offer similar performance, although PI filtering typically provides a faster response to variations in the channel delay/frequency drift, whereas linear regression is more stable.

Summary and Fundamental Limits

Throughout this section, we described the two fundamental choices that limit the attainable Wi-Sync performance: the messaging protocol and the timestamping and tuning technique.

The preference in the messaging protocol is mainly derived from the maximum tolerable Wi-Sync error of the network and its deployment range, but other aspects, such as implementation complexity and the amount of traffic, can be also considered. For wireless TSN operation, the broadcast scheme is a convenient choice for networks deployed at the edge in

small- or medium-sized scenarios. On the contrary, if the deployment scenario is large or if wireless TSN is used as a bridge between two wired TSN networks, a messaging with channel-delay compensation could be more convenient. Regarding localization, its Wi-Sync requirement is on the order of a few nanoseconds or subnanoseconds, and thus, a messaging scheme with channel compensation is compulsory.

When it comes to the class of timestamping and tuning, class A is totally out of the submicrosecond-pursued requirements and must be discarded to achieve our wireless TSN vision. Class B provides precision in the microsecond even submicrosecond range, which is not enough for high-performance wireless TSN configurations. Class C is a reasonable option for wireless TSN because it provides the best implementation complexity/timestamping performance tradeoff. Finally, class D is required only in industrial use cases that require precise localization.

To summarize, class-C timestamping and tuning combined with an appropriate messaging exchange based on the coverage range of the wireless network seems to be a convenient choice for wireless TSN based on the requirements inherited from wired TSN (0.1–1 μ s), whereas localization, with a Wi-Sync requirement in the 1-ns range, needs class-D timestamping and tuning combined with a messaging scheme with channel-delay compensation.

Progresses of Wi-Sync for Wireless TSN

Wireless TSN precision constraints cannot be met using class-A timestamping and tuning. Consequently, our review of the different Wi-Sync

solutions starts with class B. As the first example of Wi-Sync using class B, Lennvall et al. [17] present the implementation of a communication system with Wi-Sync support specifically designed for industrial wireless sensor networks. This scheme considers multihop communications and provides a synchronization precision in the range of 1–2 μ s for a single hop and in the range of 10 μ s for up to five hops. Several authors (e.g., in [20]) have also explored class-B timestamps over Wi-Fi, demonstrating a Wi-Sync in the range of 0.5 μ s. However, the Wi-Sync performance of these solutions is affected by the network and processor loads, and consequently, they suffer from low stability.

Submicrosecond Wi-Sync is feasible using optimized class-B implementations even though HW timestamping and tuning (classes C and D) greatly outperform SW ones and are widely recommended to conduct the wireless TSN research and implementation. Unfortunately, few commercial wireless cards include HW timestamping and HW clock support and, consequently, custom HW solutions over field programmable gate arrays (FPGAs) are the only alternative to implement class-C and D timestamping and tuning.

Seijo et al. [18] developed a Wi-Fi modem with a HW clock and class-C timestamping and tuning built on an FPGA platform. The attainable Wi-Sync using PTP is evaluated over different wireless channels run in a wireless channel emulator, thus obtaining the results over known and predictable conditions (see Figure 3). The results showed a Wi-Sync performance ranging from tens of nanoseconds to 100 ns depending on the wireless environment. Additionally, the authors also studied the Wi-Sync performance of class C combined with the broadcast messaging scheme. To this end, they developed this scheme on top of w-SHARP [13], and they found that the Wi-Sync accuracy was on the order of 50 ns [18] for low-range deployments. Regarding TSN synchronization over 5G NR, the analysis presented in [4] shows that the 5G NR Wi-Sync

accuracy using class-C timestamping and tuning could be below $1 \mu s$, which may be enough for a wide range of wireless TSN use cases. However, no real measurements that demonstrate this performance level over 5G have been presented.

Research on class-D timestamping and tuning has been pushed by interest in precise asset localization in industrial applications and other domains. As the first example of class D, Exel [19] presents an 802.11b modem implemented on an FPGA with a HW 100-ps precision timestamping unit that exploits the specific 802.11b modulation structure, which provides Wi-Sync in the nanosecond range. Similar solutions have been presented for other wireless technologies, such as an implementation of 802.15.4 Chirp Spread Spectrum PHY with class C for ranging purposes [22], [23]. The results were in line with the results obtained with 802.11b. The drawback of these solutions is that they require nearly ideal wireless conditions to maximize their Wi-Sync accuracy. Seijo et al. [24] developed a class-D timestamping technique that overcomes the wireless channel impairments and that can deliver subnanosecond timestamping precision in virtually any wireless condition (see Figure 4).

This timestamping unit combined with PTP and built on top of Wi-Fi has been proven to provide subnanosecond Wi-Sync performance even in high-mobility conditions without a direct line of sight. Finally, IEEE 802.11az standard [25] (802.11 amendment:

enhancements for positioning) also includes a class-D timestamping technique based on the estimation of the carrier phase shift among the communicating nodes. However, there are no works yet available that analyze its capabilities and performance.

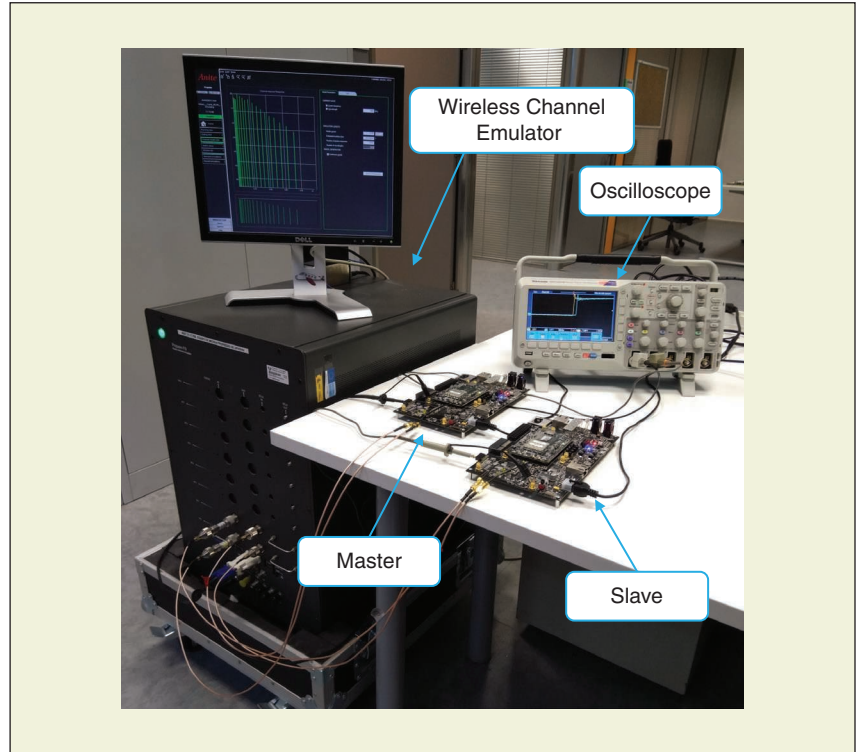


FIGURE 3 – The HW testbed used to evaluate different clock-synchronization schemes against realistic conditions [18].

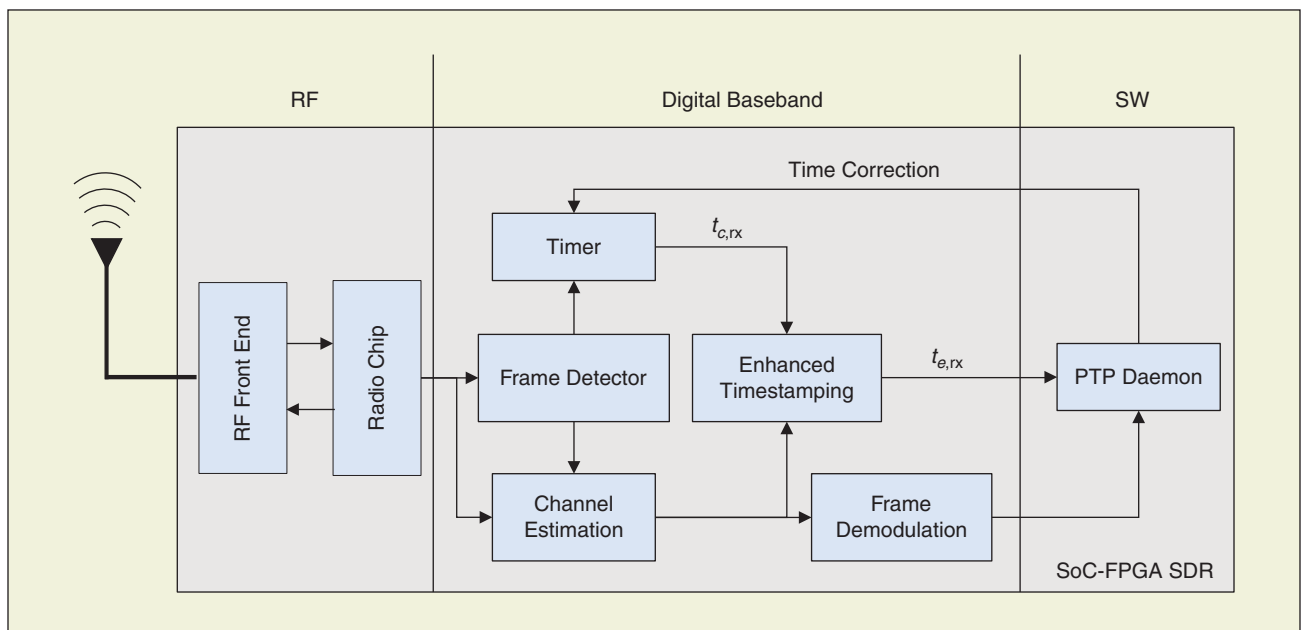


FIGURE 4 – The block diagram of a Wi-Fi receiver with subnanosecond timestamping precision. SoC-FPGA SDR: system-on-chip-field-programmable gate array SW-defined radio.

Another crucial research track is the integration of wired TSN and wireless TSN to create a large-scale hybrid TSN network.

Open Challenges and Future Directions

Wireless clock synchronization will play a key role in the coming years as the main enabler of wireless TSN. We have shown that several standards currently support Wi-Sync mechanisms and protocols, although to massively enable Wi-Sync, several research and implementation challenges are still wide open. We think that the future research tracks described in the following paragraphs should conduct the next steps in wireless clock synchronization.

In the implementation domain, only a few Wi-Fi devices already include class-C timestamping and, to the best of our knowledge, none of them supports class-D timestamping. We expect that the 802.11az standard will be adopted in next-generation Wi-Fi devices for localization and that the adoption will also enhance Wi-Sync performance. Also, a HW clock is required to maximize Wi-Sync stability and precision, and they are not available in commercial devices at the moment. Finally, the drivers of wireless network interface cards must provide a standard interface to access the PHY timestamps and the HW clock.

Another crucial research track is the integration of wired TSN and wireless TSN to create a large-scale hybrid TSN network. Several challenges must be solved to achieve this integration (e.g., clock-synchronization integration and scheduling translation). Despite the fact that some steps have already been performed with wired TSN and narrow-band wireless protocols [26], far more research, standardization, and implementation efforts are required to achieve high-performance hybrid TSN capabilities.

Finally, we think that subnanosecond Wi-Sync can present new ways to overcome other challenging obstacles in industrial wireless communications.

For instance, wireless security [27] is a very relevant matter of study as wireless control systems are potentially vulnerable to attacks. In that sense, ultra-accurate Wi-Sync could be used in combination with the estimation of the channel state to perform PHY-layer authentication and to decide whether a frame comes from a legacy or a malicious node [28]. Hence, we encourage the research community to study the use of novel Wi-Sync techniques as a promising way to solve both direct and indirect problems in the wireless TSN research trend.

Conclusion

This article discussed the need for accurate clock synchronization in high-performance wireless networks with TSN capabilities. We began the article by stating the requirements of wireless TSN and the different compelling applications enabled by high-performance clock synchronization. We presented the various existing technologies that can be used to facilitate synchronization with different levels of accuracy, from the microsecond level to less than 1 ns. Finally, we showed that the performance of the most advanced techniques is enough to satisfy the most demanding requirements found in our wireless TSN vision.

However, commercial wireless cards show a modest adoption of Wi-Sync capabilities, and hence the options to implement accurate Wi-Sync are reduced to custom solutions based on programmable HW (i.e., FPGA) platforms. We expect that the 802.11be standard (marketed as Wi-Fi 7) and the next 5G releases will adopt some of the required technologies to enable our wireless TSN vision, including accurate Wi-Sync, time-aware scheduling, scheduling orchestration with wired TSN, power-consumption optimization, and ultrahigh reliability.

Finally, we also expect the unification of different wired/wireless protocols into the TSN technology, leading to a communication media-independent TSN paradigm.

Acknowledgments

The work of Iñaki Val and Óscar Seijo is partially supported by the B-INDUSTRY5G (ELKARTEK) project of the Basque Government (Spain). Zhibo Pang's work is partially funded by the Swedish Foundation for Strategic Research through project number APR20-0023. The authors would like to thank the anonymous reviewers for their valuable comments that resulted in a significant improvement of the article. The corresponding author is Óscar Seijo.

Biographies

Óscar Seijo (oseijo@ikerlan.es) earned his Ph.D. degree in telecommunications engineering from the University of Oviedo, Oviedo, Spain, in 2021. Currently, he is a researcher with the Communication Systems Group of Ikerlan, Mondragón, 20500, Spain. During his Ph.D., he worked on the design and development of clock-synchronization techniques for wireless time-sensitive networking (TSN) communication systems. His current research interests include high-performance physical/medium access control layer design for wireless TSN communication system, wireless clock synchronization, and digital signal processing.

Iñaki Val (ival@ikerlan.es) earned his Ph.D. degree from the Department of Signals, Systems and Radio-communication at the Polytechnic University of Madrid, Madrid, Spain, in 2011. Since 2001, he has been with the Communications Systems Group of Ikerlan, Mondragón, 20500, Spain and, currently, he is the team leader of the group. Previously, he was with Fraunhofer IIS of Erlangen (Germany) as an invited researcher (2005–2006). His research activities include the design and implementation of digital wireless communications systems for industrial communications and digital signal processing. He is a member of

the IEEE Industrial Electronics Society and a Senior Member of IEEE.

Michele Luvisotto (michele.luvisotto@hitachi-powergrids.com) earned his Ph.D. degree in information engineering from the University of Padova, Padova, Italy, in 2017. He joined ABB Corporate Research Center, Västerås, SE72219, Sweden, in 2017, where he has led and taken part in several research projects aimed at high-performance wireless communications for industrial applications. Since 2020, he has been with Hitachi ABB Power Grids, where he currently manages a research team working on digital technologies for electrical power grids. He has been a guest editor for *IEEE Transactions on Industrial Informatics* and *IEEE Access*. His research interests include low-latency and ultrareliable wireless communications, the Industrial Internet of Things, smart grids, and power systems digitalization.

Zhibo Pang (pang.zhibo@se.abb.com) earned his Ph.D. degree in electronic and computer systems from KTH Royal Institute of Technology, Stockholm, Sweden, in 2013. He is currently a senior principal scientist at ABB Corporate Research Sweden, Västerås, SE72178, Sweden; an adjunct professor at the University of Sydney, Sydney, Australia; and affiliated faculty at KTH Royal Institute of Technology. He is an associate editor of *IEEE Transactions on Industrial Informatics*, *IEEE Journal of Biomedical and Health Informatics*, and *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*. He is a member of the IEEE Industrial Electronics Society and a Senior Member of IEEE.

References

[1] "Industrial wireless time-sensitive networking: RFC on the path forward," Avnu Alliance, Beaverton, OR, White Paper, 2018. Accessed: Feb. 04, 2020. [Online]. Available: <https://avnu.org/wp-content/uploads/2014/05/Industrial-Wireless-TSN-Roadmap-v1.0.3-1.pdf>

[2] D. Cavalcanti, J. Perez-Ramirez, M. M. Rashid, J. Fang, M. Galeev, and K. B. Stanton, "Extending accurate time distribution and timeliness capabilities over the air to enable future wireless industrial automation systems," *Proc. IEEE*, vol. 107, no. 6, pp. 1132–1152, 2019. doi: 10.1109/JPROC.2019.2903414.

[3] A. Mahmood, M. I. Ashraf, M. Gidlund, J. Torsner, and J. Sachs, "Time synchronization in 5G wireless edge: Requirements and solu-

Novel Wi-Sync techniques as a promising way to solve both direct and indirect problems in the wireless TSN research trend.

- tions for critical-MTC," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 45–51, 2019. doi: 10.1109/MCOM.001.1900379.
- [4] I. Godor, et al. "A look inside 5G standards to support time synchronization for smart manufacturing," *IEEE Commun. Stand. Mag.*, vol. 4, no. 3, pp. 14–21, 2020. doi: 10.1109/MCOM-STD.001.2000010.
- [5] A. Mahmood, R. Exel, H. Trsek, and T. Sauter, "Clock synchronization over IEEE 802.11 - A survey of methodologies and protocols," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 907–922, 2017. doi: 10.1109/TII.2016.2629669.
- [6] Z. Wu, Y. Zhang, Y. Yang, C. Liang, and R. Liu, "Spoofing anti-spoofing technologies global navigation satellite system: A survey," *IEEE Access*, vol. 8, pp. 165,444–165,496, Sept. 2020. doi: 10.1109/ACCESS.2020.3022294.
- [7] G. Cena, I. C. Bertolotti, S. Scanzio, A. Valenzano, and C. Zunino, "Evaluation of EtherCAT distributed clock performance," *IEEE Trans. Ind. Informat.*, vol. 8, no. 1, pp. 20–29, 2012. doi: 10.1109/TII.2011.2172434.
- [8] A. Nasrallah et al., "Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G Ull research," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 88–145, 2019. doi: 10.1109/COMST.2018.2869350.
- [9] "Use cases IEC/IEEE 60802," IEEE, 2018. [Online]. Available: <https://grouper.ieee.org/groups/802/1/files/public/docs2018/60802-industrial-use-cases-0818-v11.pdf>
- [10] M. Luvisotto, Z. Pang, and D. Dzung, "High-performance wireless networks for industrial control applications: New targets and feasibility," *Proc. IEEE*, vol. 107, no. 6, pp. 1074–1093, 2019. doi: 10.1109/JPROC.2019.2898993.
- [11] "Integration of industrial ethernet networks with 5G networks," 5G-ACIA, Frankfurt, Germany, White Paper, 2019. [Online]. Available: https://5g-acia.org/wp-content/uploads/2021/04/5G-ACIA_Integration-of-Industrial-Ethernet-Networks-with-5G-Networks.pdf
- [12] Y. H. Wei, Q. Leng, S. Han, A. K. Mok, W. Zhang, and M. Tomizuka, "RT-WiFi: Real-time high-speed communication protocol for wireless cyber-physical control applications," in *Proc. - Real-Time Syst. Symp.*, 2013, pp. 140–149. doi: 10.1109/RTSS.2013.22.
- [13] Ó. Seijo, J. A. López-fernández, I. Val, and S. Member, "w-SHARP: Implementation of a high-performance wireless time-sensitive network for low latency and ultra-low cycle time industrial applications," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3651–3662, May 2021. doi: 10.1109/TII.2020.3007323.
- [14] Y. Wang and K. C. Ho, "TDOA source localization in the presence of synchronization clock bias and sensor position errors," *IEEE Trans. Signal Process.*, vol. 61, no. 18, pp. 4532–4544, 2013. doi: 10.1109/TSP.2013.2271750.
- [15] M. Ibrahim et al., "Verification: Accuracy evaluation of WiFi fine time measurements on an open platform," in *Proc. Annu. Int. Conf. Mob. Comput. Networking, MOBICOM*, 2018, pp. 417–427. doi: 10.1145/3241539.3241555.
- [16] D. M. Anand, D. Sharma, Y. Li-Baboud, and J. Moyné, "EDA performance and clock synchronization over a wireless network: Analysis, experimentation and application to semiconductor manufacturing," in *Proc. IEEE Int. Symp. Precis. Clock Synchronization Meas. Control Commun. ISPCS '09*, 2009, pp. 81–86. doi: 10.1109/ISPCS.2009.5340200.
- [17] T. Lennvall, J. Åkerberg, E. Hansen, and K. Yu, "A new wireless sensor network TDMA timing synchronization protocol," in *Proc. 2016 IEEE 14th Int. Conf. Ind. Informatics*, pp. 606–611. doi: 10.1109/INDIN.2016.7819233.
- [18] Ó. Seijo, I. Val, J. A. López-fernández, and M. Vélez, "IEEE 1588 clock synchronization performance over time-varying wireless channels," in *Proc. IEEE Int. Symp. Precis. Clock Synchronization Meas. Control Commun. ISPCS*, 2018, pp. 1–6. doi: 10.1109/ISPCS.2018.8543078.
- [19] R. Exel, "Clock synchronization in IEEE 802.11 wireless LANs using physical layer time-stamps," in *Proc. IEEE Int. Symp. Precis. Clock Synchronization Meas. Control Commun.*, 2012, pp.1–6.
- [20] A. Mahmood, R. Exel, and T. Sauter, "Delay and jitter characterization for software-based clock synchronization over WLAN using PTP," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1198–1206, 2014. doi: 10.1109/TII.2014.2304413.
- [21] F. M. Anwar and M. B. Srivastava, "Precision time protocol over LR-WPAN and 6LoWPAN," in *Proc. IEEE Int. Symp. Precis. Clock Synchronization Meas. Control Commun. ISPCS*, 2017, pp. 1–6. doi: 10.1109/ISPCS.2017.8056739.
- [22] C. M. De Dominicis, P. Pivato, P. Ferrari, D. Maccli, E. Sisinni, and A. Flammini, "Timestamping of IEEE 802.15.4a CSS signals for wireless ranging and time synchronization," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 8, pp. 2286–2296, 2013. doi: 10.1109/TIM.2013.2255988.
- [23] P. Ferrari et al., "Timestamping and ranging performance for IEEE 802.15.4 CSS systems," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 5, pp. 1244–1252, 2014. doi: 10.1109/TIM.2013.2286958.
- [24] Ó. Seijo, J. A. López-fernández, H.-P. Bernhard, and I. Val, "Enhanced timestamping Method for sub-nanosecond time synchronization in IEEE 802.11 over WLAN standard conditions," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5792–5805, Sept. 2020. doi: 10.1109/TII.2019.2959200.
- [25] *IEEE Standard for Information Technology-Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements—Part 11: WLAN MAC and PHY Specifications—Enhancements for Positioning*, IEEE Standard 802.11az, 2020.
- [26] C. Cruces, R. Torrego, A. Arriola, and I. Val, "Deterministic hybrid architecture with time sensitive network and wireless capabilities," in *Proc. IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA*, 2018, vol. 2018-Sept, pp. 1119–1122. doi: 10.1109/ETFA.2018.8502524.
- [27] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A survey on wireless security: Technical challenges, recent advances, and future trends," *Proc. IEEE*, vol. 104, no. 9, pp. 1727–1765, 2016. doi: 10.1109/JPROC.2016.2558521.
- [28] F. Pan, Z. Pang, M. Luvisotto, M. Xiao, and H. Wen, "Physical-layer security for industrial wireless control systems: Basics and future directions," *IEEE Ind. Electron. Mag.*, vol. 12, no. 4, pp. 18–27, 2018. doi: 10.1109/MIE.2018.2874385.



Data-Driven Edge Computing

A Fabric for Intelligent Building Energy Management Systems

ZHISHU SHEN, JIONG JIN,
TIEHUA ZHANG, ATSUSHI TAGAMI,
TERUO HIGASHINO, and
QING-LONG HAN

Digital Object Identifier 10.1109/MIE.2021.3120235
Date of current version: 10 November 2021

Building energy management systems (BEMSs) have been successfully adopted as key control units for modern structures to maintain energy efficiency and provide a comfortable thermal environment for occupants. Recent advances in information and communication technology toward “Industry 4.0” are enhancing the utility of BEMSs. However, challenges, such as how to process the exponentially growing amount of heterogeneous data generated in buildings, need to be addressed

to realize “Building 4.0,” which encompasses next-generation smart systems that provide user-centric services. In this article, we propose BEMS-Edge, a framework that integrates seamless, real-time information acquisition, transmission, interpretation, and action in intelligent BEMSs. The primary components, including the Internet of Things (IoT), cloud/edge computing, big data analytics, and artificial intelligence (AI), converge to create a data-driven edge computing fabric offering a range of benefits, such as real-time data analytics and cost savings.



The effectiveness of BEMS–Edge is verified by an established, real-world BEMS testbed.

The Role and Challenges of BEMSs

According to the International Energy Agency, buildings were responsible for 28% of global energy-related carbon dioxide emissions in 2018 [1]. Buildings' energy consumption will increase due to the necessity to construct more homes and offices to accommodate the rapid growth of the world population. BEMSs are needed in regions with harsh climates. Currently, buildings in such regions are generally equipped with fewer BEMSs, while it is forecast that these areas will experience the greatest population expansion. Hence, the expected growth rates for BEMSs are much higher than those of the global population. As a result, an efficient BEMS is in high demand to monitor and control a variety of building services, including heating, ventilation, and air conditioning (HVAC) and lighting, occupant activity detection, energy measurement, intrusion/fire alarms, and water supplies.

As an essential element of smart buildings, BEMSs are designed to reduce energy use while meeting requirements such as indoor comfort for occupants. Current BEMSs are computer-based, automated systems designed to support building management through services. Although BEMSs play an important role in managing those services, their operation and maintenance are costly and inefficient [2]. Moreover, they are mainly based on centralized and static control via central computers located at management stations. For example, various classical control techniques, such as on/off control, are used, in which the process is regulated based on a given upper/lower threshold [3]. Accordingly, tuning parameters is rather cumbersome for on/off control, especially for services, such as HVAC, that are designed to cope with time-varying environmental conditions. Due to a lack of detailed data for building states, current BEMSs fail to offer real-time data processing [4].

Challenges, such as how to process the exponentially growing amount of heterogeneous data generated in buildings, need to be addressed to realize “Building 4.0.”

Realizing Intelligent BEMSs Toward Building 4.0

Advanced information and communication technologies have been deployed in various industrial fields [5]. They can also be exploited toward Building 4.0. By improving hardware and operations, along with user interaction, Building 4.0 will create an innovative ecosystem and transform the entire industry. This innovation will ultimately enhance all aspects of essential building phases, including system operation and maintenance, as well as user experiences.

Efforts toward realizing Building 4.0 support the development of emerging technologies, including the IoT and cloud computing. To enhance data analytics for BEMSs, the literature discusses the advantage of introducing a cloud-based computing paradigm. This arrangement transfers collected sensor data to the cloud, where machine learning is harnessed to generalize overall system behavior [2], [6]. In terms of implementation, an IoT-enabled smart building system is developed, and its stability and robustness are confirmed [7]. A real-time digital model of an office building is created by analyzing building information modeling (BIM) and data collected from an IoT-enabled sensor network [8]. It is verified that the cloud-based computing paradigm can achieve energy savings up to 20% on HVAC installed in an experimental building [9]. Despite the fact that cloud computing offers exceptional big data processing capability, with increasing quantities of heterogeneous building data, it is challenging for the paradigm to achieve real-time results, especially considering issues such as communication overhead and network congestion [10].

Motivation

As a distributed computing paradigm that brings data computing, storage,

and network functions closer to end users, edge computing has been a feasible solution in various fields, including 5G networks and smart manufacturing [11]. Its development is motivated by the following two essential goals:

- a framework empowered by edge-based computing that is realized on top of existing BEMSs
- a data processing scheme that utilizes collaborative edges located near user sites for information sharing to reduce the amount of data transmitted to the cloud while improving data analytics.

The primary aim of this article is to explore new research opportunities in utilizing edge computing in BEMSs. BEMS–Edge is proposed to provide a data-driven edge computing fabric: in-network computing edges that provide an intermediate function, including the network proxy and data processing, between the cloud and massive sensors. The main contributions of this article include the following:

- A framework that performs comprehensive data processing, from information acquisition to decision making, is proposed. Edge intelligence that utilizes AI is introduced to enhance the data-driven analytics.
- Hybrid edge–cloud analytics paradigms for BEMS–Edge are studied. These paradigms assign distinct roles to the edges and cloud to achieve data-driven processing that meets BEMS service requirements.
- Verification based on a real-world BEMS testbed demonstrates that BEMS–Edge can achieve satisfactory data analytics in real time.

Overview of BEMS–Edge

As a fundamental framework for realizing Building 4.0, BEMS–Edge provides

An efficient BEMS is in high demand to monitor and control a variety of building services, including heating, ventilation, and air conditioning and lighting.

an edge computing fabric for intelligent BEMSs to perform comprehensive information processing, which is realized on top of a combination of software and data communication/processing hardware.

Vision

Figure 1 illustrates the vision for BEMS-Edge. The system improves traditional BEMSs to achieve seamless, real-time information acquisition, transmission, interpretation, and action. The basic data flow is as follows:

- 1) Sensors in every room collect data reflecting building conditions, such as temperatures and luminous intensities.
- 2) The edge aggregates sensor data and forwards processed information to the cloud for further analytics and storage. It acts as a gateway that supports communication among

sensing devices through various protocols and that can also be used for data filtering and analytics. By harnessing edge intelligence [11], building tasks can be partially or entirely processed in the form of in-network computing via collaborative edges that manage local, real-time data. Meanwhile, employing other available edges and communication gateways can resolve issues when designated edges are temporarily unavailable due to unexpected incidents.

- 3) The cloud receives information from the edges and analyzes the collected data, if necessary. Based on the results, it relays actionable information to BEMS administrators, which consist of integrated machines, programs, and human operators. Accurate data analytics can support BEMSs to achieve

automatic control over building equipment, while human operators carry out regular reviews of system settings to gradually reduce room set points, operating times, and energy consumption [12].

Key Components

Ubiquitous Energy-Efficient Sensing

Ubiquitous sensors that support wireless communication have attracted increasing attention because of their ability to solve issues related to traditional hard-wired meters, which fail to provide sufficient building data to BEMSs. However, simultaneously operating a large number of sensors placed throughout buildings increases BEMS energy consumption, especially when transmitting data [13]. To resolve this, BEMS-Edge conducts sensing in an energy-efficient manner by applying lightweight messaging protocols, such as constrained application protocol or message queuing telemetry transport.

