Proceedings & Highlights of the 3rd Emerging Tech Conference Edge Intelligence (ETCEI 2024)[1]

Volume 03 | 2024

DOI: 10.63438/LTPZ5982

[1] https://conference.hetia.org/emerging-tech-conference-edge-intelligence-2024/

Publication Information

Published by
Hellenic Emerging Technologies Industry Association (HETiA)

Volume DOI: 10.63438/LTPZ5982

This volume has been assigned a Digital Object Identifier (DOI) to ensure persistent identification and long-term accessibility.

Individual full papers associated with ETCEI 2024 have been assigned separate Digital Object Identifiers (DOIs) to support persistent citation and academic indexing. Poster abstracts are published without individual DOIs.

Preface

This volume of the *Emerging Tech Conference – Edge Intelligence 2024 (ETCEI 2024) Proceedings* includes 24 peer-reviewed full papers and 7 accepted posters presented at the 3rd Emerging Tech Conference – Edge Intelligence, held on October 17–18, 2024, in Volos, Greece. The conference was co-organized by the Hellenic Emerging Technologies Industry Association (HETiA) and the University of Thessaly. The volume has been assigned a Digital Object Identifier (DOI: 10.63438/LTPZ5982), while individual full papers published in connection with ETCEI 2024 have been assigned separate DOIs to ensure persistent identification and citation.

ETCEI 2024 was organized under the scientific leadership of the Conference Co-Chairs, Assoc. Prof. Nestor Evmorfopoulos (Electrical and Computer Engineering Department, University of Thessaly) and Asst. Prof. Gregory Doumenis (Autonomous Systems Laboratory, Department of Informatics & Telecommunications, University of Ioannina), who coordinated the development of a program designed to strengthen the interface between academic research, technological innovation, and industrial deployment.

The conference took place in a global context defined by accelerated digital transformation, rapid advances in artificial intelligence, renewed strategic emphasis on semiconductor technologies, and increasing demand for resilient, secure, and energy-efficient digital infrastructures. Within this evolving landscape, Edge Intelligence has emerged as a foundational architectural paradigm enabling distributed processing, real-time analytics, and intelligent cyber-physical systems across critical sectors.

All submitted contributions underwent a peer-review process conducted by members of the Technical Program Committee. The overall acceptance rate for full papers was approximately **69%**, reflecting the commitment of ETCEI 2024 to maintaining high scientific standards while supporting emerging research directions. In addition to full papers, selected poster presentations were included to encourage early-stage innovation and interdisciplinary exchange.

The scientific program addressed a broad spectrum of internationally relevant topics, including Edge and Fog Computing applications, Edge AI and machine learning, hardware and circuit design for edge systems, IoT-oriented architectures and software processes, sensor technologies, energy optimization and storage, 5G integration, and simulation and modeling techniques for intelligent distributed infrastructures. The contributions collected in this volume reflect the growing integration of the Greek research ecosystem within the broader European and global technology landscape.

By facilitating interaction among researchers, industry leaders, startups, and policymakers, ETCEI 2024 strengthened collaboration pathways and reinforced its position as an outward-looking platform within the international emerging technology community.

The Editors acknowledge the Organizing Committee of ETCEI 2024 for their dedication and contribution to the successful realization of the conference. Appreciation is extended to the Local Organizing Committee of the University of Thessaly for their operational coordination, as well as to the reviewers for ensuring the scientific rigor and integrity of the proceedings. Special thanks are

extended to the volunteers and supporting staff whose commitment facilitated the smooth conduct of the event. The contribution of Ms. Konstantina Tsiapali (HETiA BoD) to the coordination and preparation of this volume is gratefully acknowledged.

October 17-18, 2024
Volos

Edge Intelligence 2024 Proceedings

On behalf of HETiA
Dr. Emmanouil Zervakis

Scientific Editor
Prof. George Stamoulis

ETCEI 2024:
A Flagship Industry–Academia Platform on Emerging Technologies

The 3rd Emerging Tech Conference – Edge Intelligence (ETCEI 2024) was held on October 17–18, 2024, in Volos, Greece. The conference was co-organized by the Hellenic Emerging Technology Industry Association (HETiA) and the University of Thessaly.

ETCEI 2024 provided a forum for researchers, industry representatives, technology developers, startups, and policymakers to present and discuss developments in Edge Intelligence and related emerging technologies. The program was structured to facilitate the exchange of research results, methodological approaches, and implementation experiences across academic and industrial domains.

The conference included scientific presentations, keynote addresses, invited talks, technical sessions, and moderated discussions. Topics addressed encompassed Edge and Fog Computing applications, Edge AI and machine learning, hardware and circuit design for edge systems, IoT-oriented architectures, energy optimization and storage systems, 5G-enabled integration, and modeling and simulation techniques for distributed intelligent infrastructures.

Aims & Objectives

ETCEI 2024 aimed to foster structured interaction between academia, research institutions and industry through peer-reviewed research contributions and validated technological developments in Edge Intelligence and emerging digital systems.

Particular attention was given to technological maturity, system-level integration, and alignment with contemporary European and international developments in artificial intelligence, semiconductor technologies, and advanced digital infrastructures.

The conference agenda included:

- √ Keynote speeches by invited experts
- √ Presentations of peer-reviewed full papers and accepted posters
- √ Thematic technical sessions and panel discussions
- √ Networking opportunities supporting research–industry collaboration
- √ A Career and Innovation Session facilitating interaction between students, early-stage researchers, and technology-oriented enterprises

Day 1: Keynote Addresses, Technical Sessions and Industry Dialogue

The first day of the 3rd Emerging Tech Conference (ETCEI 2024) brought together academic researchers, industry representatives and public sector stakeholders, establishing the thematic focus of the conference on Artificial Intelligence, Edge Intelligence and industrial innovation.

The opening session included a conference overview delivered by Prof. George Stamoulis and an address by Dr. Manolis Zervakis, President of HETiA.


Figure 1. Prof. George Stamoulis and Dr. Manolis Zervakis during the opening session of ETCEI 2024.

Keynote Address

The day commenced with the keynote speech entitled:

"Supercharging AI for a GenAI World – Infrastructure Architecture", by **Dionysios Diamantopoulos (IBM Research Europe)**.

The keynote addressed infrastructure architecture requirements for large-scale Generative AI systems, focusing on scalability, distributed computing environments and the evolving cloud–edge continuum.

Figure 2. Dionysios Diamantopoulos delivering the keynote address.

Technical Sessions

Four technical sessions (S1.1–S1.4) were conducted throughout the day, covering:

- √ Machine Learning & AI
- √ Internet-of-Things & AI
- √ Field Devices
- √ Innovative Approaches

The sessions included peer-reviewed paper presentations addressing topics such as federated learning, edge AI benchmarking, embedded systems, FPGA implementations, energy harvesting systems, nanosatellite communications, digital twins and 5G optimization.

Figure 3. Paper presentations during the technical sessions.

Workshop

The workshop **"AI and Industrial Applications"** was delivered by Sotiris Bekos (Qubber) and Dimitris Karagounis (Oliveex), presenting applied AI use cases and industrial deployment perspectives

Figure 4 AI and Industrial Applications workshop session at ETCEI 2024.

Invited Speech

An invited talk on industrial optical inspection systems was delivered by **Sotirios Bekos** (Qubber)

The presentation addressed current technologies, market demands and implementation challenges in industrial vision systems.

Figure 5. Sotirios Bekos delivering the invited speech at ETCEI 2024.

Panel Discussion

A panel discussion entitled: **"Using AI to Boost and Bolster the Agrifood Sector"**, brought together:

- √ Yiorgos Bontzios (Agroverse)
- √ George Varvarelis (CNH)
- √ Sotiris Bantas (Centaur)
- √ Dionysis Bochtis (Farm-B)

Moderator: Ioannis Anastasiou, Vice Governor of Digital Governance and Citizen Services for the Region of Thessaly.

The discussion addressed AI-driven transformation in agriculture and food production, including automation, predictive analytics and digital technologies across the agrifood value chain.

Figure 6. Panel discussion on AI applications in the agrifood sector.

Poster Session

The afternoon poster session provided an interactive platform for presenting emerging research results in areas including energy storage systems, 5G security, nanotechnology infrastructure, cloud–edge vision platforms and energy harvesting technologies.

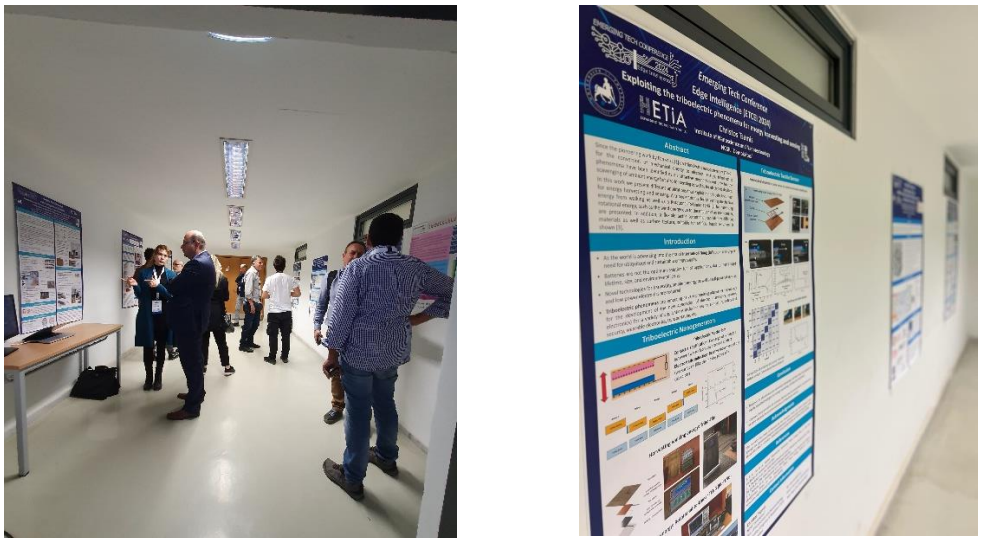


Figure 7. Poster Presentation of ETCEI 2024.

Networking Reception

The first day concluded with a networking cocktail at the 1910 Lifestyle Hotel in Volos, fostering interaction among participants from academia, industry and the public sector.

Figure 8. Networking reception of ETCEI 2024.

Day 2: Advanced Systems, Career & Innovation Session and Industry Dialogue

The second day of ETCEI 2024 focused on advanced electronic systems, semiconductor technologies, artificial intelligence infrastructures and innovation-driven industry engagement.

Technical Session – Circuits, Systems & CAD

The morning technical session (S1.5 – Circuits/Systems & CAD) included peer-reviewed presentations addressing interconnect design and electromagnetic analysis, compact modeling of nanoscale semiconductor devices, model order reduction techniques, RF and mmWave oscillator design methodologies, and automated low-noise amplifier design for wireless sensor nodes.

The session reflected current research activity in circuit design, semiconductor modeling and CAD methodologies for next-generation electronic systems.

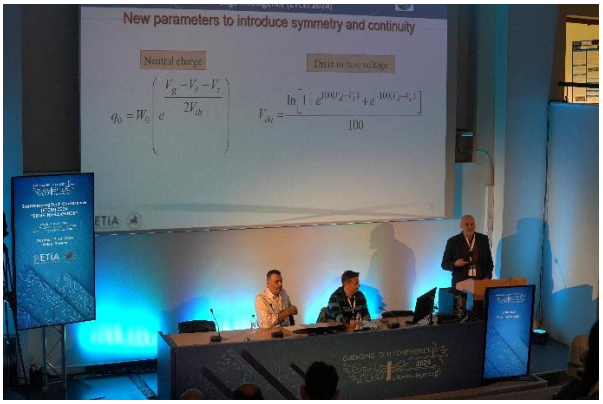


Figure 9. Paper presentations during Session S1.5 – Circuits, Systems & CAD.

Invited Speech – NVIDIA

An invited talk entitled: **"Accelerating Innovation: NVIDIA's Role in Shaping the Future of AI and High-Performance Computing",** was delivered by Paraskevas Bakopoulos, Director of NVIDIA Greece.

The presentation addressed GPU-accelerated computing, AI infrastructure ecosystems and high-performance computing platforms supporting large-scale artificial intelligence applications.

Figure 10. Paraskevas Bakopoulos delivering the invited talk on AI and HPC infrastructures.

Career & Innovation Session

A dedicated Career & Innovation Session brought together representatives from industry and innovation organizations, facilitating interaction between students, early-stage researchers and technology-oriented enterprises.

The session addressed high-specialization skills, recruitment perspectives and emerging opportunities in semiconductor technologies and AI-driven industries. Participating organizations included JOIST Innovation Park, Centaur, KENOTOM, CAPGEMINI and POWER FACTORS.

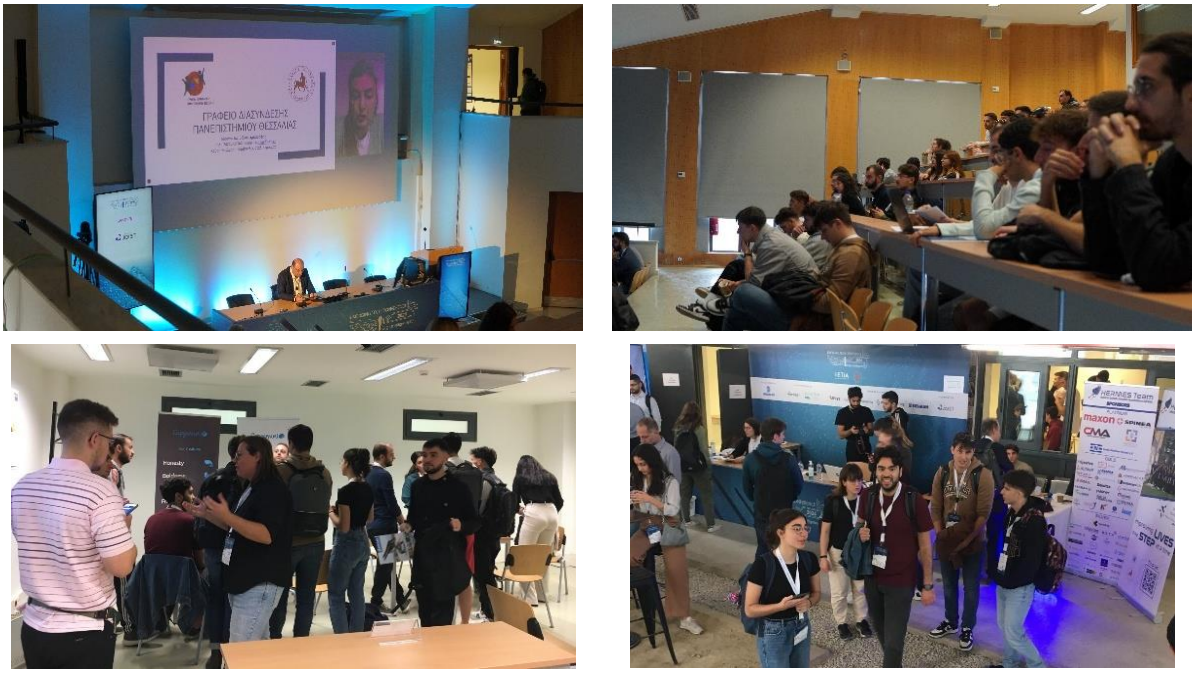


Figure 11. Career & Innovation Session at ETCEI 2024.

Invited Speech – 5G Radios

An invited talk entitled: **“The Adventure of 5G Radios in a Greek Perspective”** was delivered by Manolis Frantzeskakis, CEO of Argosemi.

The presentation examined development challenges and implementation perspectives for 5G radio systems within the Greek innovation ecosystem.

Figure 12. Manolis Frantzeskakis delivering the invited talk on 5G radio systems at ETCEI 2024.

In Conversation – AI in Semiconductor Manufacturing

A discussion session entitled:

“AI Challenges and Opportunities in Semiconductor Manufacturing”, featured Alexander Ypma and Dimitra Gkorou (ASML, The Netherlands).

The session addressed data analytics frameworks and artificial intelligence integration in advanced semiconductor manufacturing environments.

Figure 13. Discussion session on AI challenges and opportunities in semiconductor manufacturing (ASML).

Round Table – Silicon Entrepreneurs: Leveraging the CHIPS Act

A round table discussion entitled:

“Silicon Entrepreneurs: Leveraging the CHIPS Act”

brought together representatives from EDALOGIC, Alma Technologies, KENOTOM and Silicon Highway Technologies, moderated by Dr. Sotiris Bantas (Centaur).

The discussion focused on semiconductor entrepreneurship, ecosystem development and opportunities associated with European CHIPS Act initiatives.

Figure 14. Round table discussion on semiconductor entrepreneurship and CHIPS Act opportunities.

Exhibition Area

Throughout the conference, a dedicated Exhibition Area operated in parallel with the technical sessions, hosting technology exhibitors and interactive demonstrations.

In addition to participating companies and technology providers, two student research teams from the University of Thessaly presented demonstrator systems developed within academic innovation activities. The exhibits included a Formula Student race car prototype and an exoskeleton system, showcasing applied engineering design, embedded systems integration and mechanical–electronic co-development.

The exhibition space provided an opportunity for direct interaction between researchers, students, industry representatives and visitors, reinforcing the practical and application-oriented dimension of ETCEI 2024.

Figure 16. Exhibition Area featuring industry exhibitors and student research teams from the University of Thessaly.

Closing and Outlook

The conference concluded with closing remarks by the co-chairs and a summary report delivered by Prof. George Stamoulis. The announcement of the forthcoming Edge Intelligence 2025 event marked the continuation of the ETCEI initiative.

Figure 15. Closing session of ETCEI 2024.