Meanwhile, using sleep mode for certain sensors is a direct approach to reduce the energy consumed during

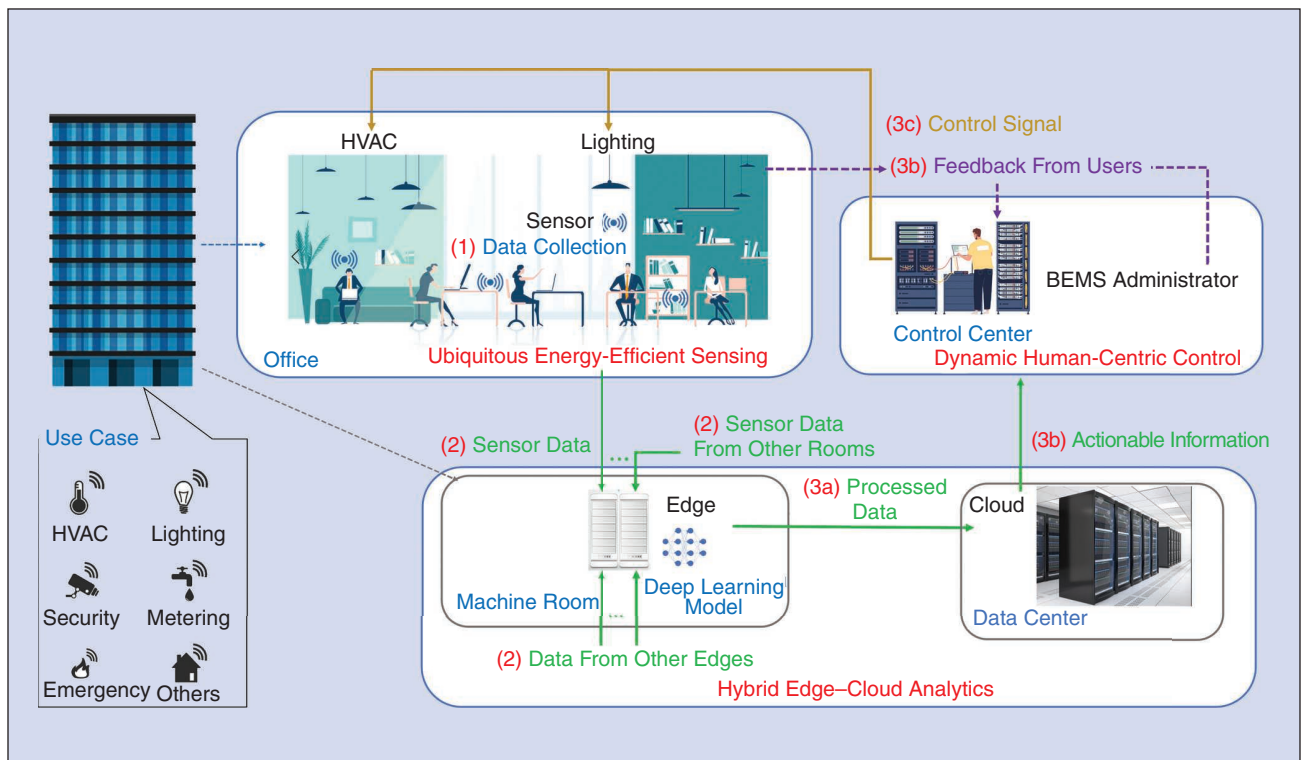


FIGURE 1 – The BEMS-Edge architecture to enable real-time IoT data analytics.

long-term monitoring. However, with this approach, data analytics might degrade due to a lack of valid sensor information. Our solution is to utilize an edge fabric to learn time series correlations within available data and extrapolate missing information from the sensors in sleep mode [14]. In this scheme, a deep autoencoder built on a long short-term memory (LSTM)-based, sequence-to-sequence structure is applied as the core of the data analytics at the edge fabric. LSTM is empowered by different cell states and gates capable of capturing underlying correlations among time series sensor data. The obtained complementary information can be processed at the same edge for a given BEMS application, e.g., hotspot detection. The output of each edge is aggregated to the cloud to infer an overall building status.

Hybrid Edge-Cloud Analytics

The proliferation of the IoT has created concerns about how to process big data, especially for BEMSs that generate massive amounts of redundant sensor information. As indicated previously, delays caused by transferring these data to the cloud and waiting for control feedback are a critical concern for establishing intelligent BEMSs. It is important to integrate the advantages of cloud and edge computing to provide satisfactory data analytics with minimum response times. Emerging edge intelligence leveraging AI can enhance analytics by executing deep learning training and inference phases at edges in the vicinity of end users. As an analytical tool, deep learning involves the generation of a model by learning correlations and dependences among data. The model is used to forecast intended knowledge by analyzing sensor data. To cope with changes in sensing environments, it is necessary to update the model by analyzing the most recent information stored at an edge through a given period.

Dynamic User-Centric Control

Meeting requirements for user-centric control is crucial for BEMSs. In this article, the term *control* refers to

By improving hardware and operations, along with user interaction, Building 4.0 will create an innovative ecosystem and transform the entire industry.

decision-based actions. For instance, based on the results of real-time data analytics, electric loads can be dynamically controlled according to system usage patterns and building environment statuses. In addition to models based on BIM and information from sensors, BEMS control can be supplemented with a human-in-the-loop (HITL) approach, which integrates human factors into the data interpretation and action process [15]. The HITL method captures the knowledge of the users themselves to support an intelligent dynamic control scheme. Users can be BEMS experts, local administrators, and even occupants, who are vital participants in BEMSs. A typical HITL application is user trajectory prediction. This can control building equipment based on an area's status. In cooperation with hotspot detection, BEMSs adjust the

air conditioning in crowded areas where low-thermal-comfort statuses are registered.

Capabilities of BEMS-Edge

BEMS-Edge is expected to provide various enhanced capabilities to tackle potential challenges of implementing Building 4.0. Figure 2 summarizes the key benefits of BEMS-Edge compared with the capabilities of two similar BEMS schemes.

- *Current BEMSs*: These are connected to a central computer terminal, where the status of building systems is statically controlled based on rules defined in advance.
- *BEMS-Cloud*: Sensors are deployed for building status monitoring, while collected sensor data are forwarded to the cloud for analytics. Based on the results, administrators control services [2], [6].

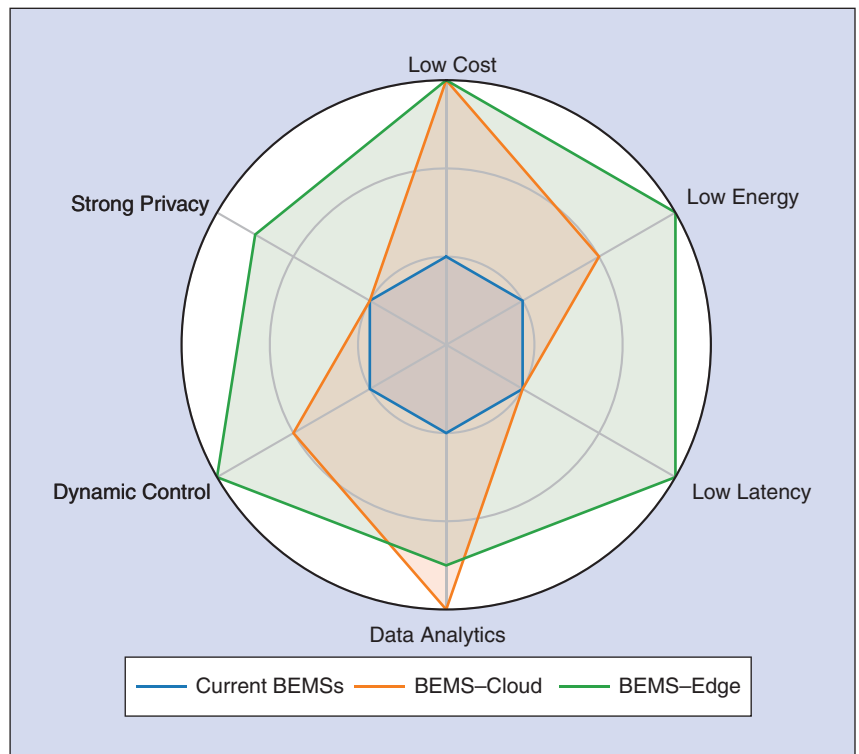


FIGURE 2 – A comparison of different BEMS capabilities.

By harnessing edge intelligence, building tasks can be partially or entirely processed in the form of in-network computing via collaborative edges that manage local, real-time data.

Cost

For BEMS-Cloud and BEMS-Edge, integrating IoT technologies reduces system installation and operation costs. Moreover, from the perspective of system diagnostics, by analyzing regularly collected sensor data, maintenance costs are lower than those of current BEMSs, which require periodic inspections due to sensing and control limitations [6].

Latency

Current BEMSs and BEMS-Cloud need to send collected sensor data to a central computer terminal or the cloud and wait for control information from administrators. On the other hand, for BEMS-Edge, since data analytics can be

performed at edges near buildings, the latency can be significantly reduced.

Data Analytics

Although BIM models are constantly being developed and upgraded, current BEMSs fail to provide satisfactory data analytics due to a lack of real-time building information [4]. BEMS-Cloud guarantees satisfactory analytics with the help of the cloud. The aggregation of data from different edge participants in the cloud enables the adoption of more complex deep learning models, leading to superior analytics. In contrast, BEMS-Edge encounters so-called data isolation if analytics are independently performed at each edge participant. Nevertheless, it has been

shown that a collaborative edge can produce analytics comparable to those achieved by centralized cloud servers, while local data privacy is preserved to some extent [16].

Dynamic Control

Current BEMSs are not designed specifically for dynamic and diverse building management. BEMS-Cloud changes this by utilizing real-time sensor data; however, latency is a critical issue for services requiring prompt control. By conducting real-time analytics at edges, BEMS-Edge offers user-centric control with adaptability to occupant behavior and changes to building layouts.

Energy Consumption

BEMS-Edge, powered by AI, can achieve accurate and dynamic control over building equipment to reduce energy use. It can further lower networking equipment energy consumption by transmitting aggregated data to the cloud.

Privacy

Although BEMS-Cloud uses the cloud to analyze sensor data to achieve acceptable analytics, it also increases risks associated with data leaks. BEMS-Edge eases this concern by processing private information via edges located in or near buildings as much as possible.

Paradigms of Hybrid Edge-Cloud Analytics for BEMS-Edge

Figure 3 illustrates four hybrid edge-cloud analytics paradigms expected to be used in Building 4.0. Here, event-driven BEMSs are designed to detect incidents from collected sensor data and react to them based on user requirements.

Paradigm 1: Analytics in the Cloud

For BEMS services designed for security and safety, providing accurate information about emergencies is the priority. Taking fires and evacuations as an example, BEMSs are required to produce valid information about where the flames are, the status of egress routes, additional hazards, and so on [17]. Analytics in the cloud are

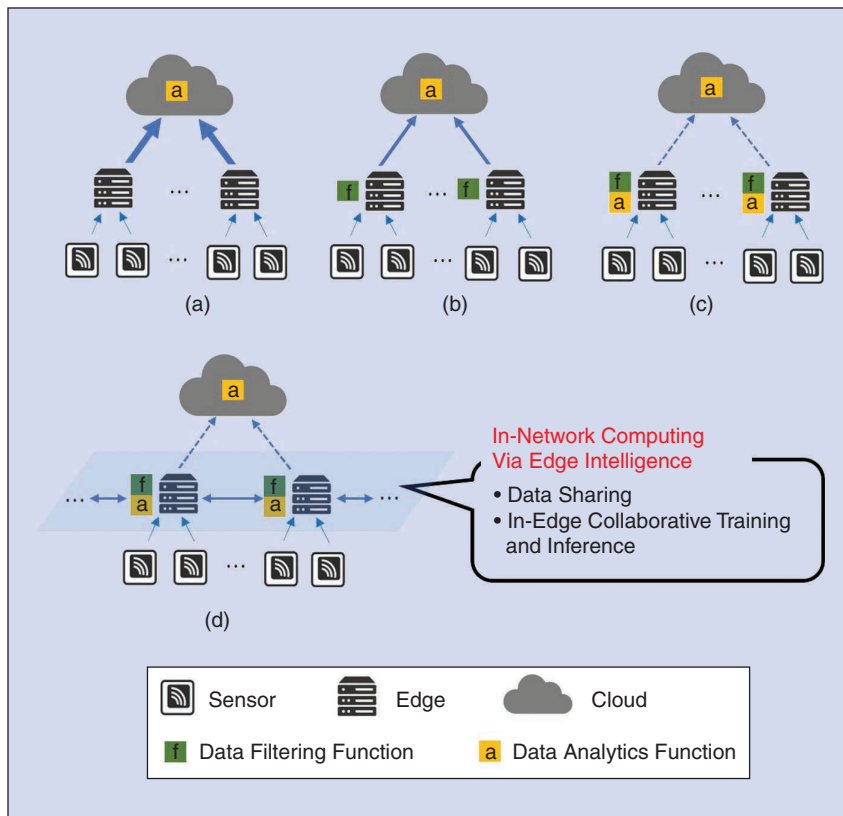


FIGURE 3 – Hybrid edge-cloud analytics paradigms for BEMS-Edge. (a) Analytics in the cloud. (b) Filtering at the edge. (c) Analytics at the edge. (d) Analytics at collaborative edges.

solution for obtaining reliable information by conducting aggregated analytics. To support this, sensor data are forwarded to the cloud via edges for communication among devices with different protocols.

Paradigm 2: Filtering at the Edges, Followed by Analytics in the Cloud

If all BEMS services are implemented through analytics in the cloud, IoT network energy costs and delays will be critical due to the data volume that must be transferred. Filtering at the edges moderates this issue because it does not directly upload all the information to the cloud. The filtering can be a simple removal of erroneous sensor data or an advanced process that creates clusters based on collected information that are sent to the cloud for outlier detection [18]. Since the edges fuse sensor data to some extent, the data volume transmitted to the cloud can be reduced. This paradigm is designed for BEMS services that require cloud services, such as smart metering, in which the cloud is used to coordinate distribution grid operations [19].

Paradigm 3: Analytics at the Edges

To reduce the information volume sent to the cloud, in this paradigm, all relevant sensor data are analyzed at the edges, with a data analytics function. Contrary to filtering at the edge, in which preprocessed results must be sent to the cloud for further analysis, analytics at the edge send only actionable information, such as the ID of a sensor that detects an anomaly, to the cloud as a notification to the administrators. Thus, latency can be significantly reduced. Analytics at the edge can be applied for services that require prompt responses. Taking HVAC control as an example, an air conditioner may take 10 min to cool a room that is too warm. Hence, it is vital to rapidly and accurately identify hotspots to maintain a comfortable environment [20]. Compared with emergency BEMS cases, raw building data do not have to be sent to the administrators via the cloud in real time. Instead, it is reasonable to execute

The proliferation of the IoT has created concerns about how to process big data, especially for BEMs that generate massive amounts of redundant sensor information.

data processing at the edges to reduce transmission congestion while achieving acceptable analytics.

Paradigm 4: Analytics at Collaborative Edges

As a complement to cloud computing, analytics at collaborative edges obtain data from neighboring edges to mitigate data isolation. Edge collaboration can incorporate in-network computing to achieve superb analytics while reducing the amount of private information that is exchanged. In summary, in-network computing breaks down into the following key elements:

- *Data sharing:* Collaboration enables authorized data sharing from edges located in different building areas and even from management systems that control various equipment. Shared data can be used for processing by deep learning models installed at the edges.

- *In-edge collaborative training and inference:* To enhance data analytics, collaboration calls for edge participants to optimize the training parameters of a machine/deep learning model. Additionally, this enables dynamic selection of the parameter/coordinator server to facilitate the training process, in which cloud involvement is no longer mandatory.

Real-World Testbed for Evaluation

Setup

A testing environment is established inside an office building in Osaka, Japan. Sensor placement information is provided in Figure 4, where 34 instruments (T&D RTR-500 series) are deployed to collect temperature data within an area of 1,800 m². Data for two months are collected at an interval

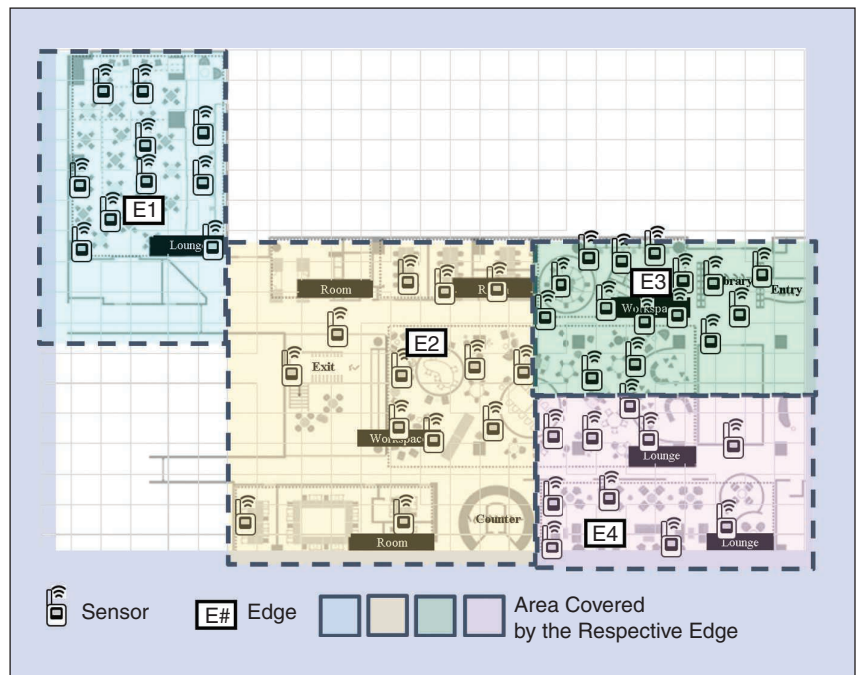


FIGURE 4 – The sensor deployment environment.

A collaborative edge can produce analytics comparable to those achieved by centralized cloud servers, while local data privacy is preserved to some extent.

of 30 min. Note that selection of the interval depends on how a building is used. Four edges are employed to process information obtained from an equal number of areas. In particular, the testbed focuses on 1) hotspot detection and 2) user trajectory prediction, as illustrated in Figure 5. Hotspot detection, which is a key component for HVAC control, determines whether the sensors detect the hotspot areas by analyzing data from nearby devices. For each time stamp, the status of all the sensors is assessed to score the hotspot detection accuracy. Based on the result, a BEMS administrator can specify an abnormal area and control the air conditioners there.

For user trajectory prediction, Bluetooth Low Energy (BLE) beacon sensors (Sanwa MM-BLEBC1) act as reference devices that send wireless signals to points (see Figure 5) users might walk through every 0.5 s. Due to the inapplicability of GPS in indoor environments,

we use BLE as the source of the wireless signals to analyze the received signal strength indicator (RSSI) for trajectory prediction: when a user walks between two points, his or her mobile phone receives RSSI signals from the beacon sensors, and it sends collected RSSI data to surrounding edge nodes to predict his or her trajectory.

Hotspot Detection

Figure 6 graphs the performance of different hybrid edge–cloud analytics paradigms. Regarding the transmitted data volume, analytics in the cloud need to send all information to the cloud for aggregated processing, while filtering at the edge reduces the amount by sorting information in advance. The accuracy of detecting anomalies in sensor data depends on the use of the edge and cloud for processing, and hence, analytics in the cloud achieve the best performance because they are centralized. Two

hybrid edge–cloud paradigms, deep reinforcement learning (Edge-DRL) [10] and plug-and-play learning (PPL) [21], are introduced for comparison. For Edge-DRL, model training and data analytics decision making are executed at the edges. Due to the limited local data pool at each edge for model training, the prediction accuracy obtained by Edge-DRL is limited. To resolve this, PPL and analytics at the edge offload model training to the cloud. Analytics at the edge excel at prediction accuracy due to the adoption of deep learning techniques, from which correlations among the massive amounts of data collected from multiple sensors in different areas are learned, while PPL uses other traditional algorithms suitable for training small data sets.

However, analytics at the edge fail to detect some hotspots due to the limited information resources available at an independent edge, at which data are collected from sensors in specific areas. Analytics at collaborative edges obtain data from neighboring edges to resolve this limitation. It is worth noting that the analytics time for these two paradigms, which are tested on a Raspberry Pi 2 Model B, is shorter than 1×10^{-3} s, which is acceptable for real-time BEMS hotspot

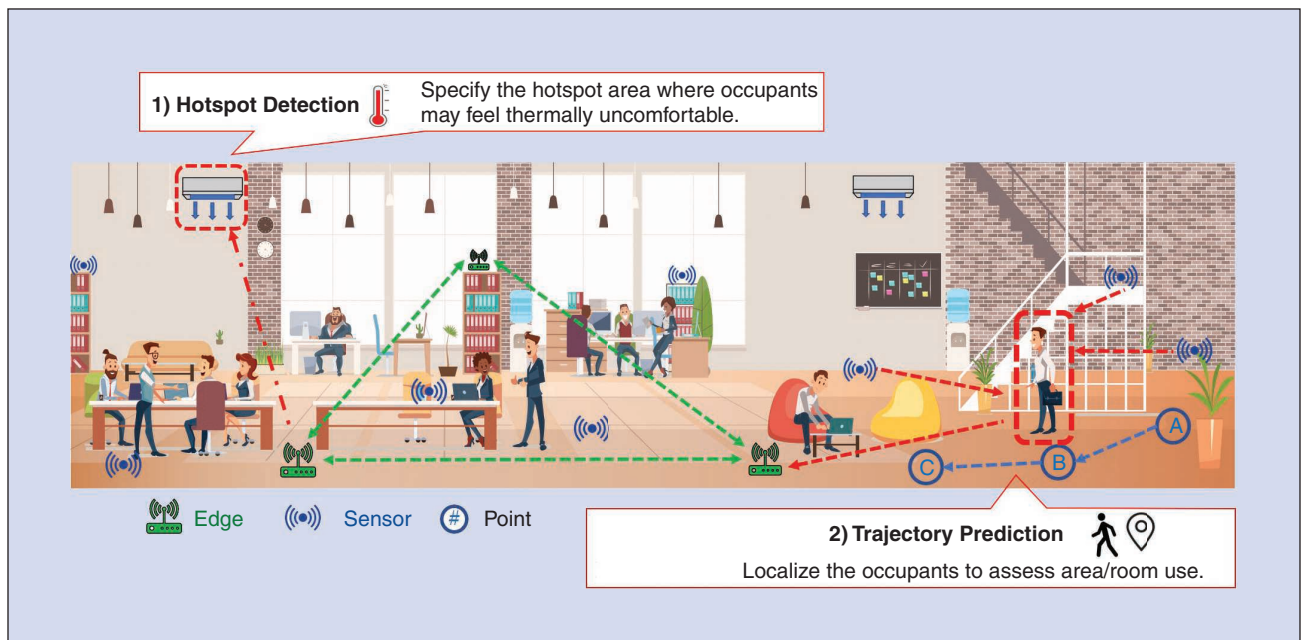


FIGURE 5 – The example BEMS use cases.

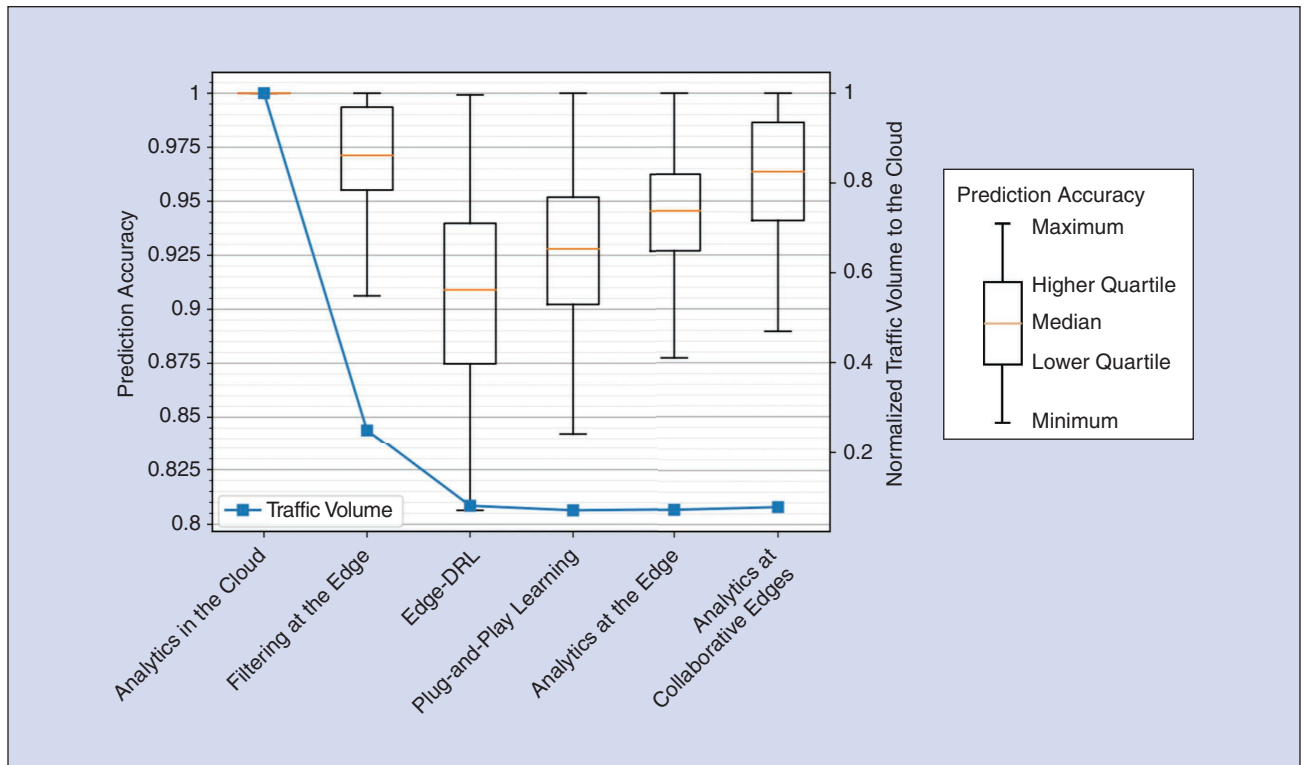


FIGURE 6 – The performance of hybrid edge–cloud analytics paradigms for BEMS–Edge.

detection. The communication expense is also simulated for each paradigm in terms of the Amazon Web Services IoT Core data publishing cost, based on the May 2021 rate in the Tokyo region. For our established testbed, the monthly cost of analytics at collaborative edges is US\$1.40, compared to US\$18.40 for analytics in the cloud. For the Tokyo Station locale (with a total floor area of 700 ha), a large business district with multiple commercial buildings with BEMSs, the estimated monthly cost of analytics at collaborative edges is US\$420, while analytics in the cloud cost US\$5,460.

Trajectory Prediction

This evaluation focuses on the use of analytics in the cloud and analytics at collaborative edges since these paradigms are expected to achieve satisfactory data analysis for life-saving services, such as emergency evacuations. To explore the underlying correlation among RSSI signals, analytics in the cloud data processing is based on a cloud-based computation, while analytics at collaborative edges

utilize in-network computing edges to train a deep learning model with sensor data. Our results demonstrate that analytics at collaborative edges attain competitive performance against analytics in the cloud in terms of user trajectory accuracy; i.e., in general, they achieve real-time (the calculation time is shorter than 1×10^{-3} s), four-point trajectory prediction with an accuracy of more than 80%. It is worth noting that they provide data analytics in the proximity of BEMS users without sharing private information to the cloud. Furthermore, by introducing data obfuscation based on ϵ -differential privacy [22], users' locations can remain private.

Conclusion

In this article, BEMS–Edge, with a data-driven edge fabric facilitating intelligent BEMSs, was proposed for Building 4.0. By incorporating the IoT, cloud/edge computing, and AI, the system enhances BEMSs in terms of costs, latency, and data analytics. Four hybrid edge–cloud analytics paradigms were studied to meet the requirements of event-driven BEMS services. A BEMS

testbed was established in a commercial building to evaluate the effectiveness of BEMS–Edge in two use cases. The results demonstrate that BEMS–Edge can achieve satisfactory data analytics for hotspot detection and trajectory prediction. A future research direction is to upgrade BEMS–Edge to support dynamic data processing, e.g., analyzing information provided by occupants' mobile devices.

Acknowledgments

This work was supported, in part, by the Research and Development of Innovative Network Technologies to Create the Future project of the National Institute of Information and Communications Technology, Japan, and the Australian Research Council Discovery Project, under grant DP190102828.

Biographies

Zhishu Shen (z_shen@ieee.org) earned his B.E. degree from the School of Information Engineering, Wuhan University of Technology, Wuhan, China, in 2009, and M.E. and Ph.D. degrees in electrical and electronic engineering and computer science from Nagoya

University, Japan, in 2012 and 2015, respectively. He is a research engineer in the architecture laboratory at KDDI Research, Inc., Fujimino-shi, Saitama, 356-8502, Japan. His research interests include network design and optimization, data learning, and the Internet of Things. He is a Member of IEEE.

Jiong Jin (jiongjin@swin.edu.au) earned his B.E. degree, with first-class honors, in computer engineering from Nanyang Technological University, Singapore, in 2006, and Ph.D. degree in electrical and electronic engineering from the University of Melbourne, Australia, in 2011. He is an associate professor in the School of Science, Computing, and Engineering Technologies, Swinburne University of Technology, Hawthorn, Victoria, 3122, Australia. His research interests include network design and optimization, edge computing and networking, robotics and automation, and cyber-physical systems and the Internet of Things as well as their applications in smart manufacturing, smart transportation, and smart cities. He is a Member of IEEE.

Tiehua Zhang (tiehuaz@ieee.org) earned his M.E. degree from the School of Computing and Information Systems, University of Melbourne, Australia, in 2015, and Ph.D. degree from the School of Software and Electrical Engineering, Swinburne University of Technology, Australia, in 2020. From 2015 to 2017, he was a software engineer in Australia, focusing on industrial projects and solutions. He is currently an artificial intelligence specialist at Ant Group, Shanghai, 200000, China. His research interests include collaborative learning/optimization, the Internet of Things, and edge intelligence. He is a Student Member of IEEE.