Conference Co-Chairs

Gregory DOUMENIS	Assistant Professor, Autonomous Systems Laboratory, Department of Informatics and Telecommunications, University of Ioannina
Nestor EVMORFOPOULOS	Associate Professor, Electrical and Computer Engineering Department, University of Thessaly

Program Committee

Christos ANDROULIDAKIS	Member of HETiA Board of Directors, Co-Founder at Alma Technologies
Sotiris BANTAS	Vice President of HETiA Board of Directors, CEO Centaur
Gregory DOUMENIS	Assistant Professor, Autonomous Systems Laboratory, Department of Informatics & Telecommunications, University of Ioannina
Nestor EVMORFOPOULOS	Associate Professor, Electrical and Computer Engineering Department, University of Thessaly
John KIKIDIS	Member of HETiA Board of Directors, Program Manager at Renesas Electronics Corp
Lena KOUGIOUMTZIDOU	Researcher, PhD, University of Thessaly
Fotis PLESSAS	Professor, Electrical & Computer Engineering Department, University of Thessaly
George STAMOULIS	Professor, Electrical & Computer Engineering Department, University of Thessaly
Konstantina TSIAPALI	Member of HETiA Board of Directors, Managing Director at Global Digital Technologies

Scientific Committee

Sotiris BANTAS	HETiA Vice President, Founder & CEO, Centaur
Michael BIRBAS	Professor, Department of Electrical and Computer Engineering, University of Patras
Matthias BUCHER	Professor, School of Electrical and Computer Engineering, Technical University of Crete
Spyros DENAZIS	Professor, Department of electrical and Computer Engineering, University of Patras
Giorgos DIMITRAKOPOULOS	Professor, Department of Electrical and Computer Engineering, Democritus University of Thrace

Gregory DOUMENIS	Conference Co-Chair, Assistant Professor, Autonomous Systems Laboratory, Department of Informatics and Telecommunications, University of Ioannina
Nestor EVMORFOPOULOS	Conference Co-Chair, Associate Professor, Electrical and Computer Engineering Department, University of Thessaly
Filippos FARMAKIS	Professor, Department of Electrical and Computer Engineering, University of Thrace
Alkiviadis HATZOPOULOS	Professor, School of Electrical and Computer Engineering, Aristotle University of Thessaloniki
Athanasios KAKAROUNTAS	Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly
Vasileios KALENTERIDIS	PhD, Senior Analog IC Designer, Thess IC
Petros KARVELIS	Assistant Professor, Department of Informatics and Telecommunications, University of Ioannina
Paris KITSOS	Professor, Electrical & Computer Engineering Department, University of the Peloponnese
Michail KIZIROGLOU	Associate Professor on Microelectronics and Microsystems Technology, Department of Industrial Engineering and Management Alexander Campus, International Hellenic University
Vasileios KONSTANTAKOS	Professor, Department of Physics, Aristotle University of Thessaloniki
Eleni KOUGIOUMTZIDOU	Researcher, PhD, University of Thessaly
Dimitris MOURTZIS	Professor, Department of Mechanical Engineering and Aeronautics, University of Patras
Thomas NOULIS	Professor, Department of Physics, Aristotle University of Thessaloniki
Dionysios REISIS	Professor, Department of Physics – Section of Electronic Physics and Systems, National & Kapodistrian University of Athens
Fotis PLESSAS	Professor, Electrical & Computer Engineering Department, University of Thessaly
Stilianos SISKOS	Professor, Department of Physics, Aristotle University of Thessaloniki
Christos SOTIRIOU	, Professor, EECE Department, University of Thessaly
Dimitrios SOUDRIS	Professor, School of Electrical and Computer Engineering, National Technical University of Athens
Nikolaos Alexandros TATLAS	Associate Professor, Department of Electrical and Electronics Engineering, University of West Attica
Christos THEOCHARATOS	PhD, COO/CIO & Co-Founder Irida Labs

George TSIATOUHAS	Professor, Department of Computer Science and Engineering, University of Ioannina
Andreas TSORMPATZOGLOU	Professor, Edge solutions towards 5G integration / Circuits and systems for Edge applications, University Of Ioannina
Fotis VARTZIOTIS	Circuits and systems for Edge applications / Hardware design and prototyping for Edge and Fog Computing for IoT, University of Ioannina


EMERGING TECH CONFERENCE
2024
Edge Intelligence
GOLD SPONSOR
HERACLES
SILVER SPONSORS
Centaur
KENOTOM
EMBEDDED ENGINEERING EXCELLENCE
BRONZE SPONSORS
Ansys
Capgemini engineering
POWERFACTORS
SIELMAN
SUPPORTER
INNOVATION PARK
JOIST

Conference Program

3rd Emerging Tech Conference (ETCEI) – 2024

Program

Thursday 17/10/2024 – DAY 1

09:00 – 09:30	Registrations + Welcome coffee		
09:30 – 10:00	Opening remarks **Prof. Charalampos BILLINIS** | Rector, University of Thessaly **Prof. George STAMOULIS** (Electrical & Computer Engineering Dep, University of Thessaly) | Conference Overview **Dr. Manolis ZERVAKIS** | President Hellenic Emerging Technologies Association *Master of Ceremony*: ***Prof. STAMOULIS***		
10:00 – 10:40	Keynote Speech: **Supercharging IA* for a GenAI World** *Infrastructure Architecture **Dionysios DIAMANTOPOULOS** | Staff Research Scientist, IBM Research Europe		
10:40 – 11:50	Paper Presentations: **S1.1 | Machine Learning & AI** Session Chairs: **Paris KITSOS & Chris THEOCHARATOS**		
	10:40 - 10:55	Session Presentation - **Georgios FLAMIS**	
	10:55 – 11:05	21	Benchmarking EdgeAI platforms using Arduino Nano 33 BLE *Rohitashva Singh Jhala and Jan Cordes*
	11:05 – 11:15	31	Personalization in Distributed tinyML Applications via Adaptive Clustered Federated Learning *Alexios Filippakopoulos, Dimitris Kastaniotis, Giorgos Andrianakos and Christos Theocharatos*
	11:15 – 11:25	15	Leveraging Empirical Mode Decomposition for Time Series Anomaly Detection in IoT and Smart City Applications *Ioannis Stasinopoulos and Ilias Theodorakopoulos*
	11:25 – 11:35	14	Edge-Optimized NILM: Combining Structured Pruning and Quantization for Energy Disaggregation *Sotirios Athanasoulias, Dimitris Karagkounis, George Samaras and Isidoros Kokos*
	11:35 – 11:45	32	Federated Learning for Workload Forecasting as enabler for Network Service Management *Dimitrios Brodimas, Dimitrios Kapolos, Eleftherios Mylonas, Alexios Birbas and Michael Birbas*
	11:45 - 11:50	Closing Remarks	
11:50 – 12:20	WORKSHOP: **AI and Industrial Applications** **Sotiris BEKOS** | Qubber, **Dimitris GARAGOUNIS** | Oliveex *Moderator:* ***Dimitrios KAREKLIDIS,*** *Journalist*		
12:20 – 13:05	Paper Presentations: **S1.2 | Internet-of-Things & AI** Session Chairs: **Gregory DOUMENIS & Sotiris BANTAS**		
	12:20 - 12:30	28	Short-term weather forecasting in maritime environments exploiting a ML-based model at the Edge *Vasiliki Naskari, Ioannis Masklavanos, Stavros Adam and Gregory Doumenis*
	12:30 - 12:40	24	Edge deep learning for low capabilities devices *Fotis Filippou, Fotis Foukalas and Theodoros Tsiftsis*
	12:40 - 12:50	7	Comprehensive comparison of YOLO-based object detection in edge applications: A use case on free parking spots *Vasilios Karvelas and Christoforos Kachris*
	12:50 - 13:00	2	Computing the Cleanness of the Photovoltaic (PV) Panels *Alexandros Tsagkaropoulos, Christoforos Vasilakis, Georgios Venitourakis, Tzouma Amrou, Georgios Konstantoulakis, Panagiotis Golemis and Dionysios Reisis*
	13:00 - 13:05	Closing Remarks	
13:05 – 14:00	Lunch Break		

14:00 – 14:20 Invited Speech: **Optical inspection systems in Industry– Current technologies, Market demands, Challenges**
Sotirios BEKOS | Managing Director at QUBBER

14:20 – 15:30 Paper Presentations: **S1.3 | Field Devices**
Session Chairs: **Thomas NOULIS & Fotis PLESSAS**

14:20 - 14:35		Session Presentation - **George KERAMIDAS**
14:35 - 14:45	33	Using AMR Sensors for Surface Residual Stress Determination in Steel Constructs *Spyros Angelopoulos, Gregory Doumenis, Aphrodite Ktena and Evangelos Hristoforou*
14:45 - 14:55	3	A novel approach on continual operation of compromised ECU functions: REWIRE Perspective *Athanasios Athanasiadis, Christoforos Koutsianoudis, Konstantinos Lamaris and Tilemachos Matiakis*
14:55 - 15:05	5	A Low Power FPGA Implementation of LDPC Encoder for Space Applications *Christos Sidiras and Vasilis Pavlidis*
15:05 - 15:15	8	Evaluation and DSP Benchmarking of the European Radiation-Hardened NG-ULTRA FPGA *Anastasios Xynos, George Lentaris and Dimitrios Soudris*
15:15 - 15:25	26	Autonomous Multi Source Energy Harvesting Multi-Sensor *Ioannis Masklavanos, Theodoros Georgiadis, Vasiliki Naskari, Fotios Vartziotis and Gregory Doumenis*
15:25 - 15:30		Closing Remarks

15:30 – 16:00 Panel Discussions: **Using AI to boost and bolster the agrifood sector**
Yiorgos BONTZIOS | Agroverse**, George VARVARELIS** | CNH**, Sotiris BANTAS** | Centaur, **Dionysis BOCHTIS** | Farm- B
*Moderator: **Ioannis Anastasiou,** Vice Governor of Digital Governance and Citizen Services for the Region of Thessaly*

16:00 – 16:45 **Poster Presentations**

11	Frequency and Procedure of Formative Assessment: Impact on Student Experience *Iris Gertner*
17	Towards a vision system for olive fruit selection for enhanced olive oil production *Dimitrios Kosmopoulos, Erion-Vasilis Pikoulis and Kostas Blekos*
18	Preliminary Hazard Analysis and Functional Safety Concept for a 2nd Life Battery Energy Storage System *Rigas Frangoudis, Dimitra Spanoudaki, Sotiris Athanasiou and Elisavet Elvanoglou*
22	SAND5G - Security Assessments for Networks and Services in 5G Networks: From 5G to Edge *Kostas Lampropoulos, Kostas Pournaras, Christos Tranoris, Odysseas Koufopavlou, Spyros Denazis and Paris Kitsos*
25	Nanotechnology processes with atomic-scale precision: Access to new tools and clean-room infrastructure for researchers and the Industry *Chloi Zormpa, Sotiris Mouchtouris, Vasilios Vamvakas and Evangelos Gogolides*
34	PerCV.ai Platform: Leveraging Cloud and Edge Computing to Build Vision AI Solutions at Scale *Thomas Charisis, Dimitris Kastaniotis and Christos Theocharatos*
35	Exploiting the triboelectric phenomena for energy harvesting and sensing *Christos Tsamis*

16:45 – 17:50 Paper Presentations: **S1.4 | Innovative Approaches**
Session Chair: **Gregory DOUMENIS**

16:45 - 17:00		Session Presentation – **Evangelos HRISTOFOROU**
17:00- 17:10	27	Charting A New Course for Food Logistics: Enhancing River and Ocean Transport Efficiency with CFD-Based Digital Twins, with examples from soybean logistics in the Amazon River *Sotiris Bantas, Miltiadis Xynopoulos, Alexandros Dimitriou, Antonis Sioutas, and Efstathios Kaloudis*
17:10 - 17:20	9	Review of Novel Battery Balancing Topologies *Evi Keramida, Sotirios Athanasiou and Fotis Plessas*
17:20 - 17:30	20	BoM Reduction for the 5G Radio Unit *Kostas Vryssas and Kostas Papathanasiou*
17:30 - 17:40	29	Development and validation of the AcubeSAT nanosatellite communications module *Georgios Kikas, Christina Athanasiadou, Ilias Kamoisis, Ioannis Dimoulios and Alkis Hatzopoulos*
17:40 - 17:50	10	Computer vision-based identification of motorcycle helmets using a TPU-based platform *Sotiris Michalakeas and Christoforos Kachris*

17:50 – 18:10 Invited Speech: **Dimitrios MOURTZIS** | Professor, Department of Mechanical Engineering and Aeronautics, University of Patras

18:10 - 18:20 Closing Remarks for Day 1 Presentations **| Gregory DOUMENIS |** Co-chair

19:00– 22:00 **Networking Cocktail** at the **1910 Lifestyle Hotel |** *Dimitriados 25 & 54 Sintagmatos, Volos, Greece*

END of DAY 1

EMERGING TECH CONFERENCE 2024 Edge Intelligence

Friday 18/10/2024 – DAY 2

09:00 – 09:30 Registrations + Welcome coffee

09:30 – 10:20 Paper Presentations: **S1.5** | **Circuits/Systems & CAD**

Session Chairs: **Nestor EVMORFOPOULOS & Christos SOTIRIOU**

09:30 – 09:40	6	Design and Electromagnetic Analysis of Interconnects in Silicon Interposer *Nefeli Metallidou, Thorsten Baumheinrich and Vasilis Pavlidis*
09:40– 09:50	19	Continuous and Symmetric Drain Current Compact Model for Nanoscale Triple-Gate FinFETs *Miltiadis K. Nakos, Andreas Tsormpatzoglou, Theodoros A Oproglidis, Dimitrios H. Tassis and Charalabos A. Dimitriadis*
09:50 – 10:00	30	A low-rank balanced truncation approach for large-scale RLCk model order reduction based on extended Krylov subspace and a frequency-aware convergence criterion *Christos Giamouzis, Dimitrios Garyfallou, Nestor Evmorfopoulos and Georgios Stamoulis*
10:00 – 10:10	13	An Efficient Design Methodology for RF and mmWave Voltage – Controlled Oscillators *Alexandros Chatzis, Anastasios Michailidis, Vasiliki Gogolou and Thomas Noulis*
10:10 – 10:20	12	Automated Low-Noise Amplifier Design Methodology for Wireless Sensor Nodes *Panagiota Tsimpou, Anastasios Michailidis, Vasiliki Gogolou and Thomas Noulis*

10:20 – 10:40 Invited Speech: **Accelerating Innovation: NVIDIA's Role in Shaping the Future of AI and High-Performance Computing**
Paraskevas BAKOPOULOS | Director of NVIDIA Greece

10:40 - 12:10 **Career Day**

10:40 – 10:50	**Liaison Office, University of Thessaly**	Chrissa Voulgari Career Counselor, Career Office - University of Thessaly
10:50 – 11:20	**Mastering the job market: High-specialization skills in emerging tech for successful HR interviews** **Angeliki Mantse,** HR Manager & **Andreas Almpanis,** Project and Networking Manager	JOIST Innovation Park
11:20 – 11:35	**CENTAUR**	Faye Beligianni, Global customer support manager
11:35 – 11:50	**KENOTOM**	Tilemachos Matiakis, Co-Founder & Technical Director
11:50 – 12:00	**CAPGEMINI**	Xanthos Vlachos, Senior Electronics Semiconductor Engineer Georgia Psarra, HR
12:00 – 12:10	**POWER FACTORS**	Eftichia Goula, Talent Acquisition Specialist Giota Karagianni, Talent Acquisition Specialist

---

12:10 - 12:30 **Students Group Participants**

12:10– 12:20	**CENTAURUS RACING TEAM**	John Athanasiadis, Head of High voltage Department & Konstantinos Bakiris, Member of High voltage Department
12:20 – 12:30	**HERMES Team**	Christos Belogiannis, Team Manager

---

12:30 – 12:45 Coffee Break

12:45 – 13:05 Invited Speech: **The Adventure of 5G Radios in a Greek Perspective**
Manolis FRANTZESKAKIS | CEO Argosemi

13:05 – 13:35 **In conversation** **AI challenges and opportunities in semiconductor manufacturing**
Alexander YPMA & Dimitra GKOROU | Data & Analytics Line Manager at ASML, The Netherlands

---

13:35 – 14:15 Round table **Silicon entrepreneurs: Leveraging CHIPS Act**

Dimitrios GARYFALLOU | EDALOGIC, **Christos ANDROULIDAKIS** | Alma Technologies, **Tilemachos MATIAKIS** | KENOTOM,
Nikos SKETOPOULOS | Silicon Highway Technologies

Moderator: ***Dr. Sotiris BANTAS,*** *CEO Centaur*

14:15 - 14:25 Closing Remarks for Day 2 Presentations | **Nestor EVMORFOPOULOS | Co-chair**

14:25 – 14:35 **Conference Summary Report**
Prof. George STAMOULIS

Edge Intelligence 2025 announcement
Dr. Emmanuel ZERVAKIS

14:35 – 15:30 **Late Lunch**

END of CONFERENCE

EMERGING TECH CONFERENCE 2024 Edge Intelligence

Contents

EMERGING TECH CONFERENCE 2024
Edge Intelligence

Papers

Session 1.1 | Machine Learning & AI

Session Chairs: **Paris KITSOS & Chris THEOCHARATOS**

Session Presentation - **Georgios FLAMIS**

Benchmarking EdgeAI platforms using Arduino Nano 33 BLE

Rohitashva Singh Jhala and Jan Cordes

Personalization in Distributed tinyML Applications via Adaptive Clustered Federated Learning

Alexios Filippakopoulos, Dimitris Kastaniotis, Giorgos Andrianakos and Christos Theocharatos

Leveraging Empirical Mode Decomposition for Time Series Anomaly Detection in IoT and Smart City Applications

Ioannis Stasinopoulos and Ilias Theodorakopoulos

Edge-Optimized NILM: Combining Structured Pruning and Quantization for Energy Disaggregation

Sotirios Athanasoulias, Dimitris Karagkounis, George Samaras and Isidoros Kokos

Federated Learning for Workload Forecasting as enabler for Network Service Management

Dimitrios Brodimas, Dimitrios Kapolos, Eleftherios Mylonas, Alexios Birbas and Michael Birbas

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 1-9

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Benchmarking EdgeAI platforms using Arduino Nano 33 BLE

Rohitashva S. Jhala and Jan Cordes

[1] *University of Oldenburg, Oldenburg, Lower Saxony, Germany*
rohitashva.jhala@uni-oldenburg.de
[2] *Deutsches Zentrum fur Luft- und Raumfahrt (DLR), Oldenburg, Lower Saxony, Germany*
jan.cordes@dlr.de

Abstract

Edge AI enables machine learning models to operate directly on resource-constrained devices such as microcontrollers, allowing for real-time data processing and enhanced data privacy. This paper benchmarks multiple Edge AI frameworks, the evaluation spans five key metrics: accuracy, memory footprint, inference time, power consumption and usability. The platforms are tested using a human-motion classification dataset and deployed on an Arduino Nano 33 BLE board. Results indicate significant variations in framework performance across these metrics, providing insights into selecting the most suitable frameworks based on specific application requirements. This research highlights the strengths and weaknesses of each framework, helping developers make more informed decisions when choosing Edge AI solutions.

1 Introduction

In an increasingly interconnected world, the ability to process information at the source, namely, the device itself has become a critical challenge [13]. The concept of Edge AI is to decentralize data processing and machine learning from cloud-based infrastructures and high-performance computers to the very source, where the data is actually generated i.e. the sensor module itself [14]. This has the potential to extend the scope of Artificial Intelligence (AI) applications to areas previously considered limited, from wearable health monitors to smart sensors in remote locations and farming, the implications for daily life are vast and varied. These devices typically have limited computational resources and are under significant constraints in terms of memory and power supply as they are usually running on batteries or power harvesting.