Atsushi Tagami (tagami@kddi-research.jp) earned his M.E. and Ph.D. degrees in computer science from Kyushu University, Japan, in 1997 and 2010, respectively. He joined KDDI Research in 1997, where he has been engaged in the research and development of performance measurements for communication networks and overlay networking. He is currently a senior manager in the architecture laboratory at KDDI Research, Inc., Fu-

jimino-shi, Saitama, 356-8502, Japan. He is a Member of IEEE.

Teruo Higashino (higashino@ist.osaka-u.ac.jp) earned his M.S. and Ph.D. degrees in 1981 and 1984, respectively, in information and computer sciences from Osaka University, Suita, Osaka, 565-0871, Japan, where he is a specially appointed professor. Since 2021, he has also been a professor at and dean of Kyoto Tachibana University, Kyoto, 607-8175, Japan. His research interests include the design and analysis of distributed systems, communication protocols, and mobile computing. He is a fellow of the Information Processing Society of Japan and the principal investigator of the Japanese government's Society 5.0 Project (2018–2023). He is a Senior Member of IEEE.

Qing-Long Han (qhan@swin.edu.au) earned his M.Sc. and Ph.D. degrees in control engineering from East China University of Science and Technology, Shanghai, in 1992 and 1997, respectively. He is a member of Academia Europaea as well as the recipient of the 2021 M.A. Sargent Medal (the highest award given by the Engineers Australia Electrical College), 2019 and 2020 IEEE Systems, Man, and Cybernetics Society Andrew P. Sage Best Transactions Paper Award, and 2020 IEEE Transactions on Industrial Informatics Outstanding Paper Award. He is a Fellow of IEEE.

References

- [1] "Tracking buildings," International Energy Agency, Paris, France, 2019. [Online]. Available: <https://www.iea.org/reports/tracking-buildings>
- [2] M. Manic, D. Wijayasekara, K. Amarasinghe, and J. J. Rodriguez-Andina, "Building energy management systems: The age of intelligent and adaptive buildings," *IEEE Ind. Electron. Mag.*, vol. 10, no. 1, pp. 25–39, 2016, doi: 10.1109/MIE.2015.2513749.
- [3] A. Afram and F. Janabi-Sharifi, "Theory and applications of HVAC control systems – A review of model predictive control (MPC)," *Building Environ.*, vol. 72, pp. 343–355, Feb. 2014, doi: 10.1016/j.buildenv.2013.11.016.
- [4] M. Dastbaz, C. Gorse, and A. Moncaster, *Building Information Modelling, Building Performance, Design and Smart Construction*. Cham: Springer Nature Switzerland AG, 2017.
- [5] G. Aceto, V. Persico, and A. Pescapé, "A survey on information and communication technologies for industry 4.0: State-of-the-art, taxonomies, perspectives, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3467–3501, 2019, doi: 10.1109/COMST.2019.2938259.
- [6] W. Tushar *et al.*, "Internet of things for green building management: Disruptive innovations through low-cost sensor technology and artificial intelligence," *IEEE Signal Process. Mag.*,

- vol. 35, no. 5, pp. 100–110, 2018, doi: 10.1109/MSP.2018.2842096.
- [7] W. Xu *et al.*, "The design, implementation, and deployment of a smart lighting system for smart buildings," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7266–7281, 2019, doi: 10.1109/JIOT.2019.2915952.
- [8] S. H. Khajavi, N. H. Motlagh, A. Jaribion, L. C. Werner, and J. Holmström, "Digital twin: Vision, benefits, boundaries, and creation for buildings," *IEEE Access*, vol. 7, pp. 1,47,406–1,47,419, Oct. 2019, doi: 10.1109/ACCESS.2019.2946515.
- [9] E. Png, S. Srinivasan, K. Bekiroglu, J. Chaoyang, R. Su, and K. Poolla, "An internet of things upgrade for smart and scalable heating, ventilation and air-conditioning control in commercial buildings," *Appl. Energy*, vol. 239, pp. 408–424, Apr. 2019, doi: 10.1016/j.apenergy.2019.01.229.
- [10] Y. Liu, C. Yang, L. Jiang, S. Xie, and Y. Zhang, "Intelligent edge computing for IoT-based energy management in smart cities," *IEEE Netw.*, vol. 33, no. 2, pp. 111–117, 2019, doi: 10.1109/MNET.2019.1800254.
- [11] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869–904, 2020, doi: 10.1109/COMST.2020.2970550.
- [12] K. Sayed and H. A. Gabbar, *Building Energy Management Systems (BEMS)*. Hoboken, NJ, USA: Wiley, 2017, ch. 2, pp. 15–81.
- [13] V. Raghunathan, C. Schurgers, S. Park, and M. B. Srivastava, "Energy-aware wireless micro-sensor networks," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 40–50, 2002, doi: 10.1109/79.985679.
- [14] T. Zhang, Z. Shen, J. Jin, A. Tagami, X. Zheng, and Y. Yang, "ESDA: An energy-saving data analytics fog service platform," in *Proc. Int. Conf. Service Oriented Comput.*, 2019, pp. 171–185, doi: 10.1007/978-3-030-33702-5_13.
- [15] M. V. Bavaresco, S. D'Oca, E. Ghisi, and R. Lamberts, "Technological innovations to assess and include the human dimension in the building-performance loop: A review," *Energy Buildings*, vol. 202, p. 109,365, Nov. 2019, doi: 10.1016/j.enbuild.2019.109365.
- [16] T. Zhang, Z. Shen, J. Jin, X. Zheng, A. Tagami, and X. Cao, "Achieving democracy in edge intelligence: A fog-based collaborative learning scheme," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2751–2761, 2021, doi: 10.1109/JIOT.2020.3020911.
- [17] H. M. Poy and B. Duffy, "A cloud-enabled building and fire emergency evacuation application," *IEEE Cloud Comput.*, vol. 1, no. 4, pp. 40–49, 2014, doi: 10.1109/MCC.2014.67.
- [18] L. Lyu, J. Jin, S. Rajasegarar, X. He, and M. Palaniswami, "Fog-empowered anomaly detection in IoT using hyperellipsoidal clustering," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1174–1184, 2017, doi: 10.1109/JIOT.2017.2709942.
- [19] M. Pau *et al.*, "A cloud-based smart metering infrastructure for distribution grid services and automation," *Sustain. Energy, Grids Netw.*, vol. 15, pp. 14–25, Sep. 2018, doi: 10.1016/j.segan.2017.08.001.
- [20] Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng, "Duty-cycling buildings aggressively: The next frontier in HVAC control," in *Proc. ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, 2011, pp. 246–257.
- [21] X. Zhang, M. Pipattanasomporn, T. Chen, and S. Rahman, "An IoT-based thermal model learning framework for smart buildings," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 518–527, 2020, doi: 10.1109/JIOT.2019.2951106.
- [22] P. Zhao *et al.*, "P3-LOC: A privacy-preserving paradigm-driven framework for indoor localization," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2856–2869, 2018, doi: 10.1109/TNET.2018.2879967.





©SHUTTERSTOCK.COM/PROFIT_IMAGE

System-on-Chip FPGA Devices for Complex Electrical Energy Systems Control

ERIC MONMASSON,
MICKAËL HILAIRET,
GIOVANNI SPAGNUOLO, and
MARCIAN N. CIRSTEĂ

Digital Object Identifier 10.1109/MIE.2021.3052179
Date of current version: 19 February 2021

Digital electronics has become a standard for controlling electrical systems. This is due to the constant improvement of the digital devices, whether in terms of density, performance, flexibility of use, or cost reduction [1]. This article looks into system-on-chip

(SoC) field-programmable gate array (FPGA) for controlling complex electrical energy systems. These devices encompass multicore floating-point microprocessors embedded with standard peripherals and an FPGA fabric that allows the design of custom peripherals and specific hardware (HW) accelerators. Thus, SoC FPGA devices

One of them is the opportunity to enlarge significantly the size of the electrical energy systems to manage.

can be regarded as a good compromise between “super” microcontrollers (very fast in terms of computation but with a fixed microarchitecture) and pure FPGAs (ideal for specific concurrent microarchitectures but limited in terms of density).

SoC FPGA architectures are discussed and compared with state-of-the-art digital signal processor (DSP) controllers, since they can also be qualified as SoC devices as they are integrating floating-point microprocessor cores and substantial peripherals. The main differences between these two groups of devices lies in the opportunity offered to the designer by the SoC FPGAs to customize the SoC device via its internal FPGA fabric. Two case studies demonstrate that with SoC FPGAs one can go beyond standard control by introducing new auxiliary functions that enhance market competitiveness. The first application concerns a fuel cell (FC) hybrid electric system controlled by passivity-based power management associated with an aging prognosis algorithm. For this application, it is shown that the time and cost constraints justify the use of a soft processor core to implement the controller.

The second application concerns the maximization of the electrical power production of a photovoltaic (PV) field operating in mismatched conditions through the dynamic reconfiguration of the PV modules. This application allows us to illustrate the ability of SoC FPGA to solve a complex optimization problem in a time that is so short that the PV field operating conditions can be considered as constant. Second, it shows the benefits of implementing C/C++ high-level synthesis (HLS)-based HW accelerators by significantly simplifying the design space exploration phase.

Finally, to generalize the lessons learned from these case studies, we

present an analysis of recent and inspiring controllers for complex electrical energy systems from which key principles for designing the next generation of SoC FPGA-based smart controllers are derived.

Embedded Digital Controllers and SoC: Evolution and Trends

Due to their ability to execute control algorithms of ever increasing complexity in a very short time, using cheap components, digital controllers took preference over the analog ones. Microcontrollers and DSPs are used [2], however, FPGA-based controllers also have some advantages [3]. DSPs and microcontrollers are flexible (C-based programming), low cost, and with a highly performing floating-point arithmetic logic unit. DSP controllers integrate a high number of peripherals, all well-fitting with the control of power electronics and drives. The main disadvantage of such devices is that they are based on a fixed microarchitecture that prevents the concurrent execution of tasks that could be executed in parallel. This significantly limits their timing performance, leading to the introduction in the controller of one sampling period delay that reduces the control system's bandwidth and introduces more chattering into direct control of power converters.

Initially designed as a simple fabric of lookup tables and flip-flops, FPGAs then integrated DSP units and memory banks and lately the end user has been able to easily synthesize 32-b reduced-instruction-set computer (RISC) processors within the FPGA fabric [4]. FPGAs are attractive for controlling industrial systems mainly because they allow the design of dedicated controllers that are the “hardware copies” of the source control algorithms, thus including the entire potential parallelism of these

algorithms and, as a consequence, accelerating significantly their real-time executions. FPGAs can also handle the control of systems with a high number of inputs–outputs (I/Os), such as multilevel converters. Indeed, the parallelism can be inside the control algorithm and it can be intrinsic to the system to be controlled, like for multiphase motors. As no additional delay is introduced, the FPGA-based controller increases the bandwidth of the designed control loops, thus they are ideal for the direct control of power converters [3], including power electronics using the recently introduced wideband-gap power switches that are commonly driven with a switching frequency above 100 kHz [5]. Computational demanding algorithms like model predictive control are also good candidates for FPGA-based implementations because of their parallelized and highly pipelined architecture [6]. The main drawbacks of FPGAs are the lack of performing internal analog-to-digital converters (ADCs) and limited size, which make floating-point arithmetic architecture design problematic. However, Intel has introduced an FPGA with 32-b floating-point DSP units [7].

SoC devices were introduced around a decade ago, mainly due to the benefits brought to mobile phones and, more recently, to the Internet of Things (IoT) [8]. They also impacted control applications because of their impressive computational power; the parallelism of the computing tasks can also be obtained by running several tasks simultaneously on different processor cores, with the possibility to also embed a real-time Linux operating system (OS). Thus, SoC can help expand the domain of traditional control algorithms (Figure 1) and brings convergence between the worlds of DSP controllers and FPGAs.

The Texas instrument dual Delfino device [9] [see Figure 2(a)] represents a natural SoC evolution of traditional DSP controllers. It is based on dual 32-b floating-point DSP cores, with always more peripherals and dedicated arithmetic units like a Viterbi complex math unit and a trigonometric math

unit (TMU), which can be regarded as specific HW accelerators [Figure 2(a)]. With the TMU, a Park's transformation can be executed in about 100 ns, comparable to what can be achieved with an FPGA. Also, parallel computing is now possible since four tasks can be executed simultaneously, one on each DSP core and one in each of the two control law accelerators (CLA) cores. So, its clock frequency is 200 MHz but, because it is a multicore architecture, it can reach up to 800 million instructions per second. CLAs alleviate the DSP cores of low level but very time-constrained tasks, like an FPGA current/voltage controller would do. Most insulated-gate bipolar transistor-based inverter switching control functions in the 10-kHz frequency range can, therefore, be achieved.

SoC FPGAs (Xilinx Zynq, Intel FPGA Arria 10, or Intel FPGA Cyclone V devices [7]) include a dual-core ARM A9, along with powerful coprocessors like the single instruction, multiple data NEON, a set of peripherals to communicate with other boards, and highly performing FPGA fabric [Figure 2(b)]. The latter offers the designer the possibility to add custom peripheral and/or specific HW accelerators adapted

SoC FPGA can make significant contributions to the key developments in complex electrical energy systems, especially those including renewable generators and those employing hydrogen technology.

to a given application. The 32-b ARM A9 microprocessors are intended to run a powerful OS-like embedded Linux. However, these processors can also be used for bare metal applications that are more adapted to standard control solutions for electrical systems. Running at 667 MHz, they feature high computing power and high-quality internal buses used for controlling either a simple peripheral via its internal registers or for exchanging a stream of data at high rate with an FPGA-based HW accelerator [10]. SoC FPGA components can easily implement 32-b RISC processor cores within the FPGA fabric (MicroBlaze for Xilinx, NIOS II for Intel/Altera, and ARM Cortex-M1 or -M3 [11]). These features offer huge flexibility to the designer who can, thanks to the FPGA

fabric, integrate specific peripherals and/or HW accelerators plus additional 32-b RISC processor cores into the SoC architecture. Table 1 summarizes

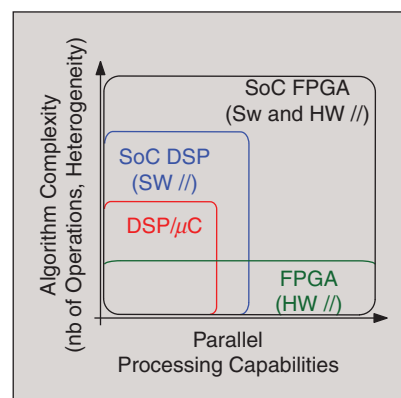


FIGURE 1 – A schematic showing the ability of each device technology to handle an algorithm's complexity and concurrency. nb: number.

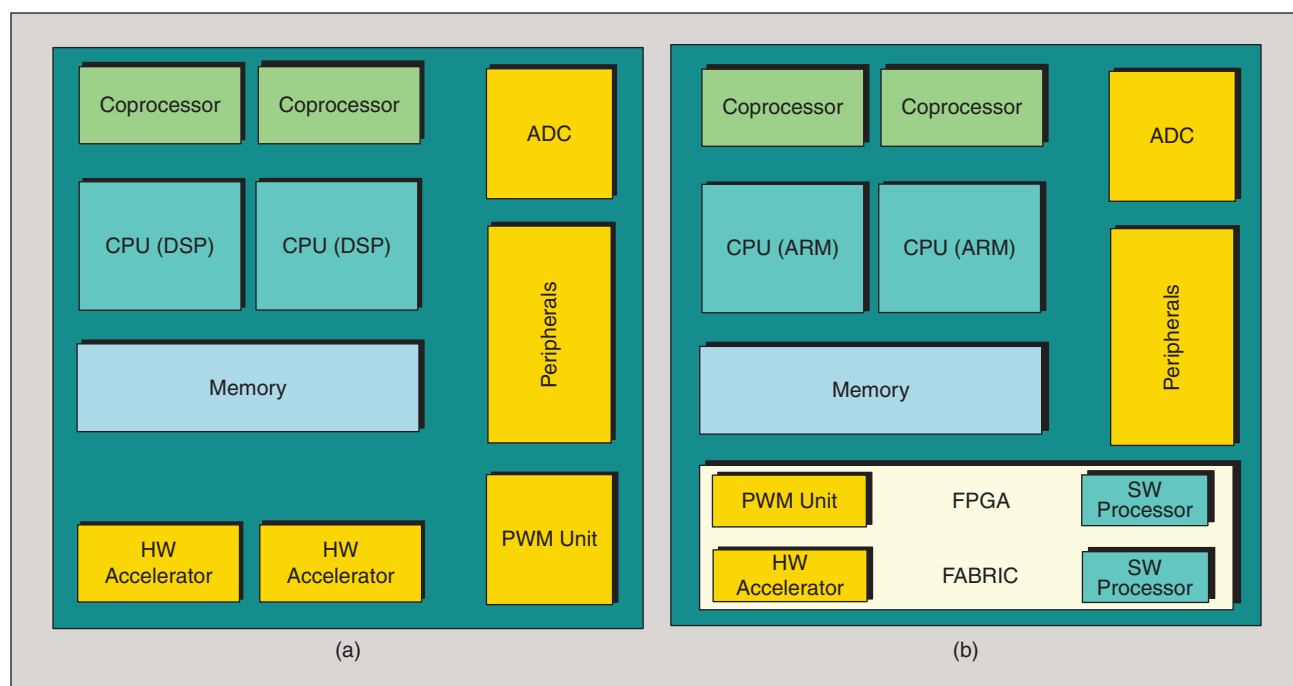


FIGURE 2 – The digital control architectures: (a) SoC DSP-based controllers and (b) SoC FPGA controllers.

Despite a number of limitations, SoC FPGA is one of the most promising digital technologies for implementing smart controllers.

pros and cons of the different types of SoC devices.

Case Studies

Control, Estimation, and Prognosis of a Hybrid FC System

Because of their large number of components like proton-exchange membrane FCs (PEMFCs), electrolyzers associated with hydrogen tanks for long-term storage [12], PV arrays, and power converters, and because of emerging possibilities in terms reinforcement of reliability offered by multistack FCs and interleaved converters [13], modern FC hybrid power systems can be considered very complex. To cope with this complexity, controllers are rapidly evolving by including always more new functionalities such as power sharing [14], impedance spectroscopy for data-based diagnosis [15], prognosis and fault system

control [16], as well as weather and power consumption forecasting.

With SoC FPGAs, the HW processor cores and the FPGA fabric are tightly coupled for such control of complex electrical systems, so that the data communication is achieved with low latency. Therefore, one critical point that needs attention is the priority interrupt management. A vectored interrupt controller (VIC) integrated in the soft-core processor NIOS II (Intel/Altera) or hard-core processors ARM Cortex-R and M is mandatory to ensure the lowest interrupt latency and constant low jitters for real-time applications, compared with general ARM Cortex-A [17]. Thanks to the VIC unit of the Cortex-R5 of Xilinx Zynq UltraScale+, this powerful component is ready to handle critical real-time applications and, due to the integration a quad-core Cortex-A53, it is also highly adapted to high computing

applications. However, considering the reduction of the costs, a soft-core processor solution, such as the NIOS II, may sometimes be a better option than using an oversized SoC FPGA due to its interesting properties: low interrupt latency and HW adaptability to the system to be controlled.

A proof of concept system, shown in Figure 3, was implemented to validate the performance of an SoC FPGA-based smart controller for a hybrid FC system composed of a FC stack and supercapacitors. It is worth mentioning that this plant is emulated in the DS1006 and DS5203 dSPACE boards [14]. All the corresponding blocks in Figure 3 are in solid blue lines. The modules related to the SoC FPGA-based controller are shown in solid red lines in Figure 3. Among them, the FC control and prognosis algorithms have been implemented in a NIOS II on a low-cost Cyclone V board (DE1-SoC Intel/Altera). Finally, all the modules shown in dashed lines, both within the plant or within the SoC FPGA-based controller, are not present in the current study but can be included in future developments, thus demonstrating the high level of scalability of the SoC FPGA-based control framework presented.

The complete HW/software (SW) system represents a hardware-in-the-loop (HIL) platform to validate the algorithms in real time [14]. The SoC FPGA architecture is composed of two pulsewidth modulation (PWM) units, an acquisition unit of six ADCs, and a soft-core base on a NIOS II. The algorithms are executed in three interrupt service routines (ISR) based on three synchronized timer events configured with a sampling time equal to 50 μs for the current loops and PWMs, 500 μs for the power management, and 1 s for a prognosis and health management (PHM) algorithms. The three ISRs use vectorized interrupts with a highest priority (0) for the current controllers (ISR0) and then priority 1 for the power management module (ISR1). The computation times are equal to 7.20 μs, 9.84 μs, and 117 μs, respectively [14].

Figure 4 shows all the main data computed in the emulated system

TABLE 1 – THE ADVANTAGES AND DISADVANTAGES OF THE DIFFERENT DIGITAL TECHNOLOGIES.

CRITERIA	DSP/μC	FPGA	SoC DSP	SoC FPGA
Algorithmic perspective	●○○	●○○	●●○	●●●
Algorithm complexity management	●○○	●○○	●●○	●●●
Rapidity, possibility of parallelism	●○○	●●○ (HW parallelism: concurrency & pipeline)	●●○ (Multicores SW parallelism)	●●● (both HW & SW parallelism)
Accuracy (floating-point capability)	●●●	●●○	●●●	SW: ●●● HW: ●●○
Connectivity	●○○	●○○	●●○	●●○
Analog interface (ADC, DAC)	●●○	●○○	●●○	●○○
Digital interface (number of I/Os)	●●○	●●●	●●○	●●●
Embedded peripherals	●●○	●●○	●●●	●●●
Embedded OS (Internet access,...)	●○○	●○○	●○○	●●●
Flexibility of use	●●○	●●○	●●○	●●○
Coding facilities	●●●	●●○	●●●	●●○
Microarchitecture adaptation, SW processor core implementation, obsolescence risk reduction	○○○	●●●	●○○	●●●
Learning curve	●●●	●●○	●●○	●○○
Cost	€○○○	€€○	€○○○	€€€○

Score: ○○○ bad, ●○○ medium, ●●○ good, ●●● very good. DAC: digital-analog converter.

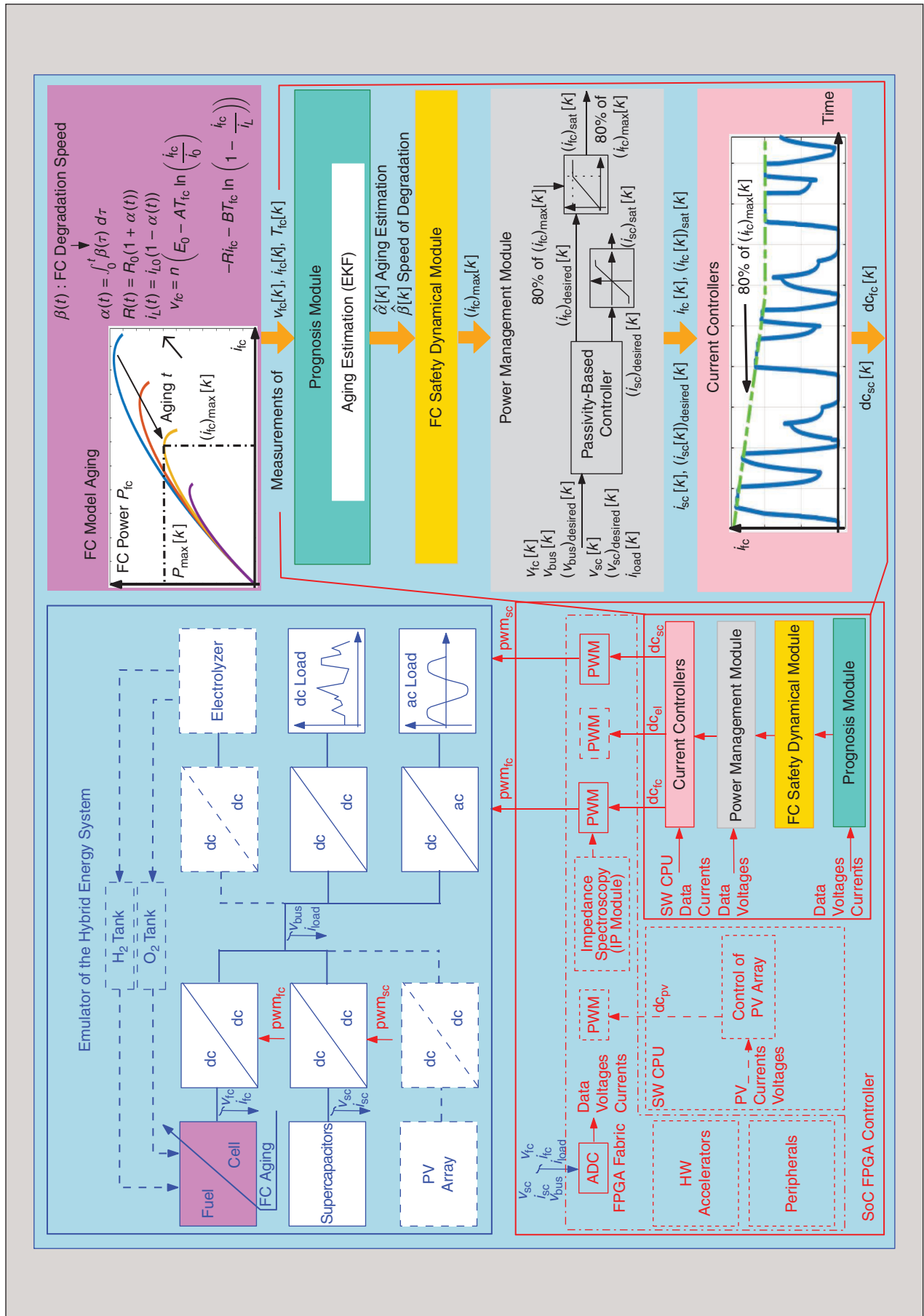


FIGURE 3 – The hybrid FC system architecture.

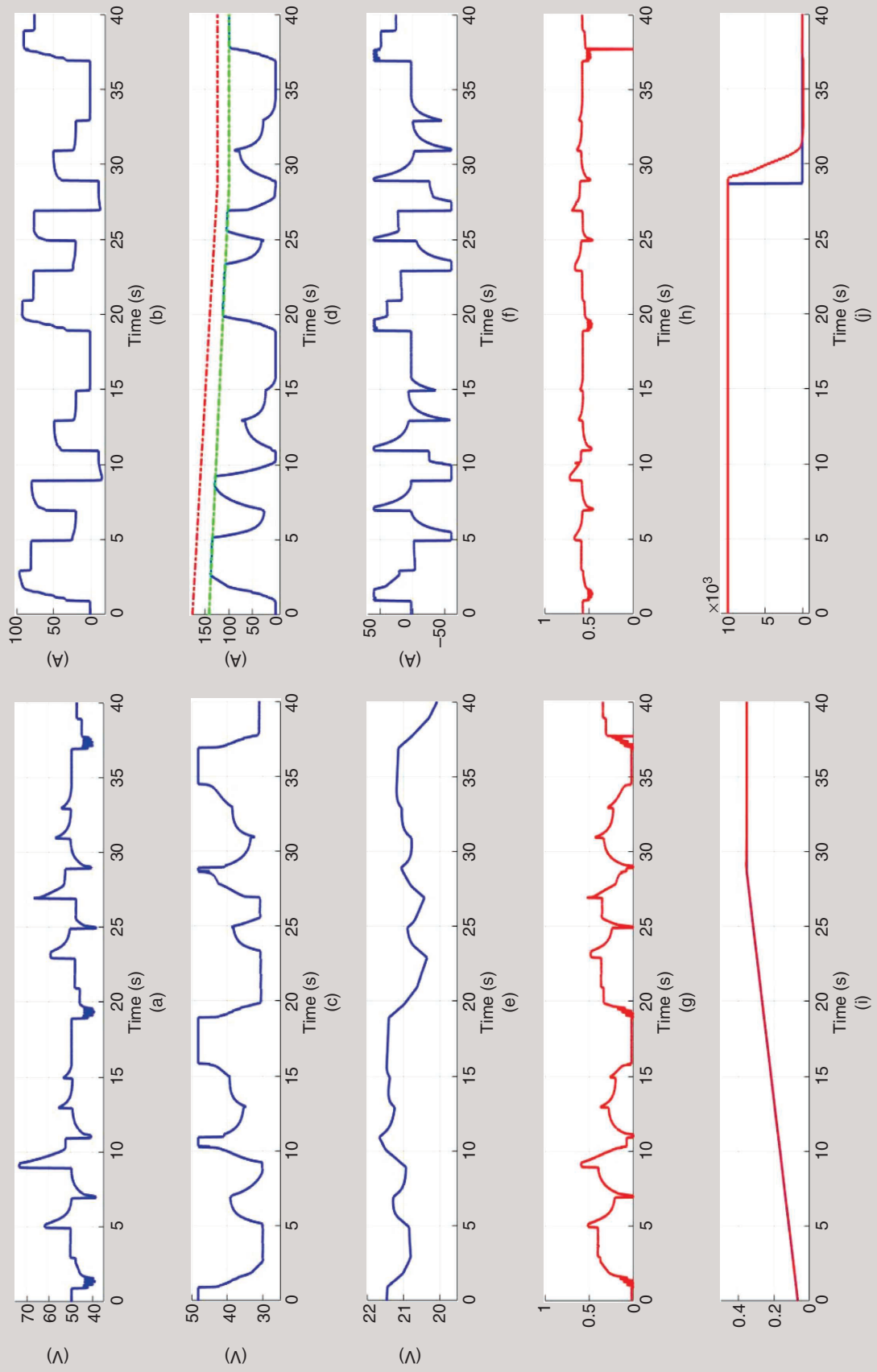


FIGURE 4 – The hybrid FC system HIL results: (a) dc bus voltage v_{bus} , (b) load current i_{load} , (c) FC voltage v_{fc} , (d) FC current i_{fc} , (e) SCs voltage v_{sc} , (f) SCs current i_{sc} , (g) duty cycle d_{cr} , (h) duty cycle d_{sc} , (i) FC SoH $\alpha(t)$ and $\beta[k]$, and (j) speed of degradation $\beta(t)$ and $\beta[k]$. SC: supercapacitor; SoH: state of health.

(blue curves) and in the NIOS II processor (red curves); these colors correspond to those chosen in Figure 3. Aging $\alpha(t)$ of the PEMFC, which has been emulated in the FC model, is estimated online ($\hat{\alpha}[k]$) by the PHM algorithm (here an extended Kalman filter) [14], [18]. The FC safety dynamic module computes the maximum FC current $(i_{fc})_{\max}[k]$ value that must not be exceeded. Notice that the FC current i_{fc} is well controlled by the SoC FPGA-based controller since it does not exceed the defined maximum current fixed to 80% of $(i_{fc})_{\max}[k]$ [see Figure 4(d)]. This means that both the speed of degradation $\beta(t)$ [see Figure 4(j)] and the aging $\alpha(t)$ [see Figure 4(i)] are well estimated by the observer implemented within the NIOS II soft-core processor [14].

Moreover, as the current controllers and peripherals implemented within the FPGA fabric (PWM, ADC) need to be tightly coupled, it appears that the soft-core processor is a valuable option to provide deterministic interrupt, minimum jitters, many possibilities of evolution, and it reduces the risk of obsolescence.

Dynamic Reconfiguration of PV Modules

Shadowing significantly affects PV arrays electrical power production and may lead to the conduction of the modules bypass diodes. Consequently, more than one maximum power point appears in the string power versus voltage (P-V) and current versus voltage (I-V) curves [19]. Depending on the actual shadowing pattern, the adoption of a system permitting change in the electrical connections among the PV modules through a suitable switching matrix [20] is useful. The reconfiguration has to be performed dynamically, because the shading pattern changes during the day, and in a short time interval during which the irradiance level received by the PV cells does not change significantly.

In [21], a theoretical analysis of the problem was proposed, and in [22], an evolutionary algorithm (EA) aimed at dynamically determining the

SoC FPGA devices are not only able to manage complex algorithm online processing, but they can also help to accelerate their execution.

best electrical configuration of the PV modules in a plant formed by more strings was presented.

Figure 5 shows a fixed shadow affecting the PV array and two EA individuals, each corresponding to a specific electrical connection of modules, to form the two parallel connected strings. The green P-V curve corresponds to the static connection, showing a peak power lower than the one the EA determines and corresponding to the configuration with blue P-V curve.

The conjoint HW/SW SoC FPGA-based implementation [23] (Figure 6) consists of the core of the EA, implemented in SW on a bare metal ARM A9 core, and of the fitness function instances that are executed in a couple of dedicated intellectual property (IP) modules within the FPGA fabric. The 12-b fixed-point representation ensures a good tradeoff between the FPGA fabric-consumed area and the loss of accuracy, which is less than 1% than the reference case based on a 32-b floating-point representation. The fitness function IP module is written in C++ and the architectural design space is explored by using the HLS approach [24]. HLS allows to design the HW accelerator through high-level languages, for example, C/C++, by generating production-quality register transfer level (RTL) code that is optimized for the targeted FPGA. The synthesis process transforms automatically a C/C++ source code in an HW description language such as VHDL or SystemVerilog. HLS accelerates verification time over RTL by raising the abstraction level for FPGA HW design. HLS designs are typically verified at a speed that is orders of magnitude faster than RTL ones. The algorithm is preliminary optimized to put into evidence the subroutines to be run in parallel and, by using a counting sort algorithm, permits saving up to 80%

of computation time with respect to the use of a standard bubble sort algorithm. The reduction in the size of the fixed-point divider leads to a 20% reduction in the latency of the fitness function.

Two practical cases, with 100 and 25 samples per module I-V curve, respectively, were implemented on the low-end, Zynq-based board (Zybo from Digilent) at an FPGA clock frequency of 125 MHz. The PV field has 24 modules divided in two parallel connected strings. The EA runs on a population of 48 individuals, for a maximum of 100 generations. The experiments revealed that, if 100 samples per curve are used, two fitness function HW accelerators can be integrated in parallel in the FPGA fabric. For the 25 samples per curve case, three IPs modules can be embedded. The acceleration rate for the 100 sample case is of 2.46 compared to an optimized full SW implementation based on a bare metal ARM A9 core running at 667 MHz, thus leading to a total execution time of 13.218 s. By comparison, the acceleration rate for the 25 sample case is of 2.80, with a total execution time of 2.374 s.

The Next Generation of Smart Controllers for Electrical Energy Systems

The electrical energy sector in Europe will be pushed by the European Union (EU) Green Deal [25] and the EU Recovery Plan [26], also in view of its integration with other energy sectors [27], [28] and with digital technologies for achieving the decarbonization goal.

We now discuss the significant contributions SoC FPGA can bring to the key future developments of renewable generators and hydrogen technology using the two applications presented in the previous section. SoC FPGA will facilitate meeting the EU expectations and targets in other fields, such as

battery management and diagnostic systems (see, for example, [29], [30]).

Monitoring and diagnostic functions will benefit from the decentralized high computational potential SoC FPGAs offer, enabled by the use of the

model-based approach for PV systems [31] and even by running data-driven approaches (for example, [32] for FC applications and [33] for PV systems).

The smart power management area will also profit from SoC FPGAs,

especially by introducing the controller digital twins (DTs) of the used static power converters. DTs have several benefits such as the possibility to make online diagnosis [34] or to study in detail the power losses of a complex

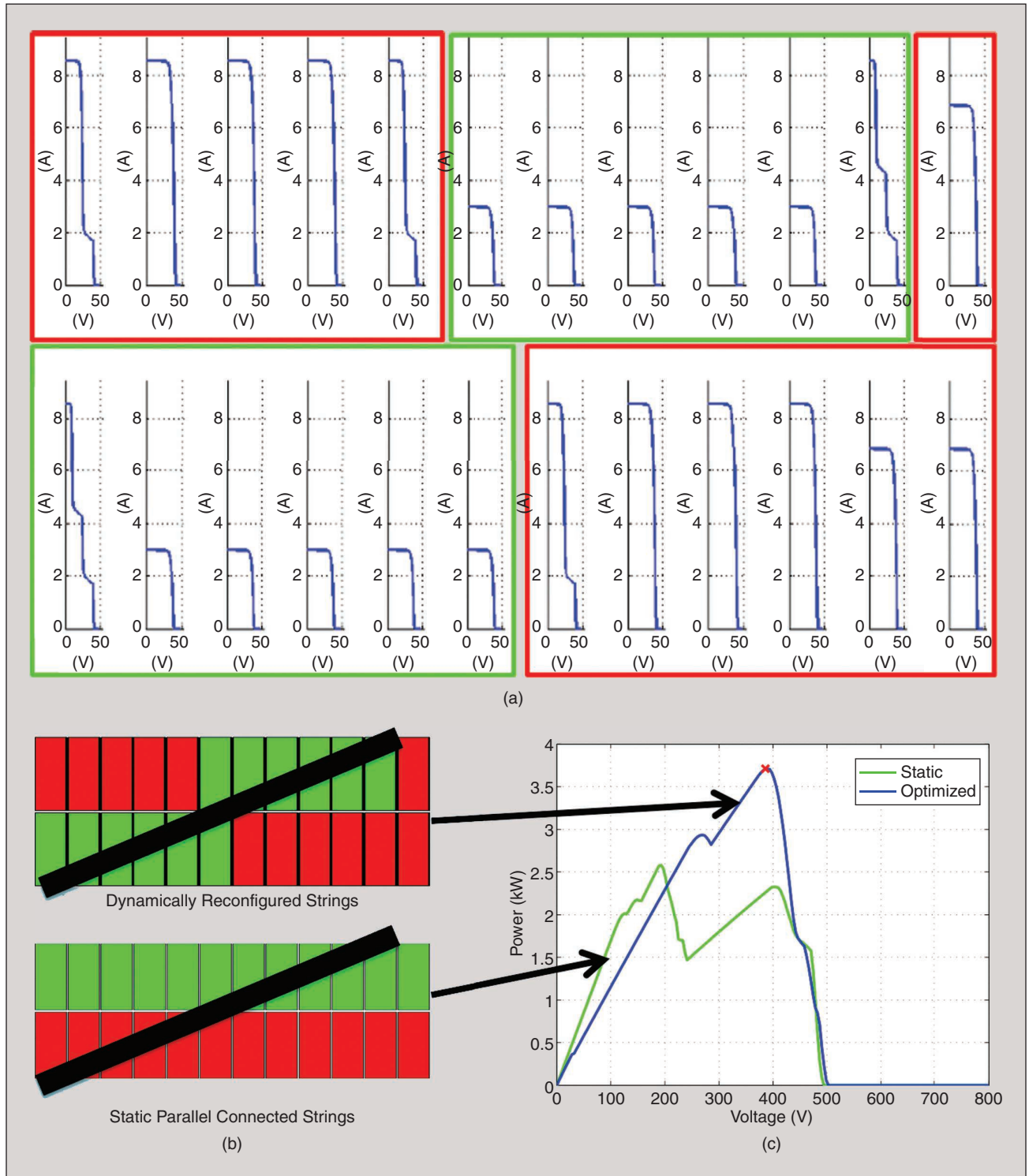


FIGURE 5 – The dynamic reconfiguration of a PV array of two parallel connected strings of 12 modules each. (a) The modules' I-V curves. (b) The strings affected by an oblique shadow. The green and red modules are series connected. (c) The P-V curves corresponding to the static and the reconfigured PV fields.

structure like a modular multilevel converter (MMC) [35]. Finally, the optimized economic dispatching of a microgrid is a good example of the ongoing mutation in terms of control algorithmic needs for modern complex electrical energy systems [36].

These recent works are good illustrations of what could be the next generation of smart controllers for complex electrical energy systems. Beyond the standard control functions (still implemented), these smart controllers will also include additional tasks like diagnosis, fault tolerant capabilities, optimization of the energy

SoC FPGAs also offer a high level of flexibility in terms of microarchitecture.

flow, and/or economical dispatching. These new functionalities can be gathered under the generic name of *smart monitoring*, and it is worth analyzing their impact on the architecture of smart controllers.

In complex electrical energy systems, the first task for smart controllers is to collect and aggregate the measurements coming from all internal

subelements. An analysis of references shows that three approaches are possible to cope with this problem. A typical approach is to use a standard serial communication like controller area network bus [29] between the low-end microcontroller that is in charge of the monitoring of a given cell and the centralized SoC FPGA-based smart controller. A second solution

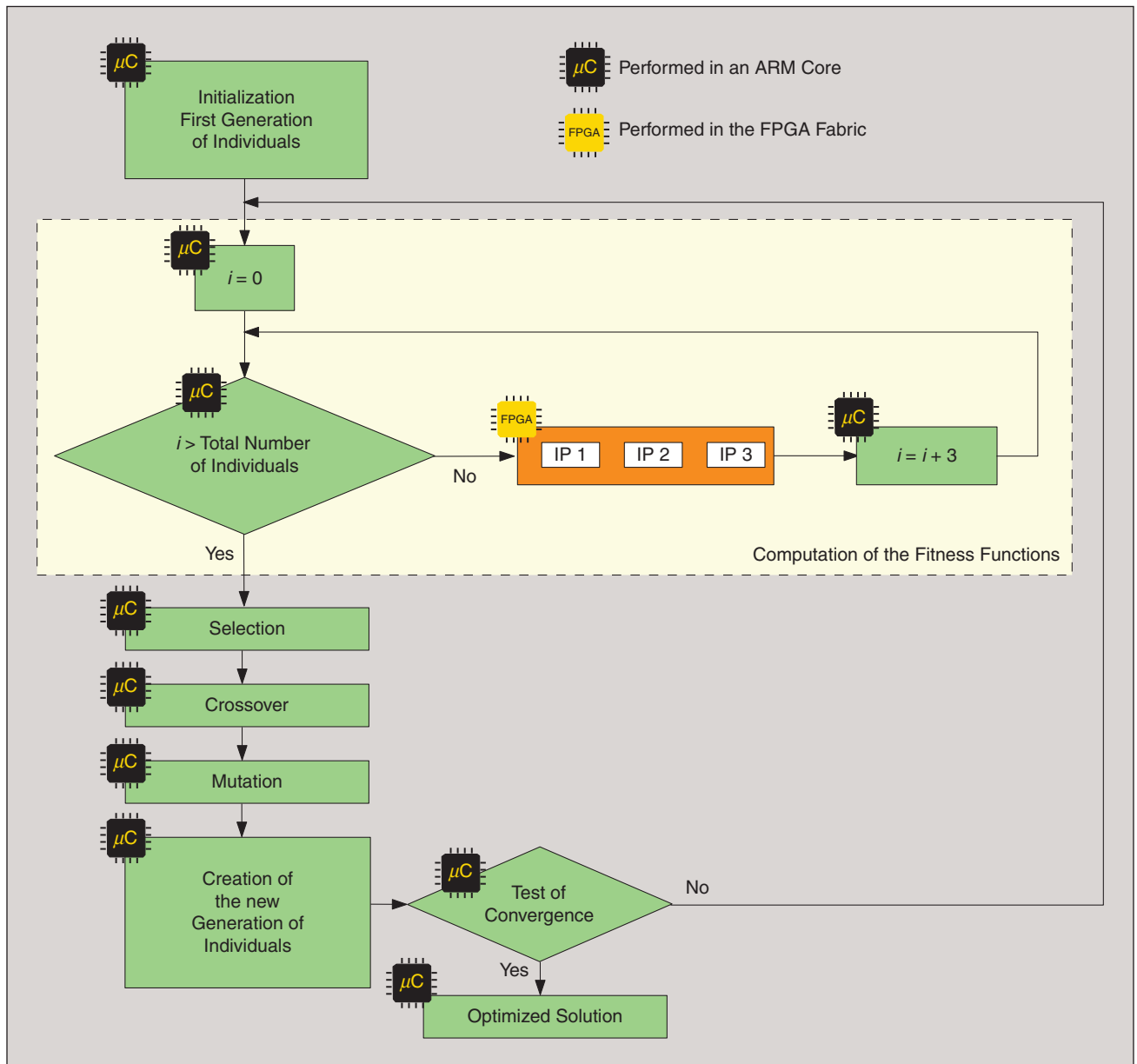


FIGURE 6 – The HW/SW implementation of the PV dynamic reconfiguration algorithm.

is to use a wireless connection (Wi-Fi or Bluetooth) [30], [31], since it offers more flexibility and scalability than a classical serial wired communication. Finally, a more radical approach is to integrate all the necessary front-end analog resources needed to measure and collect the data coming from the cells in an application-specified integrated circuit like that in [32], which also integrates the SoC FPGA-based smart controller. This solution is highly integrated but very specific and costly to design. However, and if, as expected, the market of the industrial IoT will be booming, one should expect SoC FPGA manufacturers to propose that future new devices include more analog capabilities than today. We are already seeing this with the Xilinx RF-SoC device, which is devoted to the 5G SW radio market [37].

The main tasks to be performed by the smart controller are diagnosis [31], [33], health monitoring [18], and energy management [23], [36]. Sometimes higher integration is the key objective [30], where the smart controller performs in parallel both battery management system and charger functionalities.

Depending on the time scale of these smart monitoring tasks, the controller has to apply hard real-time operating conditions (from microseconds to millisecond, with a full timing determinism, achieved by timer interruptions and with a bare metal configuration of the processor to minimize the latency) [18], [35], or soft real-time operating conditions (from seconds to hours when timing determinism is less critical). In this case, it is of great importance to execute the smart monitoring tasks as processes of a real-time OS like an embedded Linux [29], thus profiting from its communication facilities. Even when hard real-time operating conditions are mandatory, it is still possible to dedicate one core processor of the SoC FPGA to run a Linux OS, while the other one is bare metal and devoted only to the critical control tasks [38] (“asymmetric multiprocessing”).

Regarding the nature of smart monitoring strategies, most are based on

the simulation of a plant model [29]. Some of these approaches require an optimization problem that has to be solved online [23], [31], [36]. The rest of these studies, like those integrating a DT [34], [35], are based on estimators or observers [18]. However, whatever the smart controller has to execute, a stochastic optimization problem or an embedded DT, the computing load is high. Therefore, it is interesting to analyze how the HW/SW partitioning, which consists of choosing which parts of the control algorithm are implemented in a processor and which are implemented in an HW accelerator, is conducted; for the EA-based optimization, the main body of the EA is implemented in SW and the fitness function instances are implemented as HW accelerators [23], [31]. As for estimators and observers, the HW/SW partitioning is usually based on the dynamics of the model to emulate; a slow temperature estimator is naturally implemented in SW, while the battery state-of-charge estimator is done in HW [29]. In [35], because the submodule estimators of the MMCs are prone to parallelization, they are placed in the FPGA fabric. But in [34], a full FPGA implementation is performed, which results as the only choice due to the conjoint short dynamics of the emulated power converters and the complexity of the stochastic models used.

With the progress in machine learning methods, data-driven approaches are increasingly popular for the diagnosis of complex electrical energy systems. They concern classification [32] or regression techniques [33], both requiring a complex offline training process, but the online inference process may be relatively simple. However, in many neural network (NN) classification or regression problems, the trained NN is fed by new incoming data from the plant. This means that, unlike [33], the local smart controller has to implement an inferred NN. An inferred NN, as a simplified version of an optimally trained deep NN, has a reduced power and latency for meeting edge applications requirements. The deep NN is trained offline; then,

through pruning and quantization methods, the groups of artificial neurons that rarely or never fire are removed and the numeric precision of the weights is reduced, so that a reduced model size and a faster computation are achieved at the cost of minimal reductions in predictive accuracy [39]. Based on the parallel characteristics inherent in such algorithms, an FPGA-based or GPU-based implementation is highly recommended [40].

The preceding overview reveals that most smart monitoring applications are implemented in an SoC FPGA device since these heterogeneous computing platforms reached very good computing performance and enable architecture customization thanks to the FPGA fabric. With the help of a real-time OS like embedded Linux, these devices are easily connectable to Internet so they are good candidates to handle one of the biggest mutations currently experienced in digital controllers—the transformation of the “local embedded controllers” into edge computing platforms (ECPs). So the aforementioned “smart controllers” are not only able to handle locally complex control functions and smart monitoring tasks, but they can also be part of a larger control system that distributes some tasks to a remote cloud computing platform (CCP). This transformation is directly derived from the industrial IoT concept [41]. The distribution of the tasks between the ECP and the CCP can be seen as an evolution of the embedded control concept, with smart monitoring tasks processed locally. However, in [33] and [36], a different philosophy has been proposed: all the prediction tasks are achieved in advance on an hourly/daily basis and the ECP only has to compare the information received from the plant with these predictions. Thus, the computing load is clearly moved remotely into a CCP and, as consequence, the ECP can remain very light, as in [36], where a single DSP chip is sufficient to implement a decision maker based on simple tests.

To conclude, the fact that SoC FPGA-based ECPs are able to collect

data from the cell unit controllers, use it locally to execute smart monitoring tasks, and interact with a CCP where hourly/daily training of NN is achieved or where other slow supervising and storage tasks are being performed, opens new interesting lines of research. One of them is the opportunity to enlarge significantly the size of the electrical energy systems to manage [36], where the same CCP can handle the economic dispatching forecasts for several microgrids. The next step will be to integrate the possibilities for cooperation between different electrical energy systems, but reinforcing the security and the privacy of the connections between the ECP and the CCP will be a concern. Finally, any complex electrical energy system can be monitored over its entire lifespan by sending and storing on a daily basis relevant data from an ECP to a CCP. A lot of effort has to be dedicated to this topic as part of the energy IoT future research.

Conclusion

SoC FPGA can make significant contributions to the key developments in complex electrical energy systems, especially those including renewable generators and those employing hydrogen technology. We detailed some advantages and limitations through the two specific applications presented in the case studies. One concerns a FC hybrid electric system controlled by passivity-based power management associated with an aging prognosis algorithm. The other reports on the SoC FPGA implementation of a control system able to optimize online the dynamical configuration of a partially shadowed PV field.

In addition to these two case studies, we also analyzed in detail a series of recently reported results on smart controllers for complex electrical energy systems, highlighting the importance of the increasing number of smart monitoring tasks performed by this new generation of controllers, for example, diagnosis, prognosis, fault tolerant capabilities, optimization of the energy flow, and/or economical dispatching.

SoC FPGA to easily communicate with both the system to be controlled and the remote cloud services.

Despite a number of limitations like the cost (that is costs higher than for other technologies like SoC DSPs), a limited analog interface (analog-to-digital, digital-to-analog), and a designer's longer learning curve for optimal use, SoC FPGA is one of the most promising digital technologies for implementing smart controllers. By investigating with care the implications in terms of implementation of these new smart monitoring tasks, we showed that SoC FPGA devices are not only able to manage complex algorithm online processing (such as an EA optimization or a DT), but they can also help to accelerate their execution by parallelizing into customized HW accelerators several computationally demanding subtasks like fitness function calculation.

Furthermore, thanks to their highly performing FPGA fabric, SoC FPGAs also offer a high level of flexibility in terms of microarchitecture. A good illustration of this is the possibility for the designer to add one or more simple SW core processors, thus relieving the device processing system of low-level time-consuming tasks.

Finally, we pointed out another important advantage—the ability of SoC FPGA to easily communicate with both the system to be controlled, thanks to a very large number of I/Os, and the remote cloud services, because it can easily embed a Linux OS. This makes an SoC FPGA-based smart controller a high-performing ECP that is able to address incoming challenges, in terms of complexity and storage, brought on by data-driven approaches.

Biographies

Eric Monmasson (eric.monmasson@u-cergy.fr) is a full professor at CY Cergy Paris University, Cergy-Pontoise, 95031, France. His research interests include the design of field-programmable gate array-based and system-

on-chip-based digital controllers for industrial and electrical systems. He was the chair of the technical committee on Electronic Systems-on-Chip of the IEEE Industrial Electronics Society from 2008 to 2011. He is an associate editor of *IEEE Transactions on Industrial Informatics* and *IEEE Industrial Electronics Magazine*. He has authored or coauthored three books and more than 200 scientific papers. He is a Senior Member of IEEE.

Mickaël Hilairet (mickael.hilairet@univ-fcomte.fr) is a full professor at the University of Bourgogne Franche-Comté, Belfort, 90010, France, and director of the energy department at the CNRS FEMTO-ST Laboratory, Belfort, 90010, France. His research interests include industrial informatics for the control, diagnosis, and prognosis of electrical systems. He is an associate editor of *IEEE Transactions on Industrial Electronics* and *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*. He was chair of the Technical Committee on Electronic Systems on Chip of the IEEE Industrial Electronics Society in 2018–2019. He is a Member of IEEE.

Giovanni Spagnuolo (gspagnuolo@unisa.it) is a full professor at the University of Salerno, Salerno, 84084, Italy. His research interest is renewable energy systems. He is an editor of *IEEE Journal of Photovoltaics* and an associate editor of *IEEE Open Journal of the Industrial Electronics Society*. He was on the 2015 Thomson Reuters list of Most Influential Minds. He is the coauthor of five international patents. He is a Fellow of IEEE and a member of the IEEE European Public Policy Initiative Working Group on Energy.

Marcian N. Cirstea (marcian.cirstea@aru.ac.uk) earned his Ph.D. degree in electronic and electrical engineering from Nottingham Trent University, U.K., in 1996. Currently, he is a full professor and the head of the school of computing

at Anglia Ruskin University, Cambridge, CBI IPT, U.K. His research interests include design methods for digital systems design using field-programmable gate arrays (FPGAs) and power systems applications using FPGAs. He is the founder of the Technical Committee on Electronic System-on-Chip of IEEE Industrial Electronics Society, and he also coordinated a European renewable energy project. He is a Senior Member of IEEE and a fellow of the Institution of Engineering and Technology.

References

- [1] C. Buccella, C. Cecati, and H. Latafat, "Digital control of power converters—a survey," *IEEE Trans. Ind. Informat.*, vol. 8, no. 3, pp. 437–447, 2012. doi: 10.1109/TII.2012.2192280.
- [2] H. A. Toliyat, *DSP-Based Electromechanical Motion Control*. Boca Raton, FL: CRC Press, 2003.
- [3] E. Monmasson, L. Idkhajine, and M. W. Naouar, "FPGA-based controllers," *IEEE Ind. Electron. Mag.*, vol. 5, no. 1, pp. 14–26, 2011. doi: 10.1109/MIE.2011.940250.
- [4] S. M. Trimberger, "Three ages of FPGAs: A retrospective on the first thirty years of FPGA technology," *Proc. IEEE*, vol. 103, no. 3, pp. 318–331, 2015. doi: 10.1109/JPROC.2015.2392104.
- [5] J. Millan, P. Godignon, X. Perpina, A. Perez-Tomas, and J. Rebollo, "A survey of wide bandgap power semiconductor devices," *IEEE Trans. Power Electron.*, vol. 29, no. 5, pp. 2155–2163, 2014. doi: 10.1109/TPEL.2013.2268900.
- [6] W. Tu, G. Luo, Z. Chen, C. Liu, and L. Cui, "FPGA implementation of predictive cascaded speed and current control of PMSM drives with two-time-scale optimization," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5276–5288, 2019. doi: 10.1109/TII.2019.2897074.
- [7] J. J. Rodríguez-Andina, M. D. Valdés-Peña, and M. J. Moure, "Advanced features and industrial applications of FPGAs—A review," *IEEE Trans. Ind. Informat.*, vol. 11, no. 4, pp. 853–864, 2015. doi: 10.1109/TII.2015.2431223.
- [8] T. Adegbiya, A. Rogacs, C. Patel, and A. Gordon-Ross, "Microprocessor optimizations for the internet of things: A survey," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 1, pp. 7–20, 2018. doi: 10.1109/TCAD.2017.2717782.
- [9] "Tms320f2837xd dual-core delfino™ micro-controllers," Texas Instruments, Dallas, TX, 2013. Accessed: Sept. 30, 2019. [Online]. Available: <https://www.ti.com/product/TMS320F28377D>
- [10] R. F. Molanes, J. J. Rodríguez-Andina, and J. Fariña, "Performance characterization and design guidelines for efficient processor-FPGA communication in cyclone v FPGAs," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4368–4377, 2018. doi: 10.1109/TIE.2017.2766581.
- [11] Arm. [Online]. Available: <https://www.arm.com/resources/designstart/designstart-fpga>
- [12] J. Pierson et al., "Datazero: Datacenter with zero emission and robust management using renewable energy," *IEEE Access*, vol. 7, p. 146, July 2019. [Online]. Available: <https://publweb.femto-st.fr/tntnet/entries/15835/documents/author/data> doi: 10.1109/ACCESS.2019.2930368.
- [13] S. Abuzant, S. Jemei, D. Hissel, L. Boulon, K. Agbossou, and F. Gustin, "A review of multi-stack PEM fuel cell systems: Advantages, challenges and on-going applications in the industrial market," in *Proc. 14th IEEE Vehicle Power Propulsion Conf. (VPPC)*, 2017, pp. 1–6.
- [14] S. Kong, M. Bressel, M. Hilairet, and R. Roche, "Advanced passivity-based, aging-tolerant control for a fuel cell/super-capacitor hybrid system," *Control Eng. Pract.*, vol. 105, p. 104,636, Sept. 2020. doi: 10.1016/j.conengprac.2020.104636.
- [15] D. Depernet, A. Narjiss, F. Gustin, D. Hissel, and M.-c. Pera, "Integration of electrochemical impedance spectroscopy functionality in proton exchange membrane fuel cell power converter," *Int. J. Hydrogen Energy*, vol. 41, no. 11, pp. 5378–5388, Mar. 2016. doi: 10.1016/j.ijhydene.2016.02.010.
- [16] E. Dijoux, N. Yousfi Steiner, M. Benne, M.-C. Péra, and B. Grondin-Perez, "A review of fault tolerant control strategies applied to proton exchange membrane fuel cell systems," *J. Power Sources*, vol. 359, pp. 119–133, Aug. 2017. doi: 10.1016/j.jpowsour.2017.05.058.
- [17] "Real-time challenges and opportunities in SoCs," ALTERA, Intel-FPGA, San Jose, CA, White Paper, Mar. 2013. [Online]. Available: <https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/wp-01190-real-time-socs.pdf>
- [18] M. Bressel, M. Hilairet, D. Hissel, and B. Ould-Bouamama, "Model-based aging tolerant control with power loss prediction of proton exchange membrane fuel cell," *Int. J. Hydrogen Energy*, vol. 45, no. 19, pp. 11242–11254, Apr. 2020. doi: 10.1016/j.ijhydene.2018.11.219.
- [19] G. Petrone, C. R. Paja, and G. Spagnuolo, *Photovoltaic Sources Modeling*, 1st ed. Hoboken, NJ: IEEE Wiley, 2017.
- [20] G. Petrone, G. Spagnuolo, B. Lehman, Y. Zhao, C. R. Paja, and M. O. Gutierrez, "Photovoltaic arrays dynamical reconfiguration: Fighting mismatched conditions and meeting load requests," *IEEE Ind. Electron. Mag.*, vol. 9, no. 1, pp. 62–76, 2015. doi: 10.1109/MIE.2014.2360721.
- [21] M. Orozco-Gutierrez, G. Spagnuolo, J. Ramirez-Scarpetta, G. Petrone, and C. Ramos-Paja, "Optimized configuration of mismatched photovoltaic arrays," *IEEE J. Photovolt.*, vol. 6, no. 5, pp. 1210–1220, Sept. 2016. doi: 10.1109/JPHOTOV.2016.2581481.
- [22] P. Carotenuto, A. D. Cioppa, A. Marcelli, and G. Spagnuolo, "An evolutionary approach to the dynamical reconfiguration of photovoltaic fields," *Neurocomputing*, vol. 170, pp. 393–405, Dec. 2015. doi: 10.1016/j.neucom.2015.04.094.
- [23] G. Petrone, F. Serra, G. Spagnuolo, and E. Monmasson, "SoC implementation of a photovoltaic reconfiguration algorithm by exploiting a HLS-based architecture," *Mathematics Comput. Simulat.*, vol. 158, pp. 520–537, Apr. 2019. doi: 10.1016/j.matcom.2018.12.013.
- [24] J. Cong, B. Liu, S. Neuendorffer, J. Noguera, K. Vissers, and Z. Zhang, "High-level synthesis for FPGAs: From prototyping to deployment," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 30, no. 4, pp. 473–491, Apr. 2011. doi: 10.1109/TCAD.2011.2110592.
- [25] "Communication on the European Green Deal," Accessed: July 8, 2020. [Online]. Available: https://ec.europa.eu/info/publications/communication-european-green-deal_en
- [26] "EU recovery plan," European Commission, Brussels, Belgium. Accessed: July 8, 2020. [Online]. Available: https://ec.europa.eu/info/live-work-travel-eu/health/coronavirus-response/recovery-plan-europe_en
- [27] "Energy system integration strategy," European Commission, Brussels, Belgium Accessed: July 8, 2020. [Online]. Available: https://ec.europa.eu/energy/sites/ener/files/energy_system_integration_strategy.pdf
- [28] "Hydrogen strategy," European Commission, Brussels, Belgium. https://ec.europa.eu/energy/sites/ener/files/hydrogen_strategy.pdf
- [29] R. Morello et al., "Advances in Li-ion battery management for electric vehicles," in *Proc. IECON 2018 – 44th Annu. Conf. IEEE Ind. Electron. Soc.*, 2018, pp. 4949–4955. doi: 10.1109/IECON.2018.8591185.
- [30] T. Gherman, M. Ricco, J. Meng, R. Teodorescu, and D. Petreus, "Smart integrated charger with wireless BMS for EVs," in *Proc. IECON 2018 – 44th Annu. Conf. IEEE Ind. Electron. Soc.*, 2018, pp. 2151–2156. doi: 10.1109/IECON.2018.8591253.
- [31] M. Garaj, K. Y. Hong, H. S.-H. Chung, J. Zhou, and A. Lo, "Photovoltaic panel health diagnostic system for solar power plants," in *Proc. IEEE Appl. Power Electron. Conf. Expo. (APEC), Anaheim, CA*, 2019, pp. 1078–1083.
- [32] Z. Li, R. Outbibi, S. Giurgea, D. Hissel, A. Giraud, and P. Couderc, "Fault diagnosis for fuel cell systems: A data-driven approach using high-precision voltage sensors," *Renew. Energy*, vol. 135, pp. 1435–1444, May 2019. doi: 10.1016/j.renene.2018.09.077.
- [33] S. Leva, M. Mussetta, and E. Ogliari, "PV module fault diagnosis based on microconverters and day-ahead forecast," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3928–3937, 2019. doi: 10.1109/TIE.2018.2879284.
- [34] M. Milton, C. D. L. O, H. L. Ginn, and A. Benigni, "Controller-embeddable probabilistic real-time digital twins for power electronic converter diagnostics," *IEEE Trans. Power Electron.*, vol. 35, no. 9, pp. 9850–9864, 2020. doi: 10.1109/TPEL.2020.2971775.
- [35] Z. Shen and V. Dinavahi, "Real-time MPSoC-based electrothermal transient simulation of fault tolerant mmc topology," *IEEE Trans. Power Del.*, vol. 34, no. 1, pp. 260–270, 2019. doi: 10.1109/TPWRD.2018.2866520.
- [36] S. Wang, X. Wang, and W. Wu, "Cloud computing and local chip-based dynamic economic dispatch for microgrids," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 3774–3884, 2020. doi: 10.1109/TSG.2020.2983556.
- [37] B. Farley, J. McGrath, and C. Erdmann, "An all-programmable 16-nm RFSoc for digital-RF communications," *IEEE Micro*, vol. 38, no. 2, pp. 61–71, 2018. doi: 10.1109/MM.2018.022071136.
- [38] E. Zafra et al., "Efficient FPSoc prototyping of FCS-MPC for three-phase voltage source inverters," *Energies*, vol. 13, no. 5, p. 1074, Mar. 2020. doi: 10.3390/en13051074.
- [39] "Deep learning training and inference," Intel-FPGA, San Jose, CA. Accessed: Nov. 24, 2020. [Online]. Available: <https://www.intel.com/content/www/us/en/artificial-intelligence/posts/deep-learning-training-and-inference.html>
- [40] X. Ranaño, R. Fernández Molanes, C. González-Val, J. Rodríguez-Andina, and J. Fariña, "Performance evaluation of state-of-the-art edge computing devices for DNN inference," in *Proc. IECON 2020—46th Annu. Conf. IEEE Ind. Electron. Soc.*, 2020, pp. 2286–2291. doi: 10.1109/IECON43393.2020.9255055.
- [41] W. Dai, H. Nishi, V. Vyatkin, V. Huang, Y. Shi, and X. Guan, "Industrial edge computing: Enabling embedded intelligence," *IEEE Ind. Electron. Mag.*, vol. 13, no. 4, pp. 48–56, 2019. doi: 10.1109/MIE.2019.2943283.