Different frameworks have come up over the last years to make developing such Edge AI devices easier. They typically include the process of development, training, deployment, and optimization of machine learning models on these edge devices. In this work, we are comparing four Edge AI frameworks in five domains: accuracy, memory footprint, inference time, power consumption and usability. We therefore set up all four frameworks with an identical dataset and nearly identical configurations for the model to be generated. As a target edge device, we chose the Arduino Nano 33 BLE board because of its widespread adoption and inbuilt inertial measurement unit (IMU) sensor [1].

This work is structured in the sort review of related work following the explanation of the concept and

the methodology to establish comparability between the frameworks. In the Concept chapter also the dataset used and the common data preprocessing as well as the gathering of metrics to compare the frameworks are explained. The implementation chapter describes how the different frameworks have to be set up and what hurdles have to be overcome to make them comparable. The Evaluation chapter gives the results of the measured metrics and gives some additional reasoning for the determined usability score. Finally, in the Conclusion we sum up our results and give some recommendations on which framework is suitable for which use case.

2 Related Work

Osman et al. [11] did some similar work by comparing two popular frameworks: TensorFlow Lite Micro and CUBE AI. The aim was to compare the two for different TinyML applications and determine their use cases. They concluded CUBE AI is better suited for memory-limited and power-intensive TinyML applications, while TFLM offers wide availability and support for devices. Similarly, there is a need for comparing/benchmarking various other frameworks too to assess a suitable fit w.r.t. the applications as mentioned by Osman et al. in the future work.

Likewise, the article by Pratim Ray [12] presented current understanding of TinyML, the existing toolset, key enablers of the TinyML paradigm, state-of-the-art frameworks use cases and challenges. They too emphasized the need for more researchers to come forward, and develop new benchmarks with new datasets to test these frameworks and understand the use cases better.

Although Osman and Ray's work is in the field of EdgeAI, our study differentiates itself in several key ways. Unlike Osman's work which compares only two EdgeAI frameworks with both tested on different hardware platforms lacking a common reference point for benchmarking, we created a standardized benchmarking approach by using a common hardware platform, ensuring a consistent reference point for all frameworks tested. In contrast to Ray's work which is more focused on theoretical enquiry and challenges instead of practical benchmarking of frameworks with an edge device, we combine both the theoretical and practical aspects. We use the System Usability Scale (SUS) [9] based on established usability principles to evaluate the usability of the frameworks, and also benchmark them across multiple metrics like accuracy, memory consumption, inference time, and power consumption for comparison.

3 Concept

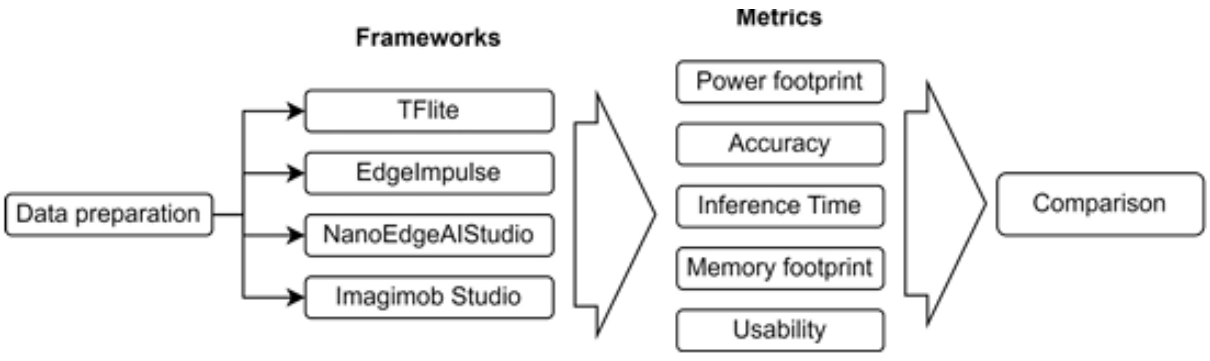


Figure 1: Concept behind the benchmarking of the frameworks.

Figure 1 shows the workflow and methodology underpinning the research. The data is used by the frameworks to train the model and once trained, the comparison is laid out across the array of metrics. In our case, the work begins with data processing where the dataset is divided into testing and training sets. Following this, the sliding window technique is employed to transform the raw inertial measurement unit (IMU) data into a series of windows [8]. Each window represents a single sample on which the neural network will be trained. The trained model is then deployed onto the Arduino Nano 33 BLE board using the Arduino integrated development environment (IDE). The inference process also involves configuring the necessary input and output functionalities to enable the BLE board to perform real-time inference. The input functionality captures new data from the sensors, while the output functionality manages the prediction results generated by the model. Accuracy and memory consumption are measured before the deployment while the inference time, current consumption, and energy are measured after the deployment. While working with the frameworks we score their usability on a system usability scale so a comparison between the usability of the frameworks can be drawn.

3.1. Dataset and Data Preprocessing

As a typical dataset for this study we used the MotionSense Dataset [3], consisting of vibration data of human motion classification recorded by an IMU. The data, recorded at 50 Hz, offers 6 different activities: Walking, Sitting, Jogging, Standing, Upstairs, and Downstairs.

The data consists of time-series signals that capture the dynamic movements of the human body along three spatial axes (x, y, z) at 50 Hz. When trying to classify human activity, such as walking, jogging, sitting etc., it's beneficial to look at segments of data over a period of time rather than individual data points. For this, we use the sliding window technique where the data is divided into smaller segments called windows. By segmenting data into overlapping windows, we can capture the temporal dependencies which are crucial in sequential data [2]. In our work, we use the window length of 2.5 seconds or 125 data points as it has proven to be reliable [2]. Before creating sliding windows, each file containing the data is first divided into test and train sets and then windows are generated for each set. This methodology also ensures that there is no overlapping of the data between test and training sets.

3.2. Metrics

We evaluated various Edge AI frameworks by assessing their performance across six critical metrics: energy consumption, inference time, model accuracy, current consumption, memory usage, and usability.

To measure the **current consumption**, we use a digital multimeter to record the current observed when the model is making predictions (active) and when the device is idle.

To measure **inference time**, we recorded the execution period per classification i.e. from the moment an input is provided to the model until an output is received. This is done by using the Profiler library [4] which will output the time taken in milliseconds for one classification.

The **energy consumption** for every classification can simply be calculated by multiplying the current consumption value and inference value of the respective framework with the voltage, which in our case is 3.3 volts

Accuracy measures the model's performance in correctly predicting the labels compared to the true labels. To measure accuracy, we consider the accuracy results were obtained by validating the trained model created within the corresponding framework using a separate test dataset.

Whereas for memory consumption, the Arduino IDE itself provides the information on the space occupied by the files used for performing inference. During the sketch upload process, the console output of the application reports the amount of flash memory and the RAM used when uploading the files to the Arduino device.

In our evaluation, we also wanted to assess the usability of each framework based on a set of established usability principles. For that, we chose the System Usability Scale (SUS) to incorporate a score-based assessment. SUS is a 10-question 5-response Likert scale which we used for a subjective assessment of our opinions about the usability of these frameworks. These questions ask the user if the given framework can fulfil each aspect of the usability principles. Wherein, the response options vary from "Strongly Disagree" to "Strongly Agree", with 0 = Strongly Disagree to 4 = Strongly Agree. To convert the responses into a range of 0 to 100, for each framework, the responses were summed and multiplied by 2.5. This resulted in overall usability score of that framework.

Specific aspects of the usability principles to be examined include: **User Control and Freedom** [10] which includes the ability to configure the training process, neural network architecture, data preprocessing and resolution, **Clarity** [10] which includes clear description of the functionalities, **Help and Documentation** [5] including diversity in tutorials/examples and response time of community forum, and **Compatibility** [7] for the hardware device.

4 Implementation

Some frameworks lack the option for full control over the network design. Consequently, we sometimes had to opt for different network designs, for example: in the case of NanoEdge AI Studio as discussed in the section 4.2.3. We tried to keep these changes minimal and stick to a standard design. For that, we set up a baseline neural network architecture with Keras and TensorFlow. The deployment methodologies for the trained network also differ among the frameworks. Some, like Edge Impulse, produce a ready-to-use package library, while others generate C++ source code that we then add to an Arduino project. This code then must follow specific guidelines to interface the model on the hardware board.

4.1. Baseline architecture of Neural Network

In order to best compare the frameworks, a fixed architecture of a neural network must be followed across all the frameworks which we call the baseline model. This model serves as a foundational template ensuring consistency and comparability.

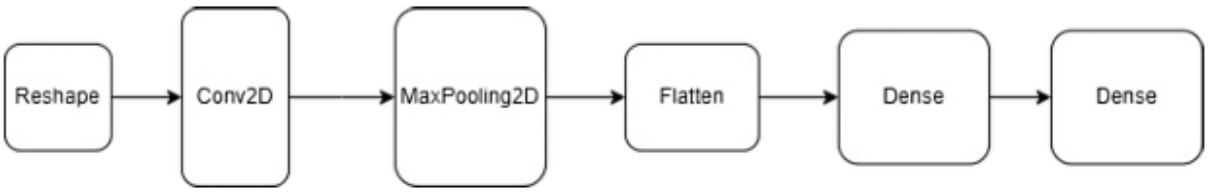


Figure 2: The baseline architecture of the Keras model

Figure 2 shows the architecture of the baseline model which was built using the Keras API in TensorFlow. The model is designed to operate on input data shaped as (125, 3), where 125 represents the number of time steps in a sliding window, and 3 corresponds to the three axes (X, Y, Z) of the IMU data.