Power Electronics

My Life and Vision for the Future

Power electronics has ushered in a new kind of industrial revolution in the 21st century because of its important roles in energy conservation, renewable energy systems (RESs), bulk energy storage, electric and hybrid electric vehicles, and smart grid applications besides its traditional role in high-efficiency energy systems. Future advances in power electronics will occur mainly in two directions: wide-bandgap (WBG) power semiconductor devices and complex smart grid systems.

Power electronics, a well-known technology, is concerned with the conversion and control of electrical energy at high efficiency with switching-mode power semiconductor devices, and its applications include dc and ac power supplies, electrochemical processes, heating and lighting control, electronic welding, power line VAR and harmonic compensation, high-voltage dc (HVdc) systems, flexible ac transmission systems (FACTSs), photovoltaic (PV) and fuel cell power conversion, solid-state circuit breakers, high-frequency heating, and motor drives. Power electronics is said to have ushered in a new kind of industrial revolution because of its important role in energy conservation, RESs, bulk energy storage, EVs and HEVs, and smart grid applications in addition to its traditional roles in industrial automation and high efficiency energy systems.

The 21st century is often defined as the golden era of power electronics

after the main technology evolution stabilized in the later part of the past century, although the momentum of technology evolution will continue in this century. Often, it is said that power electronics has brought in the third industrial revolution, where the first revolution was brought by the invention of heat engines and the second revolution was brought by the invention of transistors. In this century, power electronics will play significant role in the green energy revolution and help solve the problems of global climate change, which have devastating effects on our society. As environmental regulations are tightened and energy prices increase, power electronics applications will proliferate everywhere—industrial, commercial,

residential, transportation, aerospace, military, and utility systems.

In this article, a brief review of power electronics evolution and prognosis for the future will be given based on my knowledge and experience over the span of more than 50 years. Suffice it to say that technology prediction is always difficult because our past knowledge can only be projected to give a vision for the future. Any new invention may alter the course of a technology, which has happened many times in the history of power electronics. Since I devoted my entire life to the field of power electronics, my personal journey through it will be embedded briefly within the contents of this article [9]. Again, it is difficult to give a future perspective of this vast field in the

THE 21ST CENTURY IS OFTEN DEFINED AS THE GOLDEN ERA OF POWER ELECTRONICS AFTER THE MAIN TECHNOLOGY EVOLUTION STABILIZED IN THE LATER PART OF THE PAST CENTURY.

MY GRADUATE EDUCATION IN THE UNIVERSITY OF WISCONSIN (1958–1960)

I was selected by the Government of India under the United States Agency for International Development Program (then known as the *Technical Cooperation Mission*) to study for an M.S. degree at the University of Wisconsin, Madison. I had to sign a contract that after the completion of my studies, I would have to serve in an Indian university for a period of three years. The University of Wisconsin had a large industrial electronics laboratory, where there were experiments on thyatron dc motor drives, ignitron welding controls, high-power polyphase mercury-arc rectifiers, and so on. My research was the investigation of line current harmonics with polyphase rectifier loads. In addition to these studies, I was given intensive training on engineering education.

several pages of this article. The areas of power semiconductor devices, power converters, and machine drives, which are the main components of power electronics, will be covered in this article. In addition, some discussion on the role of power electronics in the smart grid and RESs as well as in climate change problems will also be included.

Classical Power Electronics

The history of classical power electronics [1] began at the dawn of 20th century (1902) with the invention of the glass-bulb mercury-arc rectifier by the American inventor Peter Cooper Hewitt. He later modified glass bulbs by steel tank for higher power. The introduction of the grid by Langmuir (1914) permitted conversion as well as control of electrical power. Slepian later introduced ignitron converters in 1933. The hot-cathode gas tube thyatron rectifier was invented by GE in 1926. Electrical machines, in fact, have a longer history, and many conversion and control functions were possible

with the help of machines. The advent of machines in the 19th century and the commercial availability of electrical power started around the same time.

The War on Currents (ac versus dc) was started by Thomas Edison (1847–1931) and Nikola Tesla (1856–1943), but history eventually favored the superiority of ac for general industrial applications. The era of magnetic amplifiers (MAs) with saturable core reactors (using magnetic materials like Deltamax, Supermalloy, and so on) started during the War on Currents and permitted similar conversion and control functions. The ruggedness and reliability of

MAs proved very important for military applications. It is interesting to note that E.F.W. Alexanderson of GE Corporate Research and Development, Schenectady, New York, used the MA technology to design and build a 70-kW, 100-kHz alternator to establish a radio communication link between the United States and Europe.

IN THIS CENTURY, POWER ELECTRONICS WILL PLAY SIGNIFICANT ROLE IN THE GREEN ENERGY REVOLUTION AND HELP SOLVE THE PROBLEMS OF GLOBAL CLIMATE CHANGE.

MY DOCTORAL STUDIES AND RESEARCH IN MAs (1960–1971)

After returning to India, I joined the Indian Institute of Engineering Science and Technology Shibpur (IEST Shibpur) and started teaching on industrial electronics, advanced electrotechnology, and hydroelectric plants. Simultaneously, I started my doctoral studies in projects related to MAs. Dr. Herbert Storm, an internationally known expert in MAs in GE Corporate Research And Development (GE-CRD), inspired me to do research in this area and agreed to advise me remotely on my projects. My research projects were somewhat of a hybrid, using Deltamax saturable cores, silicon (Si) power diodes, bipolar junction transistors (BJTs), and thyristors. My projects were magnetic servo amplifiers for position control with two-phase induction servomotors and multichannel telemetry encoding systems using Ramey MAs with BJTs and thyristors. After the completion of my Ph.D. degree, I did a number of research projects with graduate students in MAs before emigrating to the United States in 1971. My first article was "Electronic speed control of motors," published in 1962 in the *Journal of the Institution of Engineers (India)* [S1]. The term *power electronics* was introduced in 1960s after the invention of the thyristor.

REFERENCE

[S1] B. K. Bose, "Electronic speed control of motors," *J. Inst. Eng. (India)*, pp. 172–182, Sep. 1962.

The Era of Modern Power Electronics

Power Semiconductor Devices

The era of modern power electronics started with the advent of power semiconductor devices, which constitute the heart of power electronics. In fact, the modern power electronics evolution has been possible primarily due to device evolution. Of course, the advent of novel converter topologies, pulsewidth modulation (PWM) techniques, analytical and simulation methods, advanced CAD tools, and control and estimation techniques, along with digital control hardware and software, contributed to this evolution. The era of solid-state electronics started with the invention of transistors by Bell Laboratory in 1948, and the same laboratory also invented the PNP-triggering thyristor or silicon (Si)-controlled rectifier (SCR) in 1956. GE commercially introduced the thyristor in 1958. Si power diodes appeared slightly before in 1956.

The advent of SCRs essentially started the modern power electronics evolution. Then, the other power semiconductor devices, such as the TRIAC, gate turn-off thyristor (GTO), power bipolar junction transistors (BJTs), and power MOSFETs, gradually came. The invention of the insulated-gate bipolar transistor (IGBT) by GE was a significant milestone in the history of power semiconductors. It is interesting to note that initially, IGBTs had a thyristor-like latching problem, and the device was known as an *insulated-gate rectifier (IGR)*. Akio Nakagawa of Japan solved the latching problem and helped the commercial introduction of IGBT. The integrated gate-commutated thyristor (IGCT), which is basically a hard-driven GTO, was invented by ABB. In high-power applications, IGBTs and IGCTs are close competitors, but IGBTs are now generally preferred.

We are now on the verge of a new era of WBG power semiconductor devices, such as Si carbide (SiC) and gallium nitride (GaN), that promise higher power with higher efficiency

and higher frequency of power electronic apparatus. The commercial introduction of SiC and GaN transistors in 2010 and 2015, respectively, was another significant milestone in the history of power semiconductor devices. Their applications are now growing extensively. These, along with the next generation ultra-WBG (UWBG) devices, will bring a renaissance in power electronics.

Table 1 summarizes the material property comparison [2] of Si, 4H-SiC, and GaN. The table also includes UWBG materials like gallium oxide (Ga_2O_3) and diamond. The other potential UWBG materials, like boron nitride (BN) and aluminum-gallium-nitride (AlGaN), are not included in the table.

Currently, the 4H-SiC structure is used for device manufacturing due to its higher carrier mobility. Note that the bandgap of SiC and GaN is typically three times higher than Si, giving breakdown field strengths of these materials that are 10 times higher than Si. This means that these devices can be built with a higher blocking voltage, lower leakage current, higher T_j , and higher switching frequency. The thinner and more highly doped drift layer of SiC devices leads to a lower drift resistance and lower saturation voltage, and therefore, a low conduction loss. For a GaN device (which has a lateral structure unlike SiC), the lower conduction loss is contributed by high-electron mobility [called *high-electron mobility transistor (HEMT)*] and higher saturation velocity. The higher thermal conductivity of SiC permits efficient thermal management.

However, the thermal conductivity of GaN is low and comparable to that of Si, but lower on-state losses make this problem less severe. The UWBG materials, like Ga_2O_3 , diamond, and so on, have a higher breakdown field, and their properties are also listed in Table 1. Note that the thermal conductivity of diamond is very high, whereas this parameter is very low for Ga_2O_3 . Diamond appears to be the ultimate material because of its wider bandgap, higher carrier mobility, and higher thermal conductivity. However,

diamond is an extremely hard material. The exploration of the diamond power semiconductor is very challenging and needs the coordinated efforts of material scientists, chemical engineers, physicists, and electrical engineers. A specialist in this area comments:

Diamond has a long way to go, First, we need to develop a low-cost large area single crystal growth technology that produces wafers with a low density of crystal defects. Then, there are manufacturing challenges to overcome since it is the hardest material. For the near future, we are blessed with SiC and GaN – we will have a huge

number of challenges to solve in these two critical material technologies [2].

Table 2 shows the detailed comparison of Si, 4H-SiC, and GaN enhancement-mode power FETs [2]. The SiC MOSFET is a fast-switching voltage-controlled majority carrier unipolar device like Si MOSFET with a double-diffused metal-oxide-semiconductor (DMOS) structure. With the smaller thickness of the N-drift layer (because of the high breakdown field) and higher conductivity of the N-channel, the device has a higher voltage capability and smaller conduction drop than Si MOSFET. These properties also contribute to a higher

IMMIGRATION TO THE UNITED STATES AND START OF RESEARCH CAREER IN RENSSELAER POLYTECHNIC INSTITUTE IN MODERN POWER ELECTRONICS (1971–1976)

Going to the United States and settling in a reputed university was my lifelong ambition. I decided to emigrate to the United States in 1971 and start a new career at Rensselaer Polytechnic Institute (RPI). I joined RPI as an associate professor in electrical engineering with teaching and research in modern power electronics. The GE-CRD in Schenectady initiated this program, but getting an offer from RPI while working in India was not easy for me. The approval letter from RPI helped me to get my emigration visa (green card) quickly. RPI was a reputable private university, and the students were very brilliant. After joining RPI, I got a part-time offer from GE-CRD with a project on a thyristor high-frequency link cycloconverter. During my RPI career for five years, I did a number of innovative projects that included: developing a transistor ac switch for matrix converters; TRIAC speed control of induction motors; the series/parallel operation of TRIACs in converters; three-phase ac power control with transistors; a thyristor-saturable core self-oscillating Royer inverter; the phase-locked-loop speed control of dc motors; and so on. My GE project was extremely complex, but fortunately, it turned out to be very successful. As a reward, GE-CRD offered me a job in 1976, which I could not refuse.

TABLE 1 – A COMPARISON OF THE PROPERTY OF Si WITH WBG (4H-SiC, GaN), AND UWBG (Ga_2O_3 AND DIAMOND) MATERIALS.

	Si	4H-SiC	GaN	Ga_2O_3	DIAMOND
Bandgap E_g (eV)	1.12	3.26	3.4	4.8	5.5
Breakdown field E_B (V/cm) $\times 10^6$	0.3	3	3.5	8	10
Electron mobility μ_n (cm^2/Vs)	1,420	1,000	2,000	400	2,200
Hole mobility μ_p (cm^2/Vs)	600	100	200	100	850
Electron saturation velocity (10^6 cm/s)	10	22	25	–	–
Thermal conductivity (W/cm $^\circ\text{C}$)	1.5	4	1.3	0.2	22
Saturation drift velocity versus (10^6 cm/s)	10	20	25	–	–
Relative dielectric constant E_s	11.8	9.7	9.5	–	–

switching frequency. It has a reverse-conducting body diode, but the recovery current and recovery time are low due to the short minority carrier lifetime (like the SiC Schottky barrier diode) and are mainly contributed by the discharge of tiny junction capacitances. For this reason, the bypass diode is often omitted. The saturation resistance ($R_{DS(ON)}$) is reduced by a higher V_{GS} but increases at a higher junction temperature (positive temperature coefficient). Trench technology (developed by Infineon) can reduce conduction channel resistance of all the devices to improve the conduction efficiency (called CoolMOS, CoolSiC or CoolGaN).

Because of WBG, the device can operate at a higher T_j (up to 200 °C).

The cooling system design is less expensive due to the low thermal resistance of the device. A normally on junction field-effect transistor (JFET) with the cascoded connection of low-voltage Si MOSFET is also available, but normally off enhancement-mode devices are preferred. Currently, up to 1,700-V/350-A devices are available commercially. Higher-voltage SiC MOSFETS will have an excessive conduction loss. For this reason, higher-voltage, higher-power devices are bipolar (such as IGBT, GTO, thyristors, and so on) with conductivity modulation. For example, 15-kV SiC IGBTs have been undergoing laboratory testing for some time.

The GaN HEMT is a field-effect planar device with lateral current flow and

extremely high efficiency. The active region of the transistor is fabricated using GaN and AlGaN semiconductor materials on top of a Si substrate. The transition layers are grown for the differences of the thermal expansion coefficients between Si and GaN. The conduction path between the drain and source contacts is called *Two-dimensional (2D) electron gas (2DEG)* and is formed at the heterojunction between the GaN and AlGaN layers. The 2DEG is enhanced or depleted by the potential difference between the gate and the 2DEG below it. The HEMT conducts in the reverse direction without any body diode, i.e., there is no recovery loss. Matrix converter ac switches can be easily built with an inverse-series connection of GaN HEMT (no bypass diodes are needed) devices with a very low conduction loss.

In summary, the GaN lateral transistor structure permits very low gate and output storage charges, and correspondingly, very high switching frequency. Normally on GaN JFET is also available in a cascode structure with a series connection of Si MOSFET. Currently, 650-V/60-A, 1,200-V/30-A enhancement FETs are available, but higher-voltage, higher-power vertical devices are under development and will be available in the future. Again, in the future, bipolar devices with a higher power rating will also be available. An example application of an 80 kW GaN-based converter [3] for utility energy storage application indicates 95% less loss compared to Si IGBT and 85% less loss compared to SiC MOSFET.

In summary, the following general comments can be made for power semiconductor evolution:

- The gradual obsolescence of phase-control devices (thyristors and TRIACs) will occur.
- Si-based BJTs and GTOs are already obsolete.
- Insulated-gate self-controlled devices (power MOSFET, IGBT, and so on) will dominate.
- Si power MOSFET will remain universal for low-voltage, low-power, and high-frequency applications [switch-mode power supplies (SMPS), brushless DC motor (BLDM), and so on].

TABLE 2 – AN APPROXIMATE COMPARISON OF Si, 4H-SiC, AND GaN ENHANCEMENT-MODE POWER FETs.

	Si	4H-SiC	GaN
Voltage and current Ratings (Selected device for comparison)	30 V/15 A* (dc)	650 V/110 A [†] (dc)	600 V/31 A [‡] (dc)
Present power capability	1.2 kV/50 A	1,700 V/350 A	650 V/60 A, 1,200 V/30 A
Voltage blocking	Asymmetric	Asymmetric	Asymmetric
Gating	MOSFET	MOSFET	FET
Junction temperature range °C	-55 to +175	-55 to 200	-55 to 150
Safe operating area	Square	Square	Square
Static on-resistance $R_{DS(on)}$ mΩ 25°C	8.6	18	70
Switching frequency range, typical dV/dt (V/ns)	Up to 1 MHz –	Up to 1 MHz 50	Up to 3 MHz 200
Turn-on time (ns)	55	64	18
Turn-off time (ns)	35	74	29
Antiparallel diode	Yes	Yes	No
Reverse recovery loss	Yes	Yes	No
Snubber	Yes or no	Yes or no	Yes or no
Protection	Gate control	Gate control	Gate control
Thermal resistance R_{thjc} °C/W	5.9	0.31	1
Applications	Low voltage low power, SMPS, BLDM	VSC for EV, battery charger, PV, wind, and so forth	VSC for EV, PV, wind, and so forth
Comments	Very mature	Fast body diode	Ultrafast switching, no body diode, extremely low parasitic capacitances

*IR-6723M2DTR (dual package).

[†]ST-SCTW90N65G2V.

[‡]Infineon IG060R070D1 CoolGaN.

BLDM: brushless DC motor; VSC: voltage source converter.

- High-power IGBTs are being replaced by IGBTs.
- WBG power devices will be accepted universally in high- and medium-power converters.
- SiC devices will dominate for high-voltage, high-power applications.
- GaN devices are currently for medium power and will grow for high power in the future.
- SiC and GaN devices essentially constitute the near-term technology.
- The advent of UWBG power semiconductor devices in the next generation, such as Ga₂O₃, BN, and diamond, will provide significantly more improvement in voltage, frequency, and efficiency ratings of the devices, with the corresponding size miniaturization and higher-temperature operation in power electronics. Of course, the higher-temperature operation of passive circuit components, device packaging, and control system components is mandatory in high-temperature power electronics.
- Diamond appears to be the ultimate material that may be explored far in the next generation.
- Intelligent, integrated power modules will be increasingly available.
- The general trend is the integration of converter, control, and protection.

Power Converters

Power electronic converters are generally constituted by a matrix of controllable power semiconductor switches that perform conversion as well as control of electrical power. Traditionally, most of the phase-controlled line and load-commutated converters (LCCs) (including cyclo-converters), which have been commonly used in recent years, evolved in the classical power electronics era. Their popular applications today include thyristor-based high-power, multimewatt (multi-MW) HVdc converters in transmission systems and LCC synchronous motor drives. The advantages of this class of converters are simple topology, very high efficiency, and simplicity of control. The high efficiency is

essentially contributed by zero-current soft-switching.

However, the disadvantages are poor line displacement power factor, poor line power quality, sluggish control response, and the possibility of commutation failure due to line transients. The commutation failure in the LCC-wound field synchronous motor (WFSM) drive generates a large and dangerous pulsating torque in the machine. [Personal note: I was once involved in consulting on the failure of a 12-MW dual-cyclo-converter (CCV)-WFSM

gold-ore-grinding mill drive system in the West Australia (Kalgoorlie) grid [4], where commutation failure generated a large oscillatory pulsating torque, causing extensive damage to the machine and the ring gear system.] The har-

monic standards of the IEEE (IEEE-519) and Europe (IEC-61000) were formulated to limit the line harmonics. The line harmonics in the phase-controlled converter (PCC) can be attenuated by multipulsing with a phase-shifting transformer or by using active filters.

The line-lagging DPF problem can be solved by a static synchronous compensator (STATCOM) [static VAR compensator (SVC)]. The advent of a self-controlled thyristor inverter by forced commutation initiated the evolution of dc-link voltage-source converters for general industrial applications, including ac motor drives. William McMurray of GE was the pioneer in this area. Gradually, the advent of self-controlled devices replaced thyristor converters. Different types of PWM techniques,

ANY NEW INVENTION MAY ALTER THE COURSE OF A TECHNOLOGY, WHICH HAS HAPPENED MANY TIMES IN THE HISTORY OF POWER ELECTRONICS.

MY 11 YEARS IN THE GE CORPORATE RESEARCH LAB (1976–1987)

Having spent 16 years of my career in the university setting so far, I always felt that I had a big gap in my expertise in power electronics. I had rarely worked with large power electronic converters for industrial applications. In those days, GE-CRD was the world's top research center in power electronics. It was like Bell Lab, where transistor was invented. The images of Thomas Edison and Charles Steinmetz were everywhere in GE. My office was located in the historic Building 37, where E.F.W. Alexanderson, Gabriel Kron, Philip Alger, and so on had their laboratories. GE-CRD was then considered as the ivory tower of power electronics worldwide, and power electronic professionals from all over the world used to visit us in Schenectady. It was a thrilling experience to meet so many world-renowned scientists across the hall. After joining GE, I started working with Bill McMurray on a current-fed Auto Sequential Commutated inverter and was located in the same office. McMurray was the founding father and guru of power electronics, and the whole power electronic world bowed to him with deep respect.

During most of my time in GE, I was involved with EV and HEV projects. The EV/HEV development was the first major initiative by the U.S. Government Department of Energy after the Arab oil embargo in the 1970s. I was the principal engineer for microprocessor/digital signal processor (DSP) control development. Our first EV (ETV1) with a power transistor chopper and dc motor drive was very successful and was demonstrated before Queen Elizabeth II of England. Our last EV project (ETX II) [2] was based on an interior permanent-magnet synchronous motor (IPM-SM) drive. Gradually, IPM motor-based EV drives were accepted all over the world. My other GE projects were: a linear inductor motor drive for railroad propulsion; a switched reluctance motor (SRM) drive; a residential PV system maximum power point tracker control; a sliding mode control of induction motors (IMs); a scalar decoupled control of IM; an adaptive hysteresis-band control of an IPM motor; and so on. I published my first textbook on power electronics and ac drives in 1986 with Prentice Hall while working in GE.

such as sinusoidal pulse width modulation, HB, selected harmonic elimination, and space vector modulation, also arrived to control output voltage with improved harmonic quality. The PWM active rectifier was introduced to control line harmonics as well as DPF.

Dual PWM converters, particularly with the neutral-point-clamped type, ousted the CCVs. Active filters, which are used with PCC and diode rectifiers, are getting obsolete. The SVCs/STATCOMs are extremely important elements in power grids for lagging DPF correction of load and line P and Q control with the help of FACTS. FACTS is an extremely important element in the future smart grid. Converter soft-switching, although extremely popular in high-frequency link SMPS, is not useful in general high-power electronics, including motor drives. The matrix converter (MC), introduced in 1980s, has an attractive topology, but its future is slim in the author's opinion in spite of the advent of simple- and high-efficiency GaN ac switches. MCs have been on and off many times in the industry. The modern modular multilevel converters (MMCs), particularly using cascaded H-bridge and half-bridge (defined as MMC) topologies, are very important. The MMCs, particularly with SiC and other WBG devices, are important in utility systems with 50/60-Hz applications in HVdc converters, STATCOMs, FACTS, and variable frequency motor drives where the current is low at lower frequencies.

In summary, the following general comments can be made for converter technology evolution:

- Phase-controlled thyristor-based converters will be totally obsolete in the future.
- Voltage source converters are becoming universal.
- Soft-switched voltage source converters, particularly for motor drives and other high-power applications, show no future promise.
- Matrix converters, in spite of rich literature, are not expected for drives and other applications.

- Z-source converter [Z-source inverter (ZSI) or quasi-ZSI] applications in industry are questionable.
- Model predictive control (MPC) applications in industry are questionable.
- Multilevel MMC type (with half-bridges) converters, particularly with SiC (and other WBG and future UWBG devices), have tremendous promise for high-power applications. A lot of research is yet needed for motor drive applications.
- There is an increasing trend of real-time simulation of power electronic systems with hardware-in-loop testing.
- Converter technology, in general, is approaching saturation. Future emphasis will be on integration, automated design, and advanced control by digital signal processors (DSPs) and field-programmable gate arrays.

Motor Drives

The area of motor drives is closely associated with the evolution of power electronics, i.e., the evolution of devices, converters, control, estimation, modeling, simulation, and hardware/software implementation tools. Historically, between the two classes of dc and ac drives, the ac drives, particularly the cage-type induction motors, were traditionally used in constant-speed application, whereas dc drives were used in variable-speed applications. Although dc drives are still widely used in industry, they are tending to become obsolete because of their characteristic disadvantages. The technology advancement of power electronics has permitted variable-frequency, variable-speed ac drives (both induction and synchronous), and they are now used extensively in industry. The cage-type induction motor drives are very common because the machines are cheaper, more rugged, and reliable.

However, the efficiency of permanent-magnet SM (PMSM) drives with a high-energy neodymium-iron-boron magnet is higher, although the machine is more expensive, which makes

the lifecycle cost lower. In high-power drives, WFSM drives will remain popular because of the performance advantages with field control. Thyristor-based load-commutated inverter drives are being increasingly replaced by two-sided multilevel converters. In this context, it can be easily predicted that switched reluctance motor (SRM) drives will be totally obsolete for industrial applications. It is unfortunate that so much effort has been wasted on SRM drive technology over such a long time. The vector or field-oriented control with sensorless estimation will be increasingly popular as the industry standard with the obsolescence of scalar control.

The direct torque control has considerably improved in recent years with the drive performance approaching that of vector control. The complex MPC control is also viable with many sophistications demonstrated in the recent literature but has not yet demonstrated its performance superiority over vector control. Artificial intelligence (AI) techniques will be increasingly used, particularly for fault diagnostics and fault-tolerant control. The literature on motor drives is now diminishing, with current trends in favor of smart grid and RESs. Power electronics is now considered as an integral part of power engineering.

The future trends of motor drives can be summarized as:

- Voltage source converter drives will be universal in the future.
- Cage-type induction motor drives will remain common for general industrial applications.
- Slip power recovery drives will be increasingly obsolete in the future.
- IPM-SM drives will be increasingly popular, particularly for extended speed EV-type applications.
- For high-power applications, WFSM drives will be used with voltage source inverter-multilevel converters (particularly MMCs).
- Vector control will be universal.
- AI techniques, particularly neural networks, will be used for drive performance improvement, fault diagnostics, and fault-tolerant control.

- There is a trend toward drive system integration, particularly in the lower end of power.
- There is a diminishing trend in drive research with technology saturation.

Power Electronics in Smart Grids and RESs

Power electronics is an indispensable ingredient in modern smart grids and RESs. What is a smart grid? A smart or intelligent power grid is basically a vision of an advanced power grid of tomorrow using state-of-the-art technologies in power systems, power electronics, control systems, computers, communications, information, AI, cyber, and so on that will improve system availability, reliability, power quality, energy efficiency, and security with optimum resource utilization and economical electricity to the consumers. A micro- or minigrid is basically a local power system (ac, dc, or hybrid) that can operate autonomously or be interconnected to a grid. Considering these features, our present power grids have major deficiencies. A smart grid normally integrates large fossil and nuclear power stations, RESs, HVdc systems for economical and efficient long-distance power transmission, FACTS for unified active (P) and reactive power (Q) control in transmission circuits, and STATCOMS for VAR compensation, where most of these elements are heavily based on power electronics.

The segmentation of a large power grid by HVdc and FACTS links improves the stability management of the system. An important function of the smart grid is the supply-demand interactive energy management. If the energy demand curves (always fluctuating) are forced to follow the available generation curve, the energy storage requirement becomes minimal, and the corresponding tariff rate becomes economical. The control and protection of a large smart grid are extremely complex. The power generation has to be scheduled among the different generating units for the demand load curves, and the power flow routing has to be controlled for

optimizing system efficiency and reliability, preventing the overloading of any element of the power system. Because of the complexity of operation, the system requires dynamic modeling and real-time simulation based on supercomputers.

Currently, RESs (mainly hydro, wind, and solar) are getting tremendous emphasis all over the world, where the wind and solar PV are mainly based on power electronics. Other renewable sources, such as tidal, wave, geothermal, and biomass, will be explored systematically in the future. The main reasons are that renewables are economical, environmentally clean (green), and distributed all over the world and that they do not have the characteristic disadvantages of nuclear power. Our ultimate goal is to have 100% RESs in the future. Is that possible? It is an ambitious goal with formidable challenges. Currently, about 11% of the global energy in electrical form comes from RESs [2], which is subdivided as follows: hydro, 17%; wind, 15%; solar, 7%; and geothermal, 1%, with the remaining 60% coming from biomass.

The energy potential from wind and solar is tremendous. According to a Stanford University estimate [2], exploring only about 20% of the available wind energy (the European

Wind Energy Association estimate is 10%) can meet all the energy needs of the world. Solar energy is distributed all over the world. The solar PV cell cost is decreasing drastically in recent years. Wind or solar, or both together, can easily meet the 100% renewable goal of the world, but the main challenges are:

- 1) The sources are sporadic in nature, and individual plant capacity is small compared to fossil/nuclear plants. Bulk energy storage (usually by battery) (power electronics-based) requirements makes it expensive. Large offshore wind farms (now being emphasized in the United States) are expensive but have a consistent availability, requiring reduced storage need. Regulating frequency and voltage in a large grid with small power plants is not easy. Extensive research is needed on batteries for economical bulk energy storage.
- 2) The availability of renewable energy is regional, and it is not difficult to make 100% RESs regionally (for example, Iceland has nearly 100% RESs). However, for 100% RESs in the world, energy has to be transmitted economically and efficiently to the remote corners of the world, which may be challenging. The concept technology of the

MY EXPERIENCE IN THE UNIVERSITY OF TENNESSEE (1987–2021)

I decided to return to my university career in 1987 after spending 11 years at GE-CRD. In parallel, I also started working as the chief scientist of the newly established Power Electronics Applications Center at the Electric Power Research Institute. At the end of my GE career, I gained tremendous visibility in the world. Fortunately, a large number of brilliant visiting professors and graduate students from abroad came to work with me with the financial support of their respective governments. Many of them wanted to work in the emerging AI applications in power electronics.

At the University of Tennessee (UT), my projects included: soft-switched power conversion for ac motor drives; high-frequency nonresonant link power conversion; fuzzy-controlled wind generation systems with efficiency optimization control; converter fault studies; neural network-based control and estimation of converters; high temperature superconductivity-synchronous motor ship propulsion with a multilevel converter; and so on. In this period (1987–2021), I wrote/edited six more books in power electronics, and I guest edited two special issues of *Proceedings of the IEEE*. I got the opportunity to travel abroad extensively to give tutorials, invited seminars, and keynote addresses. The UT provided me office facilities after my official retirement in 2003.

global supergrid is already there, but the cost will be prohibitive.

- 3) Global energy consumption is increasing steadily with the rise of population and the urge of a higher living standard. The global grid has to accommodate this increasing energy demand for the future.
- 4) The proposed smart grid technology has to be extended to the global grid so that energy is economical for consumers with optimum resource utilization, reliability, power quality, energy efficiency, and security.
- 5) Linking countries across the world in such a global grid may be a formidable political problem. Despite these challenges, it is the author's belief that the world will eventually see 100% RESs, ousting the well-established fossil and nuclear power plants.

Power Electronics in Solving Climate Change Problems

Power electronics plays a significant role in solving climate change problems [5]–[7], which are such serious concerns in our society. Climate change or global warming is primarily caused by burning fossil fuels (coal, oil, and gas), which generate greenhouse gases (mainly CO₂) and trap the solar heat that raises the atmospheric temperature gradually. The harmful effects of climate change are the melting of polar ice caps and glaciers around the world; severe droughts in tropical countries near the equator; more hurricanes, tornados, rains, and floods; the increased acidity and temperature rise of sea water; the deterioration of fresh water supply; and the spread of tropical diseases. Solving climate change problems remains a challenge in the 21st century. The obvious solution of the problem is the economization of our energy consumption and replacing fossil fuel-based energy generation by RES-based clean energy (which uses power electronics). One way to promote the RESs is the implementation of a carbon tax.

Energy saving is one of the important goals in power electronics

applications. Around 65% of generated energy in the United States is consumed in motor drives, and 20% is used in lighting. About 75% of the drives are used in pumps, fans, and compressor-type drives. In this class, variable-frequency drives can save nearly 30% of energy. If LED lamps are used instead of fluorescent or traditional incandescent lamps, a substantial amount of energy can be saved. Similarly, variable-speed air-conditioners/heat pumps can save up to 30% energy. Of course, energy needs for home or space heating can be substantially reduced by proper insulation. Electric transportation, including EVs, can eliminate pollution if the energy is generated by RESs.

Considering the present trend, it can be predicted that the world will eventually have 100% EVs, eliminating internal combustion engine vehicles and HEVs. Promoting mass electric transportation, such as in Japan, can save a lot of energy. A considerable amount of energy can be saved by improving the efficiency of generation, transmission, and utilization by using the smart grid technology in the future. Unfortunately, a significant amount of energy is wasted in affluent countries like the United States because it is cheap and because of bad consumer habits. In the author's opinion, widespread energy efficiency improvement by power electronics and other methods with existing technologies can save around 20% of global energy consumption, and another 15% can be saved by the rigorous control of energy waste [2], [6]. Finally, global climate change problems are solvable by the united effort of the humanity.

Summary

In this article, the author has attempted to give a brief but comprehensive review of power electronics technology and his vision of the future progression, based on his knowledge and experience. The power semiconductor devices, converter circuits, and motor drive areas have been covered in the article. Some emphasis has been given on devices because of their

importance and dynamic advances in power electronics evolution. A brief review of classical power electronics was included in the beginning for completeness. The areas of smart grids and RESs along with climate change problems have also been included because of the impact of power electronics in these important areas. Since the author has pursued the power electronics field very aggressively during his long career in both the industrial and academic environments, a brief sketch of his career experience has been inserted within the contents of the article. Finally, in conclusion, I would like to mention that the dedicated and relentless contributions of so many scientists and engineers have made the power electronics technology so rich today.

—Bimal K. Bose

*Emeritus chair professor,
Department of Electrical
Engineering and Computer
Science, University of
Tennessee, Knoxville,
Tennessee, 37996, USA*

References

- [1] B. Brusso and B. K. Bose, "Power electronics—Historical perspective and my experience [History]," *IEEE Ind. Appl. Mag.*, vol. 20, no. 2, pp. 7–14, Apr. 2014, doi: 10.1109/MIAS.2013.2292645.
- [2] B. K. Bose, *Power Electronics and Motor Drives—Advances and Trends*, 2nd ed. New York, NY, USA, Academic, 2021.
- [3] M. Ferdowsi, P. Shamsi, and B. Baddipadiga, "Gallium Nitride (GaN) based high frequency inverter for energy storage applications," InnoCIT, St. James, MO, USA, 2017. [Online]. Available: https://eesat.sandia.gov/wp-content/uploads/2017/12/Mehdi_Ferdowsi.pdf
- [4] B. K. Bose, "Ore-grinding cycloconverter drive operation and fault," *IEEE Ind. Electron. Mag.*, vol. 5, no. 4, pp. 12–22, Dec. 2011, doi: 10.1109/MIE.2011.943022.
- [5] B. K. Bose, "Energy, environment, and advances in power electronics," *IEEE Trans. Power Electron.*, vol. 15, no. 4, pp. 688–701, Jul. 2000, doi: 10.1109/63.849039.
- [6] B. K. Bose, "Global warming," *IEEE Ind. Electron. Mag.*, vol. 4, no. 1, pp. 6–17, Mar. 2010, doi: 10.1109/MIE.2010.935860.
- [7] B. K. Bose, "Global energy scenario and impact of power electronics in 21st century," *IEEE Trans. Ind. Electron.*, vol. 60, no. 7, pp. 2638–2651, Jul. 2013, doi: 10.1109/TIE.2012.2203771.
- [8] L. G. Franquelo, I. Nagy, and C. Wen, "Honoring Dr. Bimal K. Bose [Tributes]," *IEEE Ind. Electron. Mag.*, vol. 3, no. 2, pp. 12–14, Jun. 2009, doi: 10.1109/MIE.2009.932710.
- [9] B. K. Bose, "My Life in Power Electronics," [Online]. Available: http://ethw.org/First-Hand:My_Life_in_Power_Electronics





Before Lithium-Ion Batteries: The Age of Primary Cells

Climate change mitigation and the transition toward decarbonized energy sources, widely discussed during international events such as the 2021 United Nations Climate Change Conference, in Glasgow [1], will entail a large expansion of energy storage for mobile and stationary applications, i.e., in electric vehicles (road, waterborne, air, and so on) and advanced grids (smart grids, microgrids, and similar types), respectively, in conjunction with renewable sources. Among energy storage systems, batteries are expected to play a major role, and a huge amount of funding is being allocated for their development in several countries, including the construction of gigafactories for their production. The coming years promise to become the golden age of batteries. However, if we look into the past, we realize that another golden age occurred two centuries ago.

Electrochemical effects were known in ancient times. In 1938, archeologist Wilhelm König brought forward the idea that an artifact, now known as the *Bagdad battery*, was an electrochemical cell, dating sometime between 150 BCE and 650 CE [2], but the real use of this object is questioned. Electrochemical processes were certainly known to ancient civilizations, as proved by the gold plating on the equestrian statue of the Roman emperor Marcus Aurelius (late second century CE) and the plated artifacts of the Chimú culture, which flourished in South America around the 10th century CE.

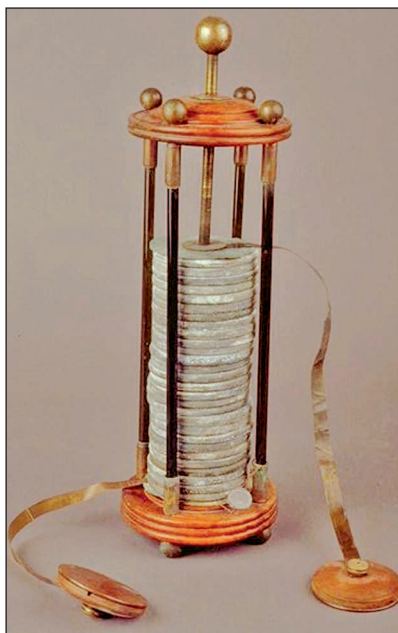


FIGURE 1 – Volta's first electrochemical battery, 1799. The anode was zinc, the cathode was copper, and the electrolyte was brine (Source: Museo della Tecnica Elettrica, Pavia, Italy; used with permission.)

In the Western world, electricity attracted growing interest from researchers after the Scientific Revolution, and several discoveries were made during the Enlightenment, but they were counterbalanced by a very small number of practical inventions [3]. The most notable of these was the lightning rod by Benjamin Franklin (in 1752), which provided protection against lightning-ignited fires in wooden buildings. At that time, electricity was relegated to electrostatic experiments performed inside laboratories, to the amazement of gaping audiences.

At the end of the 18th century, the Italian scientist Alessandro Volta (Italy,

1745–1827) was famous all across Europe and who had already received the Copley Medal from the Royal Society of London, in 1794 (considered the British forerunner of the Nobel Prize). In the framework of a dispute over animal electricity, he invented the electrochemical pile, in 1799, and communicated it to the Royal Society on 20 March 1800. In its first arrangement, the device consisted of a pile of cells, each made of two disk electrodes, one of zinc and one of copper, interposed by brine-soaked cardboard acting as an electrolyte (Figure 1) [4]. The cell soon attracted great enthusiasm throughout Europe despite the divisions of the Napoleonic Wars, and it earned Volta a harvest of honors [5]. The reason for such success was that the voltaic battery enabled the generation of a galvanic current, i.e., a flow of electricity that could persist for dozens of minutes. In our opinion, this seems a short duration, but, at that time, it was revolutionary because it was thousands (if not millions) times longer than the discharge available from electrostatic devices.

Several scientists soon replicated the voltaic battery in larger and larger sizes and used it in previously unimaginable experiments [6]. As early as 1800, the English scientists William Nicholson (1753–1815) and Antony Carlisle (1768–1840) produced water electrolysis; i.e., they decomposed water into hydrogen and oxygen, thus showing that electricity can trigger massive chemical reactions. In 1801, Humphrey Davy (United Kingdom,

1778–1829), later the recipient of the Copley Medal, in 1805, experienced the glow of a platinum strip passed by an electric current, an effect that was exploited in the incandescent bulbs developed into practical models almost 80 years later [7].

In 1802, Luigi Valentino Brugnatelli (Italy, 1761–1818), a colleague and friend to Volta at Pavia University, pioneered the electroplating of metals and non-metals (galvanoplastics). In 1802, Gian Domenico Romagnosi (Italy, 1761–1835) observed the interaction between electric and magnetic phenomena, but he did not communicate it properly, and this electromagnetic effect was rediscovered and disseminated 17 years later by Hans Christian Ørsted (Denmark, 1777–1851). Vasilij Vladimirovič Petrov (Russia, 1761–1834) produced the first electric arc between two coal electrodes, using a 4,200-cell battery (the largest built at that time), but the results were published in Russian, in 1803, and passed unnoticed in Western countries. In 1804, Francisco Salvá Campello (Spain, 1751–1828) conceived the first electric telegraph powered by a voltaic stack that used one line of two iron wires for each letter of the alphabet, with an electrolytic cell receiver producing hydrogen bubbles when the circuit was switched on. In 1806–1810, Humphrey Davy used a 250-cell battery, then the largest in the United Kingdom, to decompose several alkaline compounds and thus isolated sodium, potassium, barium, calcium, strontium, and magnesium for the first time. In 1809, Davy performed the first public demonstration of a continuous electric arc by using a 2,000-cell battery, an effect exploited in arc lamps from about 1842, in arc welding from 1885, and in arc furnaces from 1879. This is a quick summary of the major developments resulting from the voltaic cell limited to the first decade of the century.

All this interest stimulated research into more advanced electrochemical cells. As early as 1802, Johann Wilhelm Ritter (Germany, 1776–1810) built the first crude dry electrochemical battery. William Cruickshank (United Kingdom, 1745–1810) introduced the

trough battery, made of flat electrodes placed vertically and parallel in an electrolytic solution, which was a design suitable for industrial production. He also carried out very early experiments on rechargeable cells. In 1812, an improved version of a dry battery was proposed by Giuseppe Zamboni (Italy, 1776–1846) using an electrolyte consisting of a glue material placed between copper and zinc disks. However, rechargeable and dry cells remained laboratory curiosities for some eight decades.

In 1829, Antoine-César Becquerel (France, 1788–1878) identified the generation of hydrogen bubbles at a cathode (polarization) as a major cause of voltage drops in voltaic cells and built the first prototype of a depolarizing cell by using two electrolytes instead of one. This concept was recovered by John Frederick Daniell (United Kingdom, 1790–1845), in 1836. His depolarized cell had zinc and copper electrodes immersed in sulfuric acid and copper sulfate solutions, respectively, separated by a porous earthenware barrier (Figure 2). Becquerel and Daniell shared the Copley Medal of 1837 for their work on electricity.

The Daniell cell was a significant advance from the voltaic cell, providing a longer and more reliable current. It underwent several improvements, notably the version by J. Fuller, in

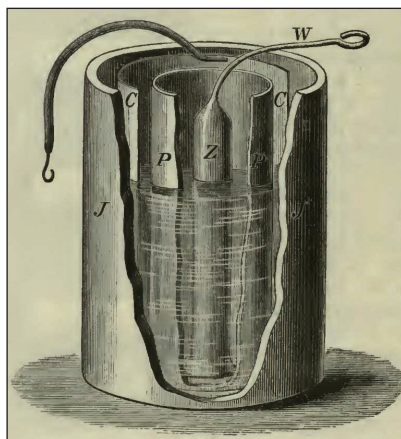


FIGURE 2 – The Daniell cell (primary, wet, and depolarized), depicted in an illustration dated 1904. The anode was zinc, the cathode was copper, and the electrolytes were sulfuric acid and copper sulfate solutions. (Source: Wikimedia Commons.)

1853 (sulfuric acid was replaced with zinc sulfate) and the gravity cell by a Frenchman named Callaud who, in the 1860s, placed the electrodes one above the other and removed the porous barrier, relying on the density gradient to keep the electrolytes separate and thus lowering the internal resistance and increasing the current. The Daniell cell was the first practical source of energy for electrical devices and powered early telegraph networks, particularly in the United States and the United Kingdom (where it was still being used as late as the 1950s). Producing a particularly stable 1.09-V electromotive force (EMF), it was also used as the voltage standard for a long time, up to the advent of the Weston cell, in 1893.

Volta passed away in 1827 and could witness only the beginning of the new science to which he had opened the door. He had thought the dynamic electricity of his cell originated from the physical properties of the electrodes, failing to identify its real source. It was Michael Faraday (United Kingdom, 1791–1867) who recognized that the electricity in the cell resulted from the chemical reactions occurring at the two electrode/electrolyte interfaces. He defined the equivalence between chemical quantity and electricity in his laws of electrolysis, in 1833. Faraday received the Copley Medals of 1833 and 1838 for his research on electricity.

In 1839, William Robert Grove (United Kingdom, 1811–1896) built the depolarized zinc–platinum cell, with sulfuric acid and nitric acid as electrolytes (Figure 3). Producing a voltage of 1.8 V (almost twice that of the Daniell cell), it was successfully used in telegraphy, particularly in the United States (up to the 1860s) despite the emission of poisonous nitric oxide fumes and the cost of platinum. In 1841, Robert Wilhelm Eberhard Bunsen (Germany, 1811–1899) replaced the platinum electrode of the Grove cell with carbon to obtain the zinc–carbon cell, with an EMF of 1.9 V (Figure 4). The Bunsen cell was cheap enough to be used in electroplating, i.e., the first industrial electrical process, developed

in Russia by Moritz Hermann Jacobi (Germany, 1801–1874), in 1838, and in the powering of arc lamps, practical versions of which were available from the following year. Among other honors, Bunsen received the Copley Medal in 1860.

All the batteries described previously, derived from the voltaic cell, were wet primary cells; i.e., they had liquid electrolytes and were not rechargeable. When the electrodes were consumed, the cells were renewed by replacing the electrodes and the electrolyte. As a consequence, even in their advanced versions, primary cells were not economical. However, up until the 1860s, they were the only widely available sources of steady-state electric power. In fact, very few magneto-electric generators had found practical applications, such as the Woolrich electrical generator (used to power the Elkington Silver Electroplating Works, in 1844), the small Siemens generators developed since 1956, and the Holmes generator (used to power the South Foreland Lighthouse, in 1859 [8], [9]).

Thus, primary batteries were the source of electricity, which, in the first six decades of the 19th century, enabled a novel research field to open and impressive results to be achieved on both the experimental and theoretical sides. They supported a new branch of science that, in the following decades, dramatically changed our view of the world. Just to mention a few, the discoveries by Georg Ohm, Gustav Kirchhoff, Michael Faraday, and Joseph Henry stemmed from primary cells [10]. In addition, primary batteries facilitated the development of electric devices and systems, such as the telegraph and telephone, which brought a revolution in telecommunications [11], [12]. They powered early electrical industrial processes, such as electroplating and electric lighting based on arc lamps, e.g., in the Opéra de Paris theater, in 1846. In other words, the voltaic battery gave birth to electrical engineering and triggered the sensational technological developments in information, energy, industry, and common

applications that characterize the world today.

In 1871, Zénobe Théophile Gramme (Belgium, 1826–1901) introduced the first reliable dc magneto-electric generator capable of delivering power levels comparable to those of steam machines with very steady dc currents. The electric energy from those dynamos was much cheaper than that from primary cells, and it made sense to use rechargeable cells, or secondary batteries, when there was such a source of inexpensive power. The lead–acid secondary battery was taken to practical production by Camille Alphonse Faure (France, 1840–1898), in 1881, building on the prototype conceived by Gaston Planté (France, 1834–1889), in 1859. Lead–acid secondary batteries were pivotal in the development of other technologies, notably electric cars in the following two decades [13], [14]. Electric grids powered by magneto-electric generators constituted other competitive technologies providing wide access to cheap electric energy [15], [16].

However, primary cells were not left behind. In 1867, Georges Leclanché

(France, 1839–1882) invented a cell with zinc and carbon electrodes, ammonium chloride as a liquid electrolyte, and manganese dioxide as a chemical depolarizer, with an EMF of 1.5 V (Figure 5). In 1881, Jules Alphonse Thiébaud (France) improved the cell by converting the zinc anode into a watertight container. However, wet cells remained

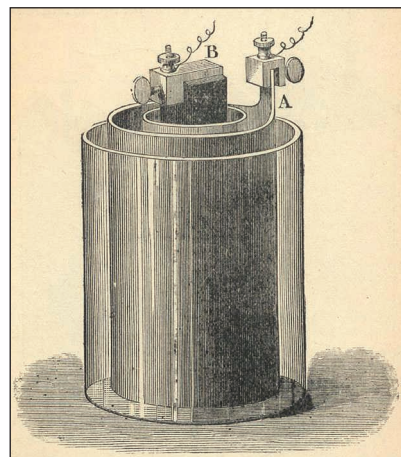


FIGURE 4 – The Bunsen cell, from *Electro-Plating and Electro-Refining of Metals*, by Arnold Philip, 1911. It was primary, wet, and depolarized. The anode was zinc, the cathode was carbon, and the electrolytes were sulfuric acid and nitric acid solutions. (Source: Wikipedia Commons.)

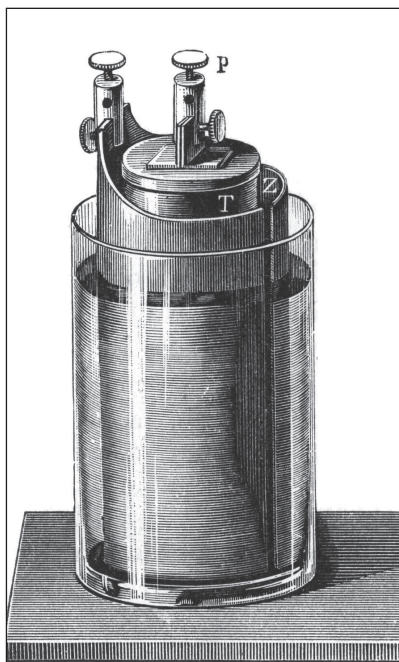


FIGURE 3 – A later model of the Grove cell, from 1897. It was primary, wet, and depolarized. The anode was zinc, the cathode was platinum, and the electrolytes were sulfuric acid and nitric acid solutions. (Source: Wikipedia Commons.)

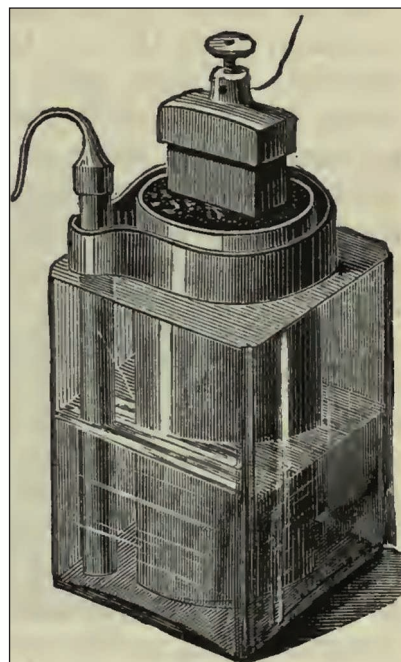


FIGURE 5 – The Leclanché cell (primary, wet, and depolarized). The anode was zinc, the cathode was carbon, the electrolyte was ammonium chloride, and the depolarizer was manganese dioxide. (Source: Wikimedia Commons.)

unsuitable for portable use. In 1886, Carl Gassner (Germany, 1855–1942) developed the Leclanché–Thiébaud cell into the dry carbon–zinc cell by immobilizing the liquid electrolyte with inert substances. As a result of these developments, the Leclanché cell obtained the gold medal at the 1889 Paris Universal Exposition, and various versions of it powered several types of small equipment in the following decades. In 1893, Edward Weston (United Kingdom/United States, 1850–1936) developed the eponymous wet primary cell that resorts to cadmium and mercury in the electrodes and cadmium sulfate in the electrolyte. It was not suitable for power uses, but it was employed for almost a century as a voltage standard, e.g., in the calibration of instruments, due to its very constant EMF of 1.0183 V.

The advent of the transistor and solid-state electronics, in 1947–1948, revolutionized the electronics market, with more compact, light, and energy-saving devices, e.g., wristwatches and portable radio receivers, which began appearing in 1952–1954 [17], and this demand promoted the development of better-performing primary batteries

[18]. Improved carbon–zinc cells produced in standard sizes (such as D, C, AA, AAA, E, and so on) dominated the market for a large part of the 20th century, with their performance improving 700% between 1920 and 1990. They are still widely used for low-power devices when cost is important, particularly in developing countries. The zinc anode was coated with mercury to protect it from corrosion, but, since mercury is environmentally hazardous, it was removed in the 1990s, and highly purified zinc is now used instead [19].

Early primary alkaline cells, which used an alkaline electrolyte instead of an acid one, were developed in France by Felix de Lalande and Georges Chaperon in 1882 to avoid corrosion issues and enabling a wider choice of materials for the battery components. The cell used zinc and copper oxide electrodes, with potassium hydroxide as an electrolyte. This battery powered the French submarine *Gymnote*, in 1888. In 1957, Karl Kordes (Austria, 1922–2011), Lewis Urry (Canada, 1927–2004), and Paul A. Marshall, of Union Carbide, built on the findings of the former and patented an alkaline dry

cell with zinc and manganese dioxide electrodes and potassium hydroxide as an alkaline electrolyte [20].

The alkaline cell was improved in the following years and overcame the dry zinc–carbon cell because of its higher energy and power densities, and, despite its higher cost, it was particularly convenient in supplying small portable devices that needed superior power levels (Figure 6). Also, in this case, a major step was the elimination of mercury from the zinc anode. The introduction of the plastic label jacket, in 1987, in place of the cardboard–steel one, supported an increase of 10–20% of the internal volume for active materials and thus of capacity. Alkaline cells remain very popular and have a vast number of uses.

Lithium is an advantageous metal for electrochemical cells, due to its high electrochemical potential and extremely low volumetric mass density, but the development of lithium cells was not straightforward [21]. Gilbert Newton Lewis (United States, 1875–1946) experimented with lithium as early as 1912, but viable primary lithium cells with a manganese dioxide cathode became available only in the 1970s, and they replaced other small cells based on chemistries such as zinc–mercury oxide, zinc–silver oxide, and zinc–manganese dioxide for niche uses, e.g., hearing aids. Primary lithium cells are still popular in some sizes, e.g., the button cells used in electronic controllers [22], [23].

Today, the performance of primary lithium cells has been largely overcome by rechargeable lithium-ion cells in their several chemistries. Their market share is expanding fast as they become more and more competitive [24]. Despite that, primary batteries still contribute 90% of the US\$50 billion global battery market. This huge demand results in about 15 billion primary batteries being disposed of every year, and their pollutant content makes them a wasteful and environmentally unfriendly technology, although the use of mercury, cadmium, and lead started to be prohibited in the 1990s. They present a very poor life cycle assessment because the manufacturing energy they consume

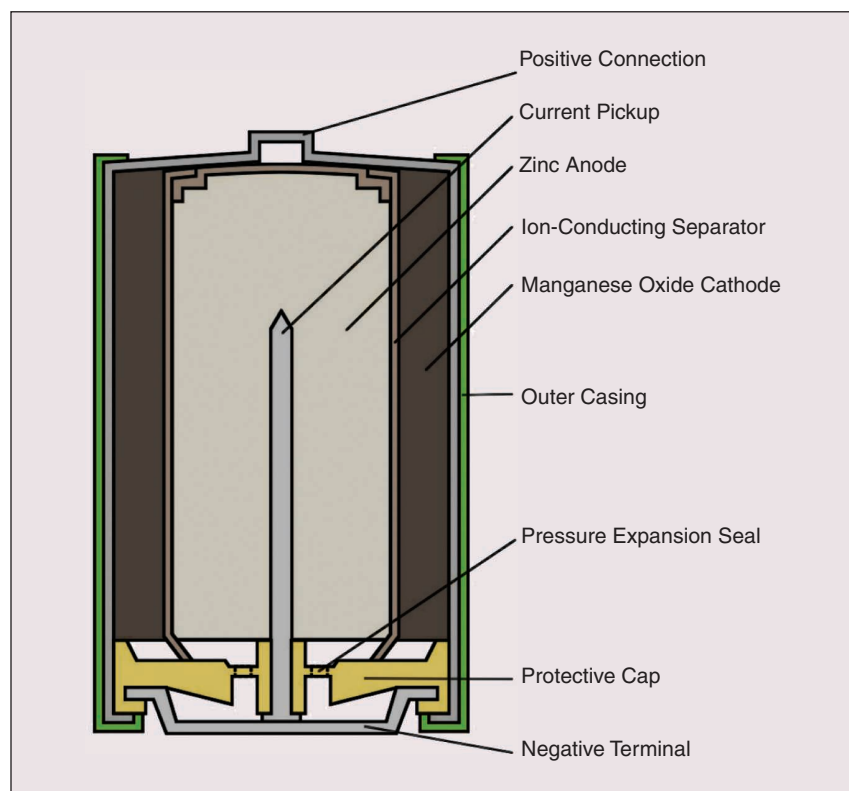


FIGURE 6 – The structure of a size D modern alkaline dry cell. (Source: Wikipedia Commons.)

is about 50 times the energy they contain [25]. Still, we are in debt to primary cells for the unique role they had in science and technology in the 19th and 20th centuries.