This architecture and configuration of the neural network is the baseline model for every framework to follow.

4.2. Frameworks

4.2.1 TensorFlow Lite for Microcontrollers

TFLM being an open-source library with Python as an interfacing language, had the capability to follow the same neural network architecture design as mentioned in the section 4.1. Once the model was trained over the data, it was validated using the test dataset. Quantization is possible in TLFM, so the model was quantized using full integer quantization i.e. by converting the weights and activations from floating-point to 8-bit integer values. After this, the quantized model was converted into the .tflite format using TensorFlow's TFLite Converter. We used the "Hello World" example from TensorFlow Lite as a reference to incorporate the IO functionalities. The .tflite model then was converted into a C array with the xdd tool[2] to be used for performing inference. Once the C file of the model was created, we included it in our Arduino program for deployment on the device.

4.2.2 Edge Impulse

Edge Impulse is a web-based developing platform that provides the interface to upload the data, build and train the model, and deploy it onto an edge device. Edge Impulse follows a block-based design, beginning with uploading the dataset and splitting it randomly between test and training sets. After this, the sliding windows are generated and a classification model is chosen in the learning block with six features representing the categories of the human motion dataset. Within the preprocessing block, we opted for the raw data option, choosing to process the data without any additional preprocessing steps following the guidelines from the baseline architecture. After training and testing in the framework, the trained model is quantized (int8) converting the weights and activations to 8-bit integer values. The framework built the quantized model into an Arduino library, which is then downloaded as a zip file. Once unzipped, It contains the trained model's parameters and an Arduino program for our BLE board, which can directly be deployed.

4.2.3 NanoEdge AI Studio

NanoEdge AI Studio is a free software developed by STMicroelectronics for Windows and Linux platforms. We fed the training dataset in the form of sliding windows as the Studio itself does not provide a tool for creating these. Once imported, the benchmark section runs an array of machine learning models online from its servers on the dataset to find the best performing one. The selection is done based on accuracy and memory consumption. Hence, the Studio does not provide any tool to access or alter the architecture of the model and the configurations of the training process. In our case, out of the best-performing models we chose Multi-Layer Perceptron (MLP) due to its similar

[2] xxd-i quantized TFmodel.tflite > model.cc

architecture of interconnected multilayered neurons. Following the benchmark, the model was tested with the test dataset to ensure accuracy and then packed into a zip file for deployment. The zip file contains the Arduino program and the C source files. These C files contain the model's weights, function declarations, enumerations, macros, and other necessary definitions required for proper model integration. The main Arduino program includes the generated C source files as headers. Additionally, we add input/output functionalities to the Arduino program for sensor reading, pre-processing inputs, and printing the outputs of the prediction.

4.2.4 Imagimob Studio

Imagimob Studio is a free software for the Windows platform where the training of the models is done over the cloud. The framework accepts the data in a very specific format [6] with two CSV f iles, one for the sensor data and the other for the label containing the "start time", "end time", and the "label name". This requires further pre-processing of the dataset. After uploading, the data is split into training, testing, and validation sets. For the model architecture and training configurations, we followed the same as specified in the baseline architecture 4.1. Imagimob performs the training process on its cloud servers which was speeding up the training process. Once trained, the Studio runs the model on validation and test sets and displays the accuracy and F1 scores. Following this, the trained model was converted into C/C++ source files using the inbuilt code generation tool. To deploy the generated files on the Nano BLE, we built an Arduino program performing IO functionalities for the model, similar to that in the NanoEdge AI Studio.

5 Evaluation

	Edge-Impulse (Enterprise)	TFLite for Microcontrollers	NanoEdge AI Studio	Imagimob Studio
Accuracy	88.58%	97.35%	94.80%	96.52%
Memory Usage-RAM (KB)	50.41	70.82	48.40	96.32
Memory Usage-Flash (KB)	156.15	359.20	128.21	845.85
Inference Time (ms)	21	20	36	56
Current Consumption (mA)	7.95	7.98	8.03	8.01
Energy Consumption (mJ)	0.50	0.47	0.86	1.34
Usability[3]	95	85	37.5	47.5

Table 1: Resulting metrics from the implementation of the frameworks.

5.1. Quantitative Results

As shown in the table 1, across the six key metrics, TFLM performs the best in terms of accuracy, inference time (per classification), and energy consumption (per classification). Although there is a

[3] The scores are derived from a system usability scale explained in detail in the section 3.2

similar current consumption figure across all the frameworks, the different inference times result in different figures for the energy consumption for a single classification. Hence, TFLM consumes the least energy per classification. As for the memory consumption, NanoEdge AI Studio utilizes the least Flash and RAM as compared to the rest of the frameworks.

5.2. Experimental results of Usability

In terms of usability scores as per system usability scale, as shown in the table 1, NanoEdge AI Studio follows the least of the usability principles mentioned in the section 3.2 while Edge Impulse adheres the most to these principles.

TensorFlow Lite for Microcontrollers, in terms of User Control and Freedom, offers complete control to configure the training process, neural network architecture, and data pre processing. With the support of numerous Python IDEs like VSCode, PyCharm, and Jupyter Notebook, the resolution of the application is highly configurable. It offers a wide range of tutorials and examples with a large and active community forum. It also offers support for a wide array of edge devices. However, in terms of clarity, the text describing the features and the deployment process is quite complex to understand at first.

Edge Impulse too provides control over the training process and the neural network archi tecture. However, it can only be accessed through Keras Expert Mode and is dependent on the output from its preceding layers. For data preprocessing, the framework provides an extensive array of tools and options like CSV Wizard. In terms of clarity, it offers a clear description of the features and options provided with the step-by-step navigation through the application interface. It also offers a wide range of examples/tutorials with deployment support for numer ous edge devices. It has a large and active community offering a response time of 2 to 3 days, as observed.

NanoEdge AI Studio does not provide control over the training process and neither does it over the neural network architecture. It also does not support generating sliding windows from the dataset and does not allow any change in the resolution of the application. NanoEdge AI Studio offers limited clarity about the features and options, particularly in the benchmark section where no information on the configuration of the library is provided. It offers limited tutorials and no support was observed from the community on the posted queries. However, the range of officially supported devices is quite extensive and varied.

Imagimob Studio provides control to configure the training process, neural network architecture and generating sliding windows. The resolution of the application is configurable and can be resized according to preferences. Imagimob Studio offers limited clarity about the features and the options available, particularly for data upload and management, where the procedure to select the data and label files separately is not clearly described. However, the processes after the data upload are adequately explained. Imagimob Studio also offers limited tutorials and no support was observed from the community on the posted queries. It offers a wide range of supported edge devices.

6 Conclusion

The four frameworks were evaluated for bringing the neural networks on resource-constrained microcontroller devices, in our case Arduino Nano 33 BLE board. Out of them, TFLM per formed the best in terms of accuracy, inference time, and energy consumption. While NanoEdge AI Studio

showed the minimum overall memory usage, flash and RAM combined. In terms of current consumption, all of the frameworks showed nearly identical figures. However, the energy consumption showed which framework used the least versus the most energy per classification. From the usability perspective, Edge Impulse showed the best result followed by TFLM, Imagimob Studio, and NanoEdge AI Studio. As different frameworks excel at different metrics, the selection of the EdgeAI framework for a specific application should be guided by the priority of the metrics for that application.

7 Future Work

Expanding the range of evaluation metrics by incorporating additional metrics such as F1 scores for each label and performing live classification tests can provide deeper insights into the model's performance before and after deployment. Evaluation of MATLAB and uTensor could not be completed due to version compatibility issues and incomplete support for our BLE hardware. Resolving these challenges can provide more frameworks for comparison.

References

[1] Arduino nano 33 ble docs. https://docs.arduino.cc/hardware/nano-33-ble/. [Accessed 13-08 2024].

[2] Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models- PubMed — pubmed.ncbi.nlm.nih.gov. https://pubmed.ncbi.nlm.nih.gov/36091986/. [Accessed 13-08-2024].

[3] GitHub- mmalekzadeh/motion-sense: MotionSense Dataset for Human Activity and Attribute Recognition (time-series data generated by smartphone's sensors: accelerometer and gyroscope) (PMC Journal) (IoTDI'19) — github.com. https://github.com/mmalekzadeh/motion-sense. [Accessed 13-08-2024].

[4] GitHub- ripred/Profiler: Easily profile your Arduino functions to see how much time they take. The output can be disabled and enabled at runtime. Very lightweight. — github.com. https://github.com/ripred/Profiler. [Accessed 13-08-2024].

[5] Help and Documentation (Usability Heuristic 10) — nngroup.com. https://www.nngroup.com/articles/help-and-documentation/. [Accessed 13-08-2024].

[6] Imagimob Studio documentation — developer.imagimob.com. https://developer.imagimob.com/data-preparation/data-collection/bring-your-own-data. [Accessed 02-09-2024].

[7] ISO 25010 — iso25000.com. https://iso25000.com/index.php/en/iso-25000-standards/iso-25010. [Accessed 13-08-2024].

[8] Sliding Window and Two Pointer Technique — linkedin.com. https://www.linkedin.com/pulse/sliding-window-two-pointer-technique-saiful-islam-rasel. [Accessed 27-08-2024].

[9] Usability form (SUS) for EdgeAI frameworks — docs.google.com. https://docs.google.com/forms/d/e/1FAIpQLScEr2S0LUd9lRLmFrMUZfQYUxy5fCXgV_1o1vdOCAbLDpNaag/viewform?vc=0&c=0&w=1&flr=0&usp=mail_form_link. [Accessed 28-08-2024].