References

- [1] UN Climate Change Conference UK 2021, "CPO26 declaration on accelerating the transition to 100% zero emission cars and vans," UK Government, London, U.K.—United Nations Climate Change, New York, NY, USA, Nov. 10, 2021. Accessed: Feb. 22, 2022. [Online]. Available: <https://ukcop26.org/cop26-declaration-on-accelerating-the-transition-to-100-zero-emission-cars-and-vans/>
- [2] P. T. Keyser, "The purpose of the Parthian galvanic cells: A first-century A.D. electric battery used for analgesia," *J. Near East Stud.*, vol. 52, no. 2, pp. 81–98, Apr. 1993, doi: 10.1086/373610.
- [3] M. Guarnieri, "Electricity in the age of enlightenment," *IEEE Ind. Electron. Mag.*, vol. 8, no. 3, pp. 60–63, Sep. 2014, doi: 10.1109/MIE.2014.2335431.
- [4] M. Guarnieri, "The big jump from the legs of a frog," *IEEE Ind. Electron. Mag.*, vol. 8, no. 4, pp. 59–61, Dec. 2014, doi: 10.1109/MIE.2014.2361237.
- [5] F. Bevilacqua and E. Giannetto, Eds. *Volta and the History of Electricity*. Milan, Italy: Università degli Studi di Pavia – Hoepli, 2003.
- [6] M. Guarnieri, "The rise of light – Discovering its secrets," *Proc. IEEE*, vol. 104, no. 2, pp. 467–473, Feb. 2016, doi: 10.1109/JPROC.2015.2513118.
- [7] M. Guarnieri, "An historical survey on light technologies," *IEEE Access*, vol. 6, pp. 25,881–25,897, Jun. 2018, doi: 10.1109/ACCESS.2018.2834432.
- [8] M. Guarnieri, "Revolving and evolving – Early DC machines," *IEEE Ind. Electron. Mag.*, vol. 12, no. 3, pp. 38–43, Sep. 2018, doi: 10.1109/MIE.2018.2856546.
- [9] M. Guarnieri, "The development of AC rotary machines," *IEEE Ind. Electron. Mag.*, vol. 12, no. 4, pp. 28–32, Dec. 2018, doi: 10.1109/MIE.2018.2874375.
- [10] M. Guarnieri, "A glance at early circuit theory," *IEEE Ind. Electron. Mag.*, vol. 15, no. 1, pp. 79–83, Mar 2021, doi: 10.1109/MIE.2021.3051783.
- [11] M. Guarnieri, "Messaging before the Internet – Early electrical telegraphs," *IEEE Ind. Electron. Mag.*, vol. 13, no. 1, pp. 38–41, Mar. 2019, doi: 10.1109/MIE.2019.2893466.
- [12] M. Guarnieri, "Creating the first web: The 19th century expansion of telegraphy," *IEEE Ind. Electron. Mag.*, vol. 13, no. 4, pp. 119–122, Dec. 2019, doi: 10.1109/MIE.2019.2946409.
- [13] M. Guarnieri, "When cars went electric – Part 1," *IEEE Ind. Electron. Mag.*, vol. 5, no. 1, pp. 61–62, Mar. 2011, doi: 10.1109/MIE.2011.940248.
- [14] M. Guarnieri, "When cars went electric – Part 2," *IEEE Ind. Electron. Mag.*, vol. 5, no. 2, pp. 46–47, Jun. 2011, doi: 10.1109/MIE.2011.941122.
- [15] M. Guarnieri, "The beginning of electric energy transmission: Part one," *IEEE Ind. Electron. Mag.*, vol. 7, no. 1, pp. 57–60, Mar. 2013, doi: 10.1109/MIE.2012.2236484.
- [16] M. Guarnieri, "The beginning of electric energy transmission: Part two," *IEEE Ind. Electron. Mag.*, vol. 7, no. 2, pp. 52–59, Jun. 2013, doi: 10.1109/MIE.2013.2256297.
- [17] M. Guarnieri, "Seventy years of getting transistorized," *IEEE Ind. Electron. Mag.*, vol. 11, no. 4, pp. 33–37, Dec. 2017, doi: 10.1109/MIE.2017.2757775.
- [18] R. A. Powers, "Batteries for low power electronics," *Proc. IEEE*, vol. 83, no. 4, pp. 687–693, 1995, doi: 10.1109/5.371974.
- [19] R. M. Dell, "Batteries: Fifty years of materials development," *Solid State Ion*, vol. 134, nos. 1-2, pp. 139–158, 2000, doi: 10.1016/S0167-2738(00)00722-0.
- [20] K. Kordesch and W. Taucher-Mautner, "Primary batteries," *Encyclopedia Electrochem. Power Sources*, vol. 134, pp. 555–564, Dec. 2009, doi: 10.1016/B978-0-44452745-5.00003-4.
- [21] B. Scrosati, "History of lithium batteries," *J. Solid State Electrochem.*, vol. 15, nos. 7-8, pp. 1623–1630, 2011, doi: 10.1007/s10008-011-1386-8.
- [22] C. A. Vincent, "Lithium batteries: A 50-year perspective, 1959–2009," *Solid State Ion*, vol. 134, nos. 1-2, pp. 159–167, 2000, doi: 10.1016/S0167-2738(00)00723-2.
- [23] M. Winter, B. Barnett, and K. Xu, "Before Li Ion Batteries," *Chem. Rev.*, vol. 118, no. 23, pp. 11,433–11,456, 2018, doi: 10.1021/acs.chemrev.8b00422.
- [24] G. E. Blomgren, "The development and future of lithium ion batteries," *J. Electrochem. Soc.*, vol. 164, no. 1, pp. A5019–A5025, 2017, doi: 10.1149/2.0251701jes.
- [25] K. Danaher, S. Biggs, and M. Jason, *Building the Green Economy: Success Stories from the Grassroots*. New York, NY, USA: Routledge, 2016.



LABORATORY TEST BENCH

FOR MOTOR DRIVE APPLICATIONS

imperix
 + SWISS MADE



online store on
imperix.com



Distinguished Lectures in Memory of Galileo Ferraris

Last year, the IEEE Italy Section and Politecnico di Torino had an opportunity to honor the memory of Galileo Ferraris (1847–1897) with the dedication of an IEEE Milestone. On 21 January 2021, a ceremony, remotely followed all around the world, was held for the Milestone, “Rotating Fields and Early Induction Motors, 1885–1888.” The plaque was inscribed with the following citation:

“Galileo Ferraris, professor at the Italian Industrial Museum (now Politecnico) of Torino, conceived and demonstrated the principle of the rotating magnetic field. Ferraris’ field, produced by two stationary coils with perpendicular axes, was driven by alternating currents phase-shifted by 90 degrees. Ferraris also constructed prototypes of two-phase AC motors. Rotating fields, polyphase currents, and their application to induction motors had a fundamental role in the electrification of the world.”

The program included an opening session with the participation of the Politecnico di Torino Rector Guido Saracco, the mayor of Livorno Ferraris; Vice Rector for Research Stefano Corgnati; Vice Rector for Culture and Communication Juan Carlos De Martin; 2021 IEEE President Susan Kathy Land; 2021–2022 IEEE Region 8 Director Antonio Luque; 2019–2021 IEEE Italy Section Chair Bernardo Tellini; and



FIGURE 1 – Galileo Ferraris, one of the prime movers of rotating field formulation in ac electrical machines.

IEEE Italy Section History Activity Coordinator Antonio Savini.

Prof. Gérard-André Capolino,
Department of Electrical Engineering,

University of Picardie “Jules Verne,” Amiens, France, and an IEEE Industrial Electronics Society (IES) Distinguished Lecturer, presented “Progress in ac Electrical Machines: From Galileo Ferraris’s Principle to the Actual Technology.” The lecture can be downloaded free from the IES Education Resource Center at <https://resourcecenter.ies.ieee.org/education/ies-dl-program/IESDLWEB0000.html>.

By the middle of the 19th century, several engineers had envisioned AC rotating electrical machines to convert electrical power into rotating mechanical forces. The first to be developed and commercialized was Zénobe Gramme’s

(continued on page 91)

**Progress in AC electrical machines:
from Galileo Ferraris principle to
the actual technology**

Gérard-André Capolino, PhD, DSc, Life Fellow IEEE
Emeritus Professor of Electrical Engineering - University of Picardie - Amiens - France
IEEE Industrial Electronics Society Distinguished Lecturer

Anciens de la Radioélectricité – IEEE France Section
Life Fellow Program
October 26, 2021

FIGURE 2 – A slide from the October 2021 presentation.



2022 IEEE International Conference on Power Electronics, Smart Grid, and Renewable Energy

The first and very successful biennial IEEE International Conference on Power Electronics, Smart Grid, and Renewable Energy (PESGRE) was held in January 2020 and included activities, discussions, and exchanges of ideas among 450 power electronics and energy systems professionals and researchers from 15 countries. The second edition was conducted in virtual mode from 2 to 5 January 2022, achieving similar success. The IEEE Industrial Applications Society (IAS), IEEE IAS/Industrial

Electronics Society (IES)/Power Electronics Society (PELS) Joint Kerala Chapter, and IEEE Kerala Section were the financial sponsors. The IES, PELS, and IEEE Power & Energy Society (PES) were the technical sponsors, and the conference was organized by the Joint Kerala Chapter (Figure 1).

PESGRE 2022 focused on the challenges, latest developments, and upcoming technologies in power electronics systems, electric drives, renewable energy resources, and smart grid operation. The theme was “Power Electronics and Renewable Energy for Sustainable Development,” and the conference opened with an inaugural address from

Prof. Liuchen Chang, PES president. There were 326 submissions from 20 countries, and after a review involving 36 technical program chairs and 331 reviewers, 176 papers were accepted for camera-ready submission. Special care was taken to ensure a thorough peer review stage, including useful feedback for the authors’ benefit. Papers presented during the conference are eligible for submission to *IEEE Transactions on Industry Applications*, subject to a further round of review.

The technical program consisted of four tutorial sessions, four keynote lectures, one student forum, and 25 technical tracks (Figure 2). The

Digital Object Identifier 10.1109/MIE.2022.3166269

Date of current version: 24 June 2022



FIGURE 1 – The conference organizers.

tutorials were delivered by well-known experts: Prof. Fei Gao (University of Technology of Belfort–Montbéliard, France), Dr. Uday Deshpande (D&V Electronics, Canada), Dr. Amitkumar K.S. (Opal-RT, Canada), Prof. Kaushik Basu (Indian Institute of Science, Bangalore), Dr. Anirban Pal (University of Nottingham, United Kingdom), and Prof. Sandeep Anand (Indian Institute of Technology, Bombay). The keynotes were presented by Prof. Kaushik Rajashekara (University of Houston), Prof. Deepak Diwan (Georgia Institute of Technology, Atlanta), Prof. Joydeep Mitra (Michigan State University, East Lansing), and Prof. H.M. Suryawanshi (National Institute of Technology, Nagpur). The technical sessions were chaired by researchers from

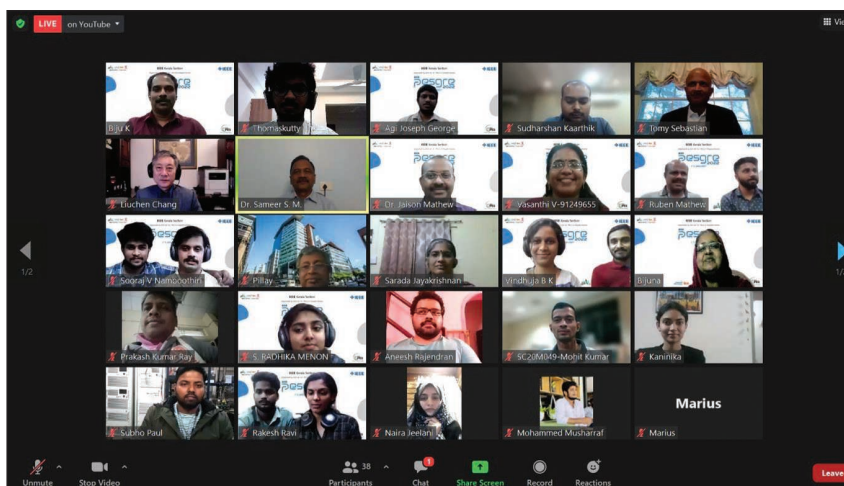


FIGURE 2 – Attendees participate in a conference session.

reputable institutes, and the conference provided an excellent forum for technical exchanges. The next edition

will be held in Kerala, India, on 17–20 December 2023.



A Two-Week Science and Technology Research Workshop for a Girls School

To promote women engineers and scientists and inspire girls in Macao to pursue a career in engineering, especially in IEEE Industrial

Digital Object Identifier 10.1109/MIE.2022.3166356

Date of current version: 24 June 2022

Electronics Society (IES)-aligned disciplines, the Macau IES Chapter organized a two-week science and technology research workshop for a school, with the theme “Electricity and Power Electronics Application in Daily Life,” in July 2021. Form 4 students from the Sacred

Heart Canossian College (English Section), a Catholic girls school in Macao, participated. The event was an opportunity to understand the basics of power electronics technology and learn research topics in the field, such as power management circuits, wireless power



FIGURE 1 – The workshop participants gather for a group photo.

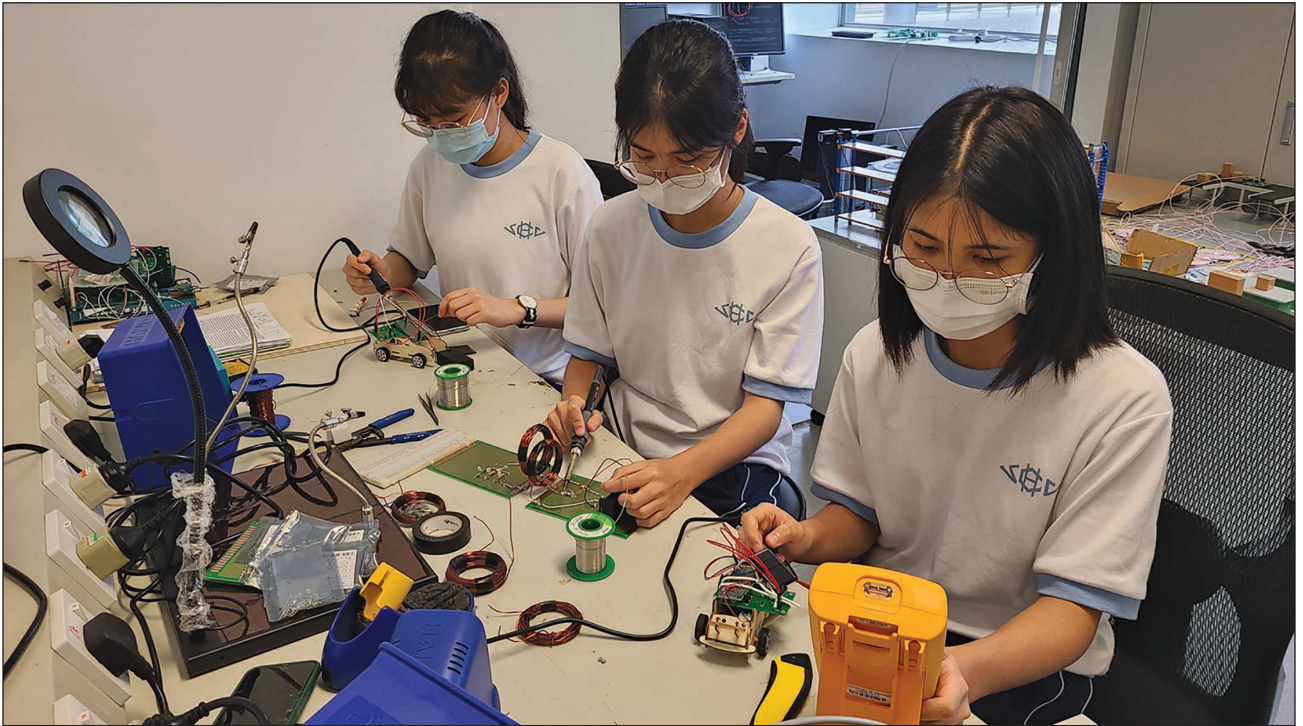


FIGURE 2 – Students solder solar model cars and wireless power transmitters.



FIGURE 3 – Students learn power electronics experimental prototypes in a laboratory.



FIGURE 4 – The participants attend the final presentation.

transfer, solar energy conversion, power quality conditioning, and so on.

The workshop featured hands-on experience with simulation software and through experiments on solar model cars and wireless power transmitters, in addition to technical lectures, visits, and a final presentation with professors and postgraduates at the State Key Laboratory of Analog and Mixed-Signal Very Large-Scale Integration and the Institute of Microelectronics, University of Macau. The students practiced scientific thinking, learned basic research methods, explored international perspectives, and tapped their scientific research potential.

The vice principal of academic affairs at Sacred Heart, Vanessa Cheong, also visited the University of Macau to participate the workshop. She believed that the event was a notable opportunity for the students to learn about science and technology and to put it into practice. She hoped the program could regularly continue.

—Chi-Seng Lam,
Macau IES Chapter chair,
University of Macau





Meet Jenifer Castillo

This month, I have the pleasure of hosting Jenifer Castillo (Figure 1), the 2021–2022 IEEE Women in Engineering (WIE) Committee Chair. She is also the IEEE Region 9 director-elect for 2022–2023, a Senior Member of IEEE, and a 17-year IEEE volunteer.

Lucia Lo Bello: Jenifer, thank you very much for accepting my invitation. Please tell us about your early years of activity in IEEE.

Jenifer Castillo: I started my volunteering as a student in the Universidad de San Buenaventura in Bogotá, Colombia. At that point, I started as a volunteer for WIE and then became the chair and soon afterward the chair of the student branch. Back then, I started to evaluate which could be my home Society and tried the first one I thought was aligned with my career: Mechatronics Engineering. After getting my degree, I was recruited by the Section chair at the time to do some microvolunteering. I am grateful to him; it was the hook that never let me leave IEEE. By then, I started in my position as product specialist for the Parker Hannifin representative in Colombia, and my professional path was starting to lean toward sales, which, at the beginning, was not what I was looking for. At that point, I tried a different Society that I even got to chair in my Section, and we won the best global Section Chapter, but it was still not fulfilling my needs.

Lo Bello: When did you realize the career benefits of being an IEEE Member?



FIGURE 1 – Jenifer Castillo.

Castillo: Besides IEEE being key to getting my first job at Parker Hannifin, when I moved to Ingersoll Rand, as an applications engineer, I had the opportunity to start working in countries outside my beloved Colombia, and up until today, I thank the experience I was having with IEEE as a volunteer because being part of such a multicultural organization has made it easier for me to work with other cultures in the company. I think that was key to the success in the sales coordinator position I was offered. At this point, I had had a good experience in the Colombia Section as the Young Professionals [at the time, Graduate of the Last Decade (GOLD)] coordinator (winning the GOLD Member and Geographic Activity Award),

membership development coordinator, and vice-chair of the Section. I have also had the pleasure to be the founder of the WIE professional chapter in Colombia after some great efforts by previous volunteers.

Lo Bello: Tell us how you became involved in multiple IEEE Societies.

Castillo: By 2014, I moved to Puerto Rico and went back to work with Parker as territory manager of the Caribbean. With Parker, though, the markets I needed to work with were very different, so I was member of the IEEE Power and Energy Society for a couple years to gather information about the Power Generation market I needed, and then the IEEE Communications Society, as there are many efforts related to the Internet of Things, as well. This was very useful and helpful for my career. At this point, I

I HAVE ALSO HAD THE PLEASURE TO BE THE FOUNDER OF THE WIE PROFESSIONAL CHAPTER IN COLOMBIA AFTER SOME GREAT EFFORTS BY PREVIOUS VOLUNTEERS.

started volunteering in the IEEE Puerto Rico and Caribbean Section and started growing there again. I focused first on students and was the Student Activities Committee (SAC) coordinator of the Section, creating a SAC team that would nurture the future leaders of the Section (and in fact, one of them was already the chair of the Section); then, I moved to Young Professionals. I was the

founder of the WIE professional affinity group (AG) from scratch, led the same Society I led in Colombia, and then I was elected chair of the Section. This was definitely determinant.

Lo Bello: How did your career in IEEE Region 9 start?

Castillo: As chair of the Section, and being able to attend the regional meetings and give some results at a Section level, I had my first opportunity to serve the Region. After many years, the Region decided to appoint an industry engagement coordinator, and the director at the time chose me. We started gathering some useful information from the Sections to understand the needs of the members in industry. Today, we have a full committee dedicated to that effort that keeps growing. By 2019, I was the secretary of the Region, learning a lot about the operations but, most of all, learning a lot from the director, Alberto Sánchez, while also being a WIE Committee member. One of the key points I worked out in WIE was the global assessment of our AG, i.e., understand their needs, activities, expectations, and it was enlightening. We do have a high volunteering quality in IEEE.

Lo Bello: What was your next step within WIE?

Castillo: I was appointed WIE committee chair, and this global position brought to me a level of experience that I cannot describe. The opportunity

to work with people from around the globe made me learn that, although we are different, we have much more common ground than we may imagine. It was so beautiful to see that the things that spark joy in Latin America are very similar to those in India, and in that same path are the needs. With this experience, and being the secretary of the Region,

THE OPPORTUNITY TO WORK WITH PEOPLE FROM AROUND THE GLOBE MADE ME LEARN THAT, ALTHOUGH WE ARE DIFFERENT, WE HAVE MUCH MORE COMMON GROUND THAN WE MAY IMAGINE.

I dared to run for Region director. Although I believe IEEE is a very diverse and inclusive organization, I also believe we still need to see more diversity in the leadership positions. Currently, I am still WIE Committee chair, and now I am IEEE Region 9 director-elect. The challenges keep growing, but this only means that I can still make a difference, that I can still nurture the future leaders, and that

IEEE keeps moving forward.

Lo Bello: What are the main items related to TAB on the current WIE agenda?


Castillo: This year, WIE is very focused on supporting the Society and Council coordinators. We have started a round of meetings to identify how to support the operational units efforts regarding women empowerment, the importance of

the pledge, and diverse and inclusive committees and, in general, to build bridges between the coordinators to replicate the best practices and strategies.

Lo Bello: How did you become involved in the IEEE Industrial Electronics Society (IES)? What is your current role in the IES?

Castillo: I came across *IEEE/ASME Transactions on Mechatronics* and was very pleased with the content. So, I found that the IES was behind it, and I paid for the membership. When I received my first magazine, I knew I had found my home Society. I could not put it down; I wanted to read the whole thing. Of course, being in contact with Yousef Ibrahim, the IES vice president for membership activities, has made everything much easier to navigate and to enjoy being part of this Society. I am currently in my first volunteering position as a member of the IES Chapters Committee, and I hope I can contribute to the Society much more in the future as I get to know it and understand what the members may need. I am glad I found the IES.

Lo Bello: And we are glad to have you on board, Jenifer! Let's exploit the momentum for WIE initiatives. Thank you very much, Jenifer.

To follow WIE initiatives, stay in touch with us on LinkedIn (<https://www.linkedin.com/groups/8609923/>). WIE is a community, and you are welcome to join. 



We want to hear from you!

Do you like what you're reading?
Your feedback is important.
Let us know—send the editor-in-chief an e-mail!





by Giuseppe Buja,
Zbigniew Krzemiński,
Marek Adamowicz,
and Marek Jasiński

S&YP + Mentors + Peace + Love = Science and Growing

In this extraordinarily difficult time, we understand better that only peace, love, and cooperation are the keys to growing in technology for humanity. Let us learn from our mentors how they grow from their hard work and international cooperation. Thanks to Prof. Giuseppe Buja and Prof. Zbigniew Krzemiński, we have unique schools of adjustable speed drives that are helping people convert electrical to mechanical power and vice versa. Let's learn how it was possible.

Prof. Giuseppe Buja

IEEE Industrial Electronics Magazine (IEM): Prof. Buja, provide us with a glimpse into your research journey.

Prof. Giuseppe Buja: My journey began in 1964 by enrolling in an electronic engineering course at the University of Padova (UNIPD) and complementing the electronic background with extra teachings from the electric engineering syllabus. Right after the “Laurea” degree, I joined UNIPD for an apprenticeship on semiconductor power circuits that at the time were composed of thyristor and diode devices. My first research involvement was in the mid-70s, when I developed the PWM (pulsewidth modulation) control of power inverters. In the

80s, I focused my research on implementation of the control of power converters in high-processing microcomputers that had just been introduced at that time, in an intriguing coincidence with power transistors. I still feel thrilled thinking about when I programmed the first-marketed DSP (digital signal processor) (Intel 2020) to control a PWM inverter built with a three-phase Toshiba IGBT bridge. In 1992, I moved to the University of Trieste, where I did research on ac electric drives. At the turn of the 2000s, I returned to UNIPD and shifted my interests from the equipment to the systems. I founded the Laboratory of Electric Systems for Automotive and Automation, giving it the mission of developing research and educational activities for mobility and energy systems by merging electric, electronics, and informatics technologies. As the times were ripe for the penetration of electric systems into vehicles, I directed the lab's research right away into drive-by-wire systems and then into the powertrains of electric vehicles (EVs). The next step was research on EV battery charging, both the wired and wireless type. Before retirement, I turned the lab's research toward conversion systems, enabling grid integration of renewable sources. With a retrospective look, I can say that my research journey has constantly evolved, chasing my innate

passion for new technologies and emerging topics.

IEM: Were your side research activities equally as lively?

Prof. Buja: Yes, really. Nowadays, it is hard to imagine how bounded the academic activities were in the early 80s. Attendance at the conferences was typically limited at the annual domestic symposium, papers were mainly published in national journals, and news on power electronics was taken almost exclusively from the library books (written, incidentally, by authors not from my country). A little at a time, I realized that this context was not productive for my research and decided to shape it in an open and international way. To begin with, I sent my research notes (by postal, no email at that time) to the author of one of my cult books on power electronics: Prof. J. Murphy at the University of Cork in Ireland. Much to my surprise, he invited me to his university, and this unique experience was the cornerstone of my side research activities. I commenced attending conferences abroad, publishing papers in international journals and to visit renowned labs. As soon as I had the chance, I enjoyed the dual experience of inviting external researchers and tutoring foreign students in my lab. An activity, however, makes me particularly proud even now. It was the attendance at conferences in Eastern European countries before the fall of

the Berlin Wall. Despite the red tape to enter them, I keep a vivid memory of the warm hospitality received by local researchers, which subsequently turned into a still-active friendship. In addition to this activity, I was honored to receive an invitation from Prof. I. Nagy from Budapest University in Hungary to help him promote the Power Electronics and Motion Control (PEMC) Council and PEMC conferences all around the world.

IEM: Let us now talk about your involvement with IEEE and specifically, with the IEEE Industrial Electronics Society (IES)?

Prof. Buja: My thinking is, if there was no IEEE, it would have to be invented. I am greatly indebted to it for many reasons. IEEE was like a gym for me as it trained me in research. Reading and deepening *IEEE Transactions* papers allowed me to stay up to date and hone my skills on my research issues. IEEE was also a springboard for me. Indeed, the publication of my first papers in *IEEE Transactions* has opened the doors of many labs to me, even in my country. Attendance at IEEE conferences was another chance for me as I personally met scholars from every country and presented my research results to an international audience. It was also a chance for me to realize how effective the social events of the conferences are in educating attendees to respect different cultures and civilizations.

I approached the IES in the late 70s, when the name of the Society was *Industrial Electronics and Control Instrumentation (IECI)*. At that time, I was researching the microcomputer control of power converters and was attracted by the IECI's initiatives on this issue. In 1980, for the first time, I attended the Society's annual meeting, which was publicized as the IECI Conference on Industrial Applications of Microcomputers. I immediately had a good feeling from the officers and, some years later, got involved in technical and administrative tasks. This did not surprise me too much because the Society had been distinguished for many years by the inclusion of outside-U.S. people in its bodies. So, in the late 80s, with the support of the

president, Prof. F. Harashima, I organized two international workshops in my country on microcomputer control of electric drives, which were well attended and honored by the presence of top researchers like Prof. B.K. Bose and Prof. K. Ohnishi. A few years later, the president, R. Begun, proposed I chair the first European edition of IECON (the Annual Conference of the IEEE Industrial Electronics Society) in 1994. I accepted all at once without weighing the relevant workload because I was confident in the help of IES officers and local volunteers, such as Prof. C. Cecati, to name one. Among the officers, I'd like to express a heartfelt tribute to Prof. R. Niedjohn, an exquisite person who visited me before the conference to offer me his strong encouragement.

IECON'94 took place in Bologna, Italy. Up until then, it was the top-attended IECON and remained in the collective memory for the high level of its scientific contributions. Subsequently, I was appointed Vice President (VP) for Small Conferences (later renamed for Workshops). It was a demanding and challenging assignment as there was a flurry of new workshops and symposia in those years. In this regard, I am glad to talk about conception of the International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives (SDEMPED), which occurred in 1996 in an airport room while waiting for a delayed flight. There I met Prof. G. Capolino, and we agreed on the increasing interest in diagnostics for power electronics equipment. At the end of the meeting, we laid out a plan to launch

the SDEMPED Symposium, which happened a year later in the charming town of Carry-le-Rouet, France. For my VP duties, I had the opportunity to strictly work with IES officers, especially the president, Prof. J.C. Hung. It was a very stimulating trial that enriched me greatly from a professional point of view. From the late 90s, I supported the IES, mainly in technical activities because of my increasing academic commitments and efforts in running my lab—charged with many research programs immediately after its start-up.