[10] Usability Principles — improvement.stanford.edu. https://improvement.stanford.edu/resources/usability-principles [Accessed 13-08-2024].

[11] Anas Osman, Usman Abid, Luca Gemma, Matteo Perotto, and Davide Brunelli. Tinyml platforms benchmarking, 2021.

[12] Partha Pratim Ray. A review on tinyml: State-of-the-art and prospects. Journal of King Saud University- Computer and Information Sciences, 34(4):1595–1623, 2022.

[13] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3:1–1, 10 2016.

[14] Raghubir Singh and Sukhpal Singh Gill. Edge ai: A survey. Internet of Things and Cyber-Physical Systems, 3:71–92, 2023.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 10-16

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Personalization in Distributed tinyML Applications via Adaptive Clustered Federated Learning

Alexios Filippakopoulos, Dimitris Kastaniotis[4], Giorgos Andrianakos and Christos Theocharatos
Irida Labs, Kastritsiou 4, 26504 Patras, Greece
{afilippak,kastaniotis,andrianakos,htheohar}@iridalabs.com

Abstract

In large-scale non-IID distributed settings, where statistical heterogeneity and task diversity among local distributions is prevalent, and devices want to maximize performance on the tasks they have data for, vanilla federated algorithms frequently fail to converge and generalize across all tasks. This study aims to highlight the need for more sophisticated approaches in such settings, by adopting a clustered and personalized federated learning approach. We empirically justify the choices made when designing the framework, while our findings suggest a significant boost in performance and training stability.

1 Introduction

From smart gadgets to cameras and sensors, IoT devices produce an immense volume of data. Due to the distributed nature of these devices, the data produced is highly decentralized and heterogeneous. Therefore, a need for AI solutions that consider this distributed nature of data, along with the constraints imposed by the IoT setting, while operating in a privacy-preserving manner, becomes increasingly apparent. Federated Learning (FL) is a popular distributed framework that allows multiple edge devices, referred to as clients constituting a network, to collaboratively train a shared global model, capable of inferring from all tasks within that network. Each client trains its model locally, on its own data, sharing only its updated parameters. Training is coordinated by a central entity, referred to as the server, whose role is to broadcast the initial weights within the network, receive the updated parameters of all participants, aggregate the global model (Federated Averaging) and broadcast it within the network (H. Brendan McMahan 2016). This constitutes a complete global training round in a federated setting. As a result, local data remain at the edge, rendering it decentralized, while the parameter-only exchange between clients and the server preserves, to an extent, both user and data privacy. However, real-world applications are defined by statistical heterogeneity among local distributions, diversity in tasks, as well as non-IID network-wide distributions. Due to these inherent problems, FL's convergence, along with the generalization capabilities of the global model, can be severely hindered, resulting in many clients being better off training solely on their local data (Farzin Haddadpour 2019). Therefore, aggregating one general purpose global model may not be sufficient for modeling all clients' distributions in large-scale heterogenous settings.

[4] *Corresponding author

To address these challenges, Personalized Federated Learning (PFL) acknowledges the benefits of training with rather inaccessible data in a privacy-preserving manner via FL and aims to equip each client with a unique and personalized model instance, tailored to its local data and client-specific tasks. The dominant approaches for achieving personalization in a federated setting, include meta learning (Chelsea Finn 2017), multi-task learning (Virginia Smith 2018), model regularization (Filip Hanzely 2020) and model interpolation (Yishay Mansour 2020), which we opt for, where obtaining personalized models is achieved by mixing the global (aggregated) model with each locally updated model. Another approach is Clustered Federated Learning (CFL) (Felix Sattler 2019), that tries to find clients with similar tasks and/or distributions and cluster them into a 'subpopulation' whose members federate explicitly with each other. Throughout this paper, the goal is to showcase the need for more sophisticated federated approaches in large-scale, distributionally heterogeneous, and task-diverse settings, while addressing some common questions that arise when utilizing the forthcoming techniques.

2 Problem Formulation

Let a population of $k \in \{1, \ldots, KK\}$ clients, each maintaining a local distribution D_k, of size $|D_k|$, on domain $\Xi := X \times Y$, where $X \in \mathbb{R}^d$ is the input domain and $Y \in \mathbb{R}^T$ is the label domain. We denote as $\bar{D} = \bigcup_{k=1}^{K} D_k$ the network-wide (global distribution), which corresponds to the joint distribution of all K local distributions D_k.

For any model $h \in \mathcal{H}$, the loss function is defined as $\ell: \mathcal{H} \times \Xi \to \mathbb{R}^+$, while the true and empirical risks for said model h at a specific local distribution D_k are denoted as $L_{Dk}(h) = E_{(x.y)\sim D_k}[\ell(h(x), y)]$ and $L_{Dk}(h)$ respectively. Thus, any client kk training its model (h_k) on its local distribution D_k will obtain the updated model $h_k^* = argmin L_{Dk}(h_k)$. In a federated setting, where clients train in a collaborative manner as previously described, the aim is to obtain one aggregated global model $\bar{h} * = argmin\, \hat{L}_D \left(\sum_{k=1}^{K} \left(\frac{|D_k|}{|D|} \right) h_k \right)$. Similarly, CFL will construct $\mathbb{C}$ clusters of similar clients before training is initiated. Then, for each cluster $cc \in \mathbb{C}$, comprised of $Kc \subset K$ clients, the cluster-level global distribution will be $\bar{D}_c = \bigcup_{k=1}^{K_c} D_k$ and we will obtain an aggregated global model $\bar{h}_c * = argmin\, \hat{L}_{D_c} \left(\sum_{k=1}^{K_c} \left(\frac{|D_k|}{|D_c|} \right) h_k \right)$.

On the other hand, PFL via model mixing constructs the personalized model of a client k with the following convex combination: $h_k * = \lambda_k^* + (1 - \lambda)\bar{h}^*$, where λ is the mixing parameter and thereby $h_k * = argmin\, \hat{L}_{D_k}(\lambda h_k^* + (1 - \lambda)\bar{h}^*)$. The primary issue of this approach stems from the mixing of the global model $\bar{h}^*$ with every client model h_k^*. Specifically, the greater the divergence between $\bar{D}$ and D_k, the more likely it is a suboptimal set of parameters for h_k^* and subsequently a worse fit for D_k to be obtained. As a result, CFL with PFL can be combined to reduce the divergence between the global and some local distributions. By constructing similarity clusters and mixing the cluster-level global models with the corresponding client-side trained models, a superior set of parameters for each goal of maximizing personalization. Hence, a mechanism for identifying and grouping similar clients into clusters, whose joint distribution will diverge less from D_k compared to $\bar{D}$, is critical, since the quality of the obtained parameters is intrinsically linked to the quality of the clusters formed.

3 Methodology

To construct the clusters before the commencement of training, Hierarchical Clustering is utilized while assuming no knowledge of a client's data, tasks or distribution. To quantify the similarities of clients, the following intuition is used. In supervised statistical learning, both empirically and theoretically, it is proven that parameters that are updated based on a fixed distribution are reflective of that distribution. Therefore, after the server transmits the initial weights throughout the network, each client pre-trains its model for some E epochs and broadcasts its updated weights to the server to be used for measuring the similarities. After the server pools all the updated weights, several distance metrics can be used to quantify this weight similarity. Here, cosine similarity is used that appears to deduce the optimal number of clusters more consistently compared to other metrics. The combination of hierarchical clustering using cosine similarity along with a threshold of t = 0.9, provided the best results in most cases; however, this highly depends on problem setting. Regarding hyperparameter, it is advisable to compare the framework's performance with respect to different values for t. As for the rest of the framework, we do not deviate from the standard FL protocol previously described. Following the cluster formation, each client trains its model instance for one epoch on its local data and transmits its updated weights to the server. Then, the server performs the personalized aggregation via model mixing, using only the weights of the clients that correspond to each cluster. Finally, the server transmits the client-specific models within the network, repeating this process until the framework converges. Figure 1 illustrates the described framework in the form of pseudocode.

/* Runs on Server-Side */
for each client $k \in \{1, 2, \dots K\}$ **do**
  Transmit h^0 // Initial weights
  Receive h_k^E // Pre-trained client weights
end for
Calculate clients' **cosine similarities** using h_k^E
Cosine Similarity **Hierarchical Clustering** to form C clusters
for each cluster $c \in \{1, 2, \dots C\}$ **do**
  $h_c^0 \leftarrow \sum_{k=1}^{|c|} \frac{|D_k|}{|D_c|} h_k^E$ // Aggregate cluster-level initial weights
  for each client $k \in c$ **do**
    Transmit h_c^0
  end for
end for
for each global round $t = 0, 1, \dots T-1$ **do**
  for each client $k \in \{1, 2, \dots K\}$ **do**
    Receive h_k^{t+1} // Client's updated weights
  end for
  for each cluster $c \in \{1, 2, \dots C\}$ **do**
    $h_c^{t+1} \leftarrow \sum_{k=1}^{|c|} \frac{|D_k|}{|D_c|} h_k^{t+1}$ // Aggregate cluster-level global mode
    for each client $k \in c$ **do**
      $h_k^{t+1} \leftarrow \lambda h_k^{t+1} + (1-\lambda) h_c^{t+1}$ // Aggregate client-specific mode
      Transmit h_k^{t+1}
    end for
  end for
end for

/* Runs on Client-Side */
$h_k^0 = h^0$ //Initialize local model with received weights
for each pretraining round $e = 0, 1, \dots E-1$ **do**
  $h_k^{e+1} = h_k^e - \eta_e \bar{\nabla} \ell(h_k^e(\mathbf{x}), y)$ // Update local model
end for
Transmit h_k^{e+1} // Pre-trained local weights
Receive h_c^0 // Cluster-level initial weights
$h_k^{e+1} = h_c^0$ // Load cluster-level initial weights
for each global round $t = 0, 1, \dots T-1$ **do**
  $h_k^{t+1} = h_k^t - \eta_t \bar{\nabla} \ell(h_k^t(\mathbf{x}), y)$ // Update local model
  Transmit h_k^{t+1} // The updated local weights
  Receive h_k^{t+1} // The aggregated personalized weights
end for

Figure 1: An overview of our framework in pseudocode.