It is evident how much I am technically, professionally, and socially satisfied to be a Member of IEEE, and specifically, the IES. It was therefore natural for me to encourage students and colleagues to join and participate in the life of the Institute and its Societies. About that, I am pleased to mention some students of mine whom I have introduced to the IES: M. Valla [postdoc (postdoctoral) student] and R. Oboe and R. Keshri (Ph.D. students).

IEM: Did you have exciting experiences while performing your side research activities?

Prof. Buja: Of course, especially while attending conferences. I would like to go off the technical rails and talk about two experiences that are etched in my memory. The first one occurred in the 80s after the gala dinner of a conference in the Tatra Mountains (Romania). I was tired after a hard trip arriving at the conference site, and I decided to anticipate my colleagues by walking back to the hotel. On the way, I found myself in front of a bear. I was terrified, but instinctively stood there



IES President, Mr. C. Einhoff, gives the E. Mittelman award to Prof. G. Buja.

with a tough face. The bear came up to me at a brisk pace and, maybe because I did not run away or had a tough face, it turned around and slowly went into the bush. When I go back to this event in my mind, I am convinced that this instinct helped me solve some problems more than long reasoning. The other experience occurred to me in the 90s at IECON in Maui, Hawaii. One day, I decided to have an afternoon of leisure and booked a tour in a submarine to enjoy the seabed around the island. When I took my place I jumped on the seat like a scholar, discovering that one of my idols, Prof. W. Leonhard, was sitting next to me. I introduced myself shyly while he greeted me warmly and put me at ease by saying that he knew my papers. To strike up a conversation with him, I asked about the implementation of field-oriented control (FOC), but he diverted the talk and told me about him. So, while colored fish passed in front of the portholes, I found out the genesis of FOC theory, I came to know that he had practiced on the drives when he was in charge of the electric equipment of a submarine during the Second World War, and much more. I listened to him all along the tour and, when we went ashore, I grasped how simple and friendly the great scientists are.

IEM: What advice do you have for young researchers?

Prof. Buja: The choice to undertake a career as a researcher in the engineering field must be dictated by passion—a great passion. Indeed, a researcher must work on “a matter in the making,” not on an established one. This means that he or she must believe strongly in the force of his or her ideas to overcome the uncertainties that inevitably arise in carrying out a research task. It is equally important that he or she be strong enough in recognizing when an idea is wrong or unpractical, and in starting over. He or she must also be conscious that, behind a good achievement, there are long times spent thinking and experimenting even if the idea was originated from an instant inspiration.

Another piece of advice to a young researcher is that he or she must go

leave the comfort zone of his or her lab to gain work experience, besides that of research, in labs abroad or in a company. Nowadays, there are a lot of opportunities that he or she should not miss. This helps him or her grow in many ways, such as in confronting his or her knowledge, assessing his or her competence, getting inspiration, and even discovering methodologies that advance his or her abilities. In this regard, I mention the research path of a student of mine, Prof. R. Oboe. Under my guidance, he spent a part of his Ph.D. course in Japan at the laboratory of Prof. Ohnishi and went back with a wealth of skills in the motion control field, which he has deeply nurtured to excel in the scientific community.

For their part, tutors must be fair in helping and evaluating their students; and first of all, must be masters of life by teaching, together with the science, the values of intellectual honesty, behavioral correctness, and the conscious use of research results.

IEM: How has retirement changed your life?

Prof. Buja: Much has changed. I continue to carry out research activities and pursue my hobbies, but now timings are reversed as most are for hobbies. I like walking outdoors, especially along the banks of the numerous rivers that pass through Padova, where I live. I like reading books about contemporary history and listening to pop music. Research activities are mainly aimed at supporting the work of students and colleagues, with hints and discussions, but sometimes I inspire them with novel concepts, sticking to my motto: “research is a state of mind.”

Prof. Zbigniew Krzemiński: Inspiring Young Engineers and Scientists to Take on Great Industrial Challenges

Support from an experienced mentor who will help time and reliably show directions for further development is very important for young scientists at the initial stage of their career. In this column, we present the profiles of mentors who have had an extraordinary impact on the development of the scientific career of young engineers and scientists.

Among them, we present Prof. Zbigniew Krzemiński from the Gdańsk University of Technology (Gdańsk Tech), Poland. Over the 47 years of his academic career, Prof. Krzemiński has mentored several dozen young professionals, including 15 promoted doctors in the field of automation, electronics, and electrical engineering. Some of them are already well-known professors, such as Prof. Haitham Abu-Rub from Texas A&M University at Qatar, who is known to the readers of this column.

In 2004, Prof. Krzemiński went beyond strictly academic activities, and together with Ph.D. students created the university start-up MMB Drives, which, going through the entire development cycle of start-ups, has become a professional technology company [2]. And just as in the scientific world, where a professor's mentees successfully solve new scientific problems, in the industrial field, a team of Ph.D. students and doctors from MMB Drives has been uncompromisingly facing the greatest challenges in almost all industries, including the power industry, oil and gas, renewable sources, rail and sea transport, electromobility, and, of course, the machine industry (see Figure 1). The headquarters of MMB Drives is shown in Figure 2. Befitting a technological company, MMB Drives has its own power sources in the form of a wind farm, photovoltaic panels, and heat pump.

It should be emphasized that during these 18 years, the MMB team did not lose its competitive gene won the 2021 Grand Challenge: Energy competition, which was implemented as a part of the Grand Challenge: Formula competition, organized in Poland by the National Center for Research and Development. Figures 3 and 4 presents the winners and MMB Team, including Prof. Krzemiński, all of whom are involved daily in engineering power electronics and control systems for renewable energy and variable-speed drives.

The “Grand Challenge: Energy” competition began with close to 200 teams of constructors from all over Poland presenting the work of compact devices for individual applications capable of converting wind energy into electricity, then storing it and returning

it most effectively. The wind turbine prototypes had to meet parameters determined in the Participant Manual, i.e., their dimensions could not exceed 2 m³ nor 200 kg. Moreover, the systems had to be esthetic and silent. The wind power plant developed by Prof. Krzemiński's team had been equipped with high-power factor wind with an axial-flux permanent-magnet generator and a silicon carbide (SiC)-based power converter system, ensuring the effective management of energy storage. The developed shape of the blades and the control algorithm ensured that during the competition the device could work without power limitations against strong winds and used maximal wind energy for low and high speeds.

The secret of Prof. Krzemiński's success consists of combining industrial and university experience. The beginning of his career dates back to the mid-1970s when Poland was a member of the so-called Socialist Bloc behind the Iron Curtain under the influence of the Soviet Union. According to the Soviet doctrine, free contacts of scientists of socialist countries with Western scientists were then forbidden. Moreover, financial support of scientific development in those years was highly ineffective due to the inefficiency and low effectiveness of the whole socialist bureaucratic machine. Therefore, scientific achievements in the field of power electronics and industrial electronics in Poland were based on only the determination of individual inventors and researchers and their direct cooperation with the industry. It was therefore necessary to develop such solutions needed for the industry which would at the same time inspire young doctoral students and scientists for the development of their careers.

Prof. Krzemiński's first achievement, when he was a doctoral student, was developing a controlled thyristor drive with an electronic controller and two parallel operating stepper motors for the textile industry. Controlling two motors in parallel was aimed at increasing the power. The solution's innovation was the use of the impulse splitter for thyristors with a flip-flop system, which was resistant to the interference

from control signals. Shortly thereafter, he was ready to face the greatest life challenge: developing the original, nonlinear control of an induction motor, called *multiscalar control* [3]. This challenge was related to the implementation of his doctoral dissertation and, as a result, determined his entire future academic career as a professor.

It has just been a decade since Blaschke proposed FOC in 1972 [4],

enabling one to control an induction motor like a separately excited dc motor. FOC, which was based on decomposition of the instantaneous stator current into two components: flux current and torque-producing current, ensured, in its first industrial applications, precise control of variable-speed drives, but assuming that the magnetic flux of the motor would be kept constant. Today, numerous modifications

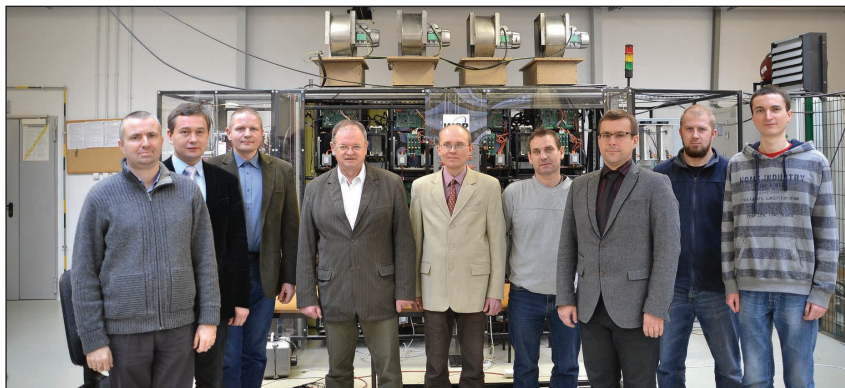


FIGURE 1 – Prof. Krzemiński's university collaborators and mentees with the world's first prototype of a medium-voltage power electronic transformer for the oil and gas industry [1].



FIGURE 2 – The headquarters of the technology company MMB Drives has its own power sources in the form of a wind farm, photovoltaic panels, and heat pump.



FIGURE 3 – The MMB team, led by Prof. Zbigniew Krzemiński. The team won the 2021 Grand Challenge: Energy competition.

and developments of FOC, thanks to the decoupling members of velocity and flux dynamics, allow for free adjustment of the induction motor excitation [5], but at that time, new methods of nonlinear control of induction motors were of particular interest.

The idea of nonlinear control was to apply the transformation of vector variables of the induction motor model to new variables, independent of the reference system—one of which was proportional to the motor torque—and then the use of linearization of the model and control obtained, which resulted in decoupling the dynamics of the regulated speed and flux of the induction motor. For linearization of the nonlinear equations of the induction motor multiscalar model, Prof. Krzemiński used the theory of structural synthesis of automatic control nonlinear systems developed by Ukrainian scientist L.M. Boychuk [6]. In multiscalar control, according to Prof. Krzemiński, the regulation of torque and flux is no longer performed in a rotating coordinate system, but instead uses cross and scalar products of stator current and rotor flux as the products contain complete information about the mutual position of the

stator current vector and rotor flux vector. The nonlinear feedback allows linearization and therefore simplification of the mechanical and electromagnetic control loops. In the second half of the 1980s, it became possible for Polish scientists to go to Western Europe. Thanks to a Robert Bosch scholarship, Prof. Krzemiński was able to attend the 1987 International Federation of Automatic Control Congress, where, for the first time, he presented his proprietary method for nonlinear control of induction motors [3].

The fall of the Berlin Wall and the liberation of Eastern Europe from the domination of the Soviet Union in 1989–1991 allowed scientists from Poland to communicate freely with their colleagues around the world. In 1998–2000, Prof. Krzemiński, in cooperation with the German company AvK SEG Kempen (later Woodward Kempen) and together with his Ph.D. student Andrzej Geniusz, actively participated in the development of multiscalar control for offshore wind turbines with double-fed induction generators [7], [8]. As a result of this cooperation, AvK SEG Kempen awarded Gdańsk Tech an order to develop a control system for a high-power

wind turbine generator. “We were invited to collaborate to develop an industrial sensorless control system for a double-fed machine used in high-power wind farms. It required the development of an original induction machine speed observer. The system was applied in practice, and the subject of cooperation was extended to the issues of generator stability in conditions occurring in a wind farm,” recalls Prof. Krzemiński. Geniusz currently works at Woodward Kempen as a specialist responsible for the Concycle series of multiscalar converters. Today, Concycle converters are installed in more than 15,500 wind turbines worldwide.

The interesting industrial challenge undertaken by Prof. Krzemiński’s team was the development of new-generation variable-speed drive solutions that reduce footprints and operating costs on oil-production platforms.

For the first time in the world, Prof. Krzemiński used a megawatt-scale, SiC-based power electronic transformer for supplying electric-submersible pump drives (see Figure 5). His team has also developed innovative down-hole monitoring systems using high-temperature electronics and created a reduced-diameter, high-efficiency, permanent-magnet motor for retrofitting existing production wells (see Figure 6). The low- and medium-voltage, variable-speed drives and control devices developed by MMB Drives have been used in more than 900 production wells worldwide.

As Prof. Krzemiński emphasizes, the team’s function implementing projects in the field of modern electric drives and power electronic converters requires specialization and continuous improvement of the members of the research team themselves. The current state of the theory of electric machines control enables high-quality regulation of selected quantities, provided that the latest achievements of electronics and power electronics are used, which are developing very quickly. This creates special conditions for conducting research, consisting of continuous improvement of the developed solutions, and testing the possibility of using electronic systems with the highest parameters.



FIGURE 4 – A joint social event with Prof. Krzemiński’s university associates and engineers and doctoral students from MMB Drives.



FIGURE 5 – (a) Prof. Krzemiński with his Ph.D. student Janusz Szewczyk at Alkhorayef Petroleum Company, an electric submersible pump (ESP) factory in Saudi Arabia. (b) The world's first power electronic transformer-based, medium-voltage, variable-speed drive with ESP installed on the test rig [1], [9].

Carrying out research as a part of every planned project requires the employment of a team of professionals with genuine scientific and research achievements, especially in the field of laboratory work. The role of the mentor, in this case, is to motivate the team to take up challenges that go beyond the achievements thus far, but also to give a helping hand when the young specialist thinks that he or she has hit a dead end. “Despite the very rapid development of technology all over the world, young specialists are sure to face many challenges—new types of engines that will be able to use battery energy more efficiently, or more efficient hydrogen production technologies. We are also working on all of this at MMB Drives. At the same time, I inspire my younger colleagues to always be accompanied by the element of fun in achieving success,” advises Prof. Krzemiński.

Also, thanks to Prof. Buja and Prof. Krzemiński, the world divided by the Berlin Wall has been unified and freedom has won out. This is our role: to make science in peace, and respect others.

References

- [1] D. S. Shanks, J. Pietryka, J. Szewczyk, J. Samsel, and Z. Krzemiński, “Systems and methods of power transmission for downhole applications,” U.S. Patent and Trademark Office, Washington, DC, USA, U.S. Patent No. 10 968 726, 2021.
- [2] A. Anna, M.-K. Ewa, K. Zbigniew, and A. Marek, “Selected financial-economic aspects of R&D in renewable energy conversion technologies. The case of University spin-off company,” in



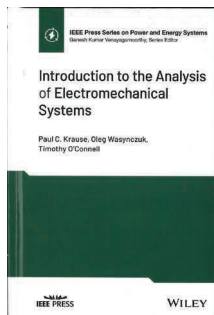
FIGURE 6 – High-precision downhole monitoring system using high-temperature electronics developed by MMB Drives.

- Proc. 5th Int. Conf. Ecol. Veh. Renewable Energies, EVER*, 2010, pp. 1–7.
- [3] Z. Krzemiński, “Nonlinear control of induction motor,” *IFAC Proc. Vol.*, vol. 20, no. 5, pp. 357–362, 1987, doi: 10.1016/S1474-6670(17)55396-3.
- [4] B. Felix, “Das Prinzip der Feldorientierung, die Grundlage für die transvector-Regelung von Drehfeldmaschinen,” *Siemens-Z.*, vol. 45, pp. 757–767, 1971.
- [5] M. P. Kazmierkowski, R. Krishnan, and F. Blaabjerg, Eds., *Control in Power Electronics*, vol. 17. San Diego, CA, USA: Academic, 2002.
- [6] L. M. Boychuk, *An Inverse Method of the Structural Synthesis of Automatic Control Nonlinear Systems*. Kyiv, Ukraine: Naukova Dumka, 1966, p. 7.
- [7] A. Geniusz and Z. Krzemiński, “Control system based on the modified multiscalar model for the double fed machine,” in *Proc. PCIM Eur. Conf.*, 2005, pp. 1–6.
- [8] A. Geniusz, “Power control of an induction machine,” U.S. Patent No. 7 423 406, 2008.
- [9] A. M. Zahrani, Y. Windiarto, and T. Orłowski, “First installation of NEMA 4 medium voltage drive in Saudi Aramco fields,” in *Proc. SPE Middle East Conf. ‘Artif. Lift’*, Manama, Bahrain, Nov. 26–27, 2014, pp. 1–8.



by Fernando A. Silva

Introduction to the Analysis of Electromechanical Systems



By Paul C. Krause, Oleg Wasynczuk, and Timothy O'Connell, IEEE Press (Wiley), 2022, ISBN-13: 978-1119829997 (hardcover), 256 pages; ISBN-10: 1119829992 (electronic)

Master's degrees related to aeronautics, aviation, avionics, space, and aerospace engineering involve a synthesis of theoretical foundations and advanced technologies, which have increasing importance in the 21st century. Modern theoretical foundations and technologies are integrated into aerial vehicles, such as airplanes, helicopters and drones, and support associated services, including air traffic management and operational and maintenance activities. Aerospace is a rapidly expanding area, with new techniques being developed and existing ones perfected and made easier to incorporate as standard technologies. Aerospace-connected master's degrees must include atmospheric, orbital, and interplanetary flight dynamics together with a wide range of knowledge, including aerodynamics, propulsion, structures, materials, manufacturing processes, hydraulics,

pneumatics, control, computing, electronics, telecommunications, machine learning, electrical actuators, and electric energy.

The new book *Introduction to the Analysis of Electromechanical Systems* is a well-written and authoritative text on electromechanics, including several industrial electronics subjects, such as electric machines, power electronics, electric drives, and electric systems. It is an excellent source of key concepts and ideas for aerospace-related master's students aiming to grasp the analytical foundations of electric machines, power electronics, electric drives, and electric power systems. These concepts are key to understanding and designing all electric aircrafts and electric vertical take-off and landing (eVTOL) vehicles using electric power to take-off, hover, flight, and land vertically.

The book is suited to an introductory course on the fundamentals of electromechanical systems. It includes six chapters on fundamental topics of electrical systems and electrical machine modeling and is oriented toward the control of electric drives through power electronics converters and electrical power systems. The chapter topics are as follows:

- basic system analysis
- fundamentals of electric machine analysis
- electric machines
- power electronics
- electric drives
- power systems.

The first chapter provides sufficient detail to enable students to understand the fundamentals of electric

systems, such as power calculations, phasor analysis, active and reactive power, magnetics, field energy and coenergy, transformers, and two- and three-phase electric systems. Chapter 2 is about the fundamentals of electric machines. It covers coupled circuits in relative motion; electromagnetic force and torque; winding configurations; rotating air gap magnetomotive force; two-pole, two- and three-phase stators; reference frame transformations; stator equations in reference frames; and instantaneous and steady-state phasors. Machines with higher numbers of pole pairs complete the chapter.

The third chapter studies electric machines, such as dc, permanent magnet dc, permanent magnet ac, and symmetrical induction machines. Attention is given to voltage, current, and torque laws, synchronous rotating frames, steady-state analysis, ac machine torque, phasors, and steady-state equivalent circuits. Concepts are illustrated using examples. Chapter 4 deals with power electronic converters. It includes switching circuit fundamentals, principles of power electronic conversion, switches and switching functions, energy storage elements, dc-dc converters, ac-dc converters, and dc-ac inverters. Examples of converter design/analysis are also provided.

Electric drives are introduced in chapter 5, which begins with the study of dc drive averaged models and block diagrams and proceeds to torque control. Brushless dc drive operation and torque control are studied next. An induction to motor drives and their torque control is included. The sixth chapter familiarizes readers with electric power

systems and synchronous generators. It introduces the most common three-phase wye and delta connections, ideal transformers, synchronous generators, reactive power and power factor correction, per-unit systems, transient stability, and three-phase faults.

Examples and end-of-chapter exercises/problems invite students to use the concepts themselves. The chapters also include important references to specialized texts. This comprehensive book, included in the IEEE Press Series on Power

and Energy Systems, offers a preface, appendixes with abbreviations, constants, conversions, trigonometric identities, a table of contents, and an index.

Paul C. Krause is a Life Fellow of IEEE, 2010 recipient of the IEEE Nikola

Tesla Award, founder of P.C. Krause & Associates, and former professor of electrical and computer engineering at Purdue University. Oleg Wasynczuk is a Fellow of IEEE, 2008 recipient of the IEEE Cyril Veinott Award, chief technical officer of P.C. Krause & Associates, and professor of electrical and computer engineering at Purdue. Timothy O'Connell is a Senior Member of IEEE, senior lead engineer at P.C. Krause & Associates, adjunct research assistant professor of

electrical and computer engineering at the University of Illinois at Urbana-Champaign, and associate editor of *IEEE Transactions on Aerospace and Electronic Systems*. The authors have written or cowritten hundreds of tech-

nical papers and many textbooks on electric machines and electrified aircraft propulsion.

Introduction to the Analysis of Electromechanical Systems is a unique and essential practical guide, mainly for aerospace, electronics, and electrical undergraduates, providing fundamental methods and hands-on examples and exercises so that students and even researchers from related fields can become familiar with methods for electromechanical systems. The book is also suited to librarians and senior undergraduate and graduate students, practicing professionals, and professors in aerospace-related fields, providing concepts to further explore new trends in all electric powered aircraft technology.

—Fernando A. Silva
*Instituto Superior Técnico, INESC-ID,
Universidade de Lisboa, Portugal*



Society News (continued from page 78)

dc motor (1871), which generated a torque on its shaft via a rotating field provided by brushes passing electrical energy to a collector. If Nikola Tesla was the first to imagine the basic principle of the induction motor (also called an *asynchronous motor*) just prior to emigrating to the United States (1879), Ferraris was at the center of ac rotating field formulation. He, in 1885, and Tesla, in 1887, both developed prototype induction motors with two-phase stator windings. However, the first three-phase induction motor, similar to what we now have in industry, was developed by Mikhail Dolivo-Dobrovolsky, in 1888.

Prof. Capolino reviewed the general principles of rotating magnetic

fields, after which he described the motors of Ferraris, Tesla, and Dolivo-Dobrovolsky and detailed their evolution through examples from General Electric and Hitachi. He developed a classification of electrical machines as key elements of the modern industry for electricity production and for electromechanical energy conversion. He displayed recently designed electrical machines that use innovative materials and novel electromagnetic structures. At the end, he described technical improvements, such as increasing energy efficiency, providing fault tolerance, and advanced winding configurations. The lecture was by attended by 190 people from around the

world, including many students. It was followed by a 15-min question-and-answer session.

A similar, though more detailed, lecture was given on 26 October 2021 for the IEEE France Section Life Member Affinity Group. Nearly 90 people, most of them IEEE members, attended the presentation, which was held remotely because of the pandemic. These IES Distinguished Lectures added a positive note to the year.

—Gérard-André Capolino
*IEEE Life Fellow
IES past president (2012–2013)*





The IEEE Industrial Electronics Society (IES) is closely monitoring the evolution of the COVID-19 pandemic. The safety and well-being of participants are our priority. For updated information regarding a specific conference, you may visit its website.

Please send information regarding conferences to Yang Shi (vp-conferences@ieee-ies.org) and Huijun Gao (vp-workshops@ieee-ies.org).

-  IES Majority Sponsored Conferences
-  Other Cosponsored Conferences
-  Technically Cosponsored Conferences

2022

JUNE

GPECOM 2022

Fourth Global Power, Energy, and Communication Conference
Cappadocia, Turkey
14–17 June 2022
<https://gpecom.org/2022/>

ICAT 2022

2022 28th International Conference on Information, Communication, and Automation Technologies
Sarajevo, Bosnia and Herzegovina
16–18 June 2022
<https://icat.etf.unsa.ba/>

SPEEDAM 2022

26th Edition of the IEEE International Symposium on Power Electronics, Electrical Drives Automation, and Motion Sorrento, Italy
22–24 June 2022
<http://www.speedam.org/>

ICCIA 2022

Seventh International Conference on Computational Intelligence and Applications
Nanjing, China
24–26 June 2022
<http://iccia.org/>

icSmartGrid 2022

Tenth International Conference on Smart Grid
Istanbul, Turkey
27–29 June 2022
<http://www.icsmartgrid.org/>

DoCEIS 2022

13th Doctoral Conference on Computing, Electrical, and Industrial Systems
Caparica, Portugal
29 June–1 July 2022
<https://doceis.dee.fct.unl.pt/>

CPE-POWERENG 2022

16th International Conference on Compatibility, Power Electronics, and Power Engineering
Birmingham, United Kingdom
29 June–1 July 2022
<https://uobevents.eventsair.com/ieee2022/>

JULY

YEF-ECE 2022

Sixth International Young Engineers Forum on Electrical and Computer Engineering
Caparica, Portugal
1 July 2022
<http://sites.uninova.pt/yef-ece>

AIM 2022

2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics
Sapporo, Japan
11–15 July 2022
<https://www.aim2022.org/>

INDIN 2022

2022 IEEE 20th International Conference on Industrial Informatics
Perth, Australia
25–28 July 2022
<https://2022.ieee-indin.org/>

HSI 2022

15th International Conference on Human System Interaction
Melbourne, Australia
29–31 July 2022
<https://hsi2022.welcometohsi.org/>

AUGUST

IAI 2022

Fourth International Conference on Industrial Artificial Intelligence
Shenyang, China
24–27 August 2022
<http://journal13.magtech.org.cn/iai2022/home>

SEPTEMBER

ICEM 2022

25th International Conference on Electrical Machines
Valencia, Spain
5–8 September 2022
<https://icem2022.com/>

SEST 2022

2022 International Conference on Smart Energy Systems and Technologies
Eindhoven, The Netherlands
5–8 September 2022
www.sest2022.org/

AE 2022

27th International Conference on Applied Electronics
Pilsen, Czech Republic
6–7 September 2022
<https://www.appel.zcu.cz/>

ETFA 2022

2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation
Stuttgart, Germany
6–9 September 2022
<https://2022.ieee-etfa.org/>

PEMC 2022

2022 IEEE 20th International Power Electronics and Motion Control Conference
Brasov, Romania
25–28 September 2022
<https://ieee-pemc2022.org/>

RW 2022

Resilience Week 2022
Washington, D.C., United States
26–29 September 2022
www.resilienceweek.com

OCTOBER

IECON 2022/ICELIE 2022

48th Annual Conference of the IEEE Industrial Electronics Society and Ninth IEEE International Conference on E-Learning in Industrial Electronics
Brussels, Belgium
18–21 October 2022
<http://www.iecon2022.org>

IROs 2022

IEEE/RSJ International Conference on Intelligent Robots and Systems
Kyoto, Japan
23–27 October 2022
<https://iros2022.org/>

NOVEMBER

INDEL 2022

24th International Symposium on Industrial Electronics and Applications
Banja Luka, Bosnia and Herzegovina
9–11 November 2022
www.indel.etfbl.net

DECEMBER

ICIEA 2022

17th IEEE Conference on Industrial Electronics and Applications
Chengdu, China
16–19 December 2022
<http://www.ieeeciea.org/2022/>

2023

JANUARY

SII 2023

2023 IEEE/SICE International Symposium on System Integration
Atlanta, Georgia, United States
17–20 January 2023
<https://www.sice-si.org/SII2023/>

MARCH

ICIT 2023

2023 IEEE International Conference on Industrial Technology
Orlando, Florida, United States
8–10 March 2023
Home page: TBA

ICM 2023

2023 IEEE International Conference on Mechatronics
Loughborough, United Kingdom
15–17 March 2023
<https://www.welcometoimago.com/icm2023/>

APRIL

WEMDCD 2023

Sixth IEEE Workshop on Electrical Machines Design, Control, and Diagnosis
Newcastle upon Tyne, England
13–14 April 2023
Home page: TBA

MAY

ICPS 2023

2023 IEEE Sixth International Conference on Industrial Cyber-Physical Systems
Wuhan, China
8–11 May 2023
Home page: TBA

IEMDC 2023

IEEE Electrical Machines & Drives Conference
San Francisco, California, United States
14–17 May 2023
Home page: TBA

JUNE

ISIE 2023

2023 IEEE 32nd International Symposium on Industrial Electronics
Helsinki, Finland
19–21 June 2023
Home page: TBA

AIM 2023

2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics
Seattle, Washington, United States
28M–30 June 2023
Home page: TBA

JULY

INDIN 2023

2023 IEEE 21st International Conference on Industrial Informatics
Lemgo, Germany
18–20 July 2023
Home page: TBA

IESES 2023

2023 IEEE Third International Conference on Industrial Electronics for Sustainable Energy Systems
Shanghai, China
26–28 July 2023
Home page: TBA

AUGUST

SDEMPED 2023

14th Edition of the IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics, and Drives
Chania, Greece
28–31 August 2023
Home page: TBA

SEPTEMBER

ETFa 2023

2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation
Sinaia, Romania
12–15 September 2023
Home page: TBA

OCTOBER

IECON 2023

49th Annual Conference of the IEEE Industrial Electronics Society
Singapore
16–19 October 2023
Home page: TBA

2024

JUNE

ISIE 2024

2024 IEEE 33rd International Symposium on Industrial Electronics
Ulsan, South Korea
18–21 June 2024
Home page: TBA

OCTOBER

IECON 2024

50th Annual Conference of the IEEE Industrial Electronics Society
Chicago, Illinois, United States
13–16 October 2024
Home page: TBA

2025

AUGUST

SDEMPED 2025

15th Edition of the IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics, and Drives
Grapevine, Texas, United States
24–27 August 2025
Home page: TBA



Recruit a Member. *Earn Rewards!*

Take advantage of our Member Get-A-Member Program today!

Your personal and professional experiences with IEEE make you uniquely qualified to help bring in new members. With the Member Get-A-Member (MGM) Program you can get rewarded for word-of-mouth referrals. Earn incentives and awards while helping to grow IEEE membership.

Earn up to US\$90 on your membership renewal dues!



Learn more about
the MGM Program at
www.ieee.org/mgm