4 Experiments

Throughout experimentations, the positive effect of combining CFL and PFL in real-world applications is shown, where no prior knowledge of clients' data and tasks is assumed. Task diversity and distribution heterogeneity is present within the federation, and maximizing personalization is the

goal. Throughout our experiments, the following issues are addressed: (1) the number of layers' weights for computing the similarities, (2) the pre-training time for the weights to adequately reflect the underlying distribution and (3) the advantages of combining PFL and CFL with respect to the framework's performance regarding personalization. For the first two tasks, 'similarity experiments' are designed that do not relate with the federated setting, while for the third issue the federated protocol is simulated.

4.1. Model Architecture and Dataset

Regarding model architecture, a CNN is utilized that is comprised of 5 ReLU activated convolutional layers, 2 max pooling layers and 1 fully connected layer, deriving in a total model size of 511 KBs. Regarding the datasets used throughout the experiments, we opted for CIFAR-100 comprised of 100 classes and CIFAR-10 comprised of 10 classes, where each class on both datasets has 600 32x32 RGB images. To study the non-IID setting, we must pathologically create one. We define a parameter $a_1 \in [0,1]$, where for instance $a_1 = 0.8$ means that 80% of the total data of a client are for one class and the rest $1 - a_1 = 20\%$ of data are reserved for other classes. Similarly, we also denote $a_2 \in [0,1]$, where for instance if $a_1 = 0.8$ and $a_2 = 0.6$, this means that 80% of a client's data is reserved for one class, $(1 - a_1) * a_2 = 12\%$ is reserved for the second class and $(1 - a_1) - ((1 - a_1) * a_2) = 8\%$ are reserved for other classes. As a result, we can pathologically create a non IID environment with imbalanced local distributions, that closely resembles a real-world scenario.

4.2. Experimental Setup

For the similarity experiments, CIFAR-10 with 10 clients and 3 tasks per client was used, with $a_1 = 0.6$ and $a_2 = 0.8$. For the federated setting simulation, 16 clients were employed as well as two pathological Non-IID settings, with imbalanced local distributions, using CIFAR-100 classes with $a_1 = 0.6$ and $a_2 = 0.8$, as well as with $a_1 = 0.7$ and $a_2 = 0.8$. Firstly, simulating a scenario where all tasks within a cluster are shared by all its clients, two tasks were assigned per client for a total of 16 tasks with the optimal number of clusters being 8. Conversely, a scenario was simulated in which, within each cluster there are tasks that do not concern all clients. Here, 3 tasks were assigned per client for a total of 20 tasks within the network, with 4 optimal clusters. The implementation was carried out using PyTorch for training and Python's socket interface for networking. Training lasted 500 global training rounds, using cross-entropy loss and stochastic gradient descent with a learning rate of 0.01.

5 Results on Federated Simulation Experiments

During the federated simulation experiments, the following issues were addressed: (1) the way to determine a good value for the mixing parameter λ and (2) an effectiveness comparison of clustered PFL in large-scale heterogeneous settings w.r.t. FL and PFL. Figure 2 illustrates the validation loss per training round, comparing both different mixing values (λ), as well as the effectiveness of clustering. Based on the two graphs, a striking difference in oscillations is observed regarding validation loss w.r.t. clustering. Our analysis outcome is that clustering significantly reduces these oscillations and smoothens the loss's convergence, by preventing the models from drifting too much from their local distributions. Additionally, clustered clients reach lower convergence points, indicating a better fit on the data and highlighting the effectiveness of the method in such settings. Regarding the mixing parameter λ, no specific value outshines regarding validation performance, with all of them converging to roughly the same point.

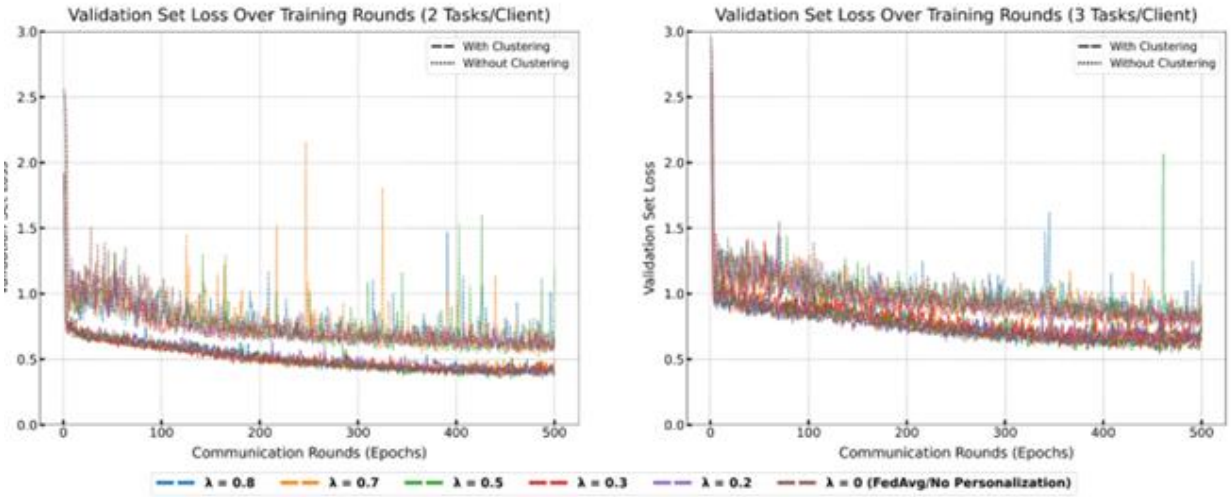


Figure 2. Validation loss per training round, comparing both different mixing values (λ), as well as the effectiveness of clustering.

Figure 3, on the other hand, illustrate the test accuracy per training round, comparing both different mixing values (λ), as well as the effectiveness of clustering. It is obvious that the clustering framework steadily outperforms its counterparts, reaching an average of +12% and +11% in the two graphs respectively, in terms of test set accuracy. Furthermore, the striking difference between FedAvg (no clustering λ=0 or brown dotted) compared to all other approaches, highlights the need for personalization in large-scale heterogeneous embedded settings and the reason that one global model is usually not sufficient. Regardless of clustering or not, no particular λ stands out compared to others, besides FedAvg, which does not personalize and is the worst performing. Interestingly though, on the 2-tasks experiment, the highest recorded precision belongs to FedAvg (see Figure 4). This suggests, that in Non-IID and imbalanced settings, constructing only one global model will also result in a bias towards the majority tasks, leading to fewer predictions for underrepresented tasks, which tend to be correct and hence inflate the model's precision. Lastly, a general rule for choosing an appropriate λ value is the following. As λ increases we incorporate more of the locally trained model in each aggregation and vice versa. Thus, the more we expect a local distribution to diverge from its cluster level or global distribution, the greater λ should be.

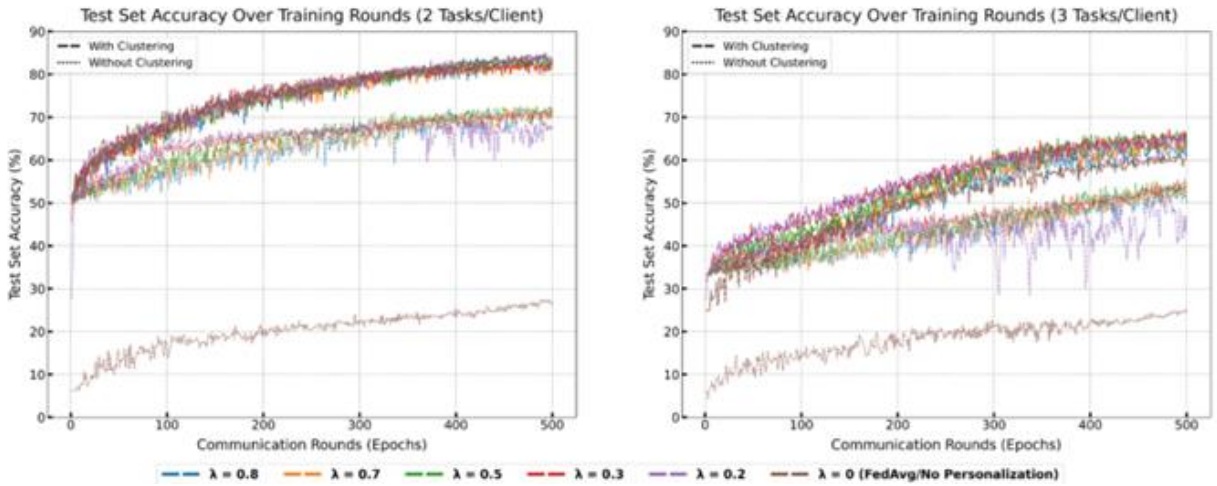


Figure 3. Test accuracy per training round, comparing both different mixing values (λ), as well as the effectiveness of clustering

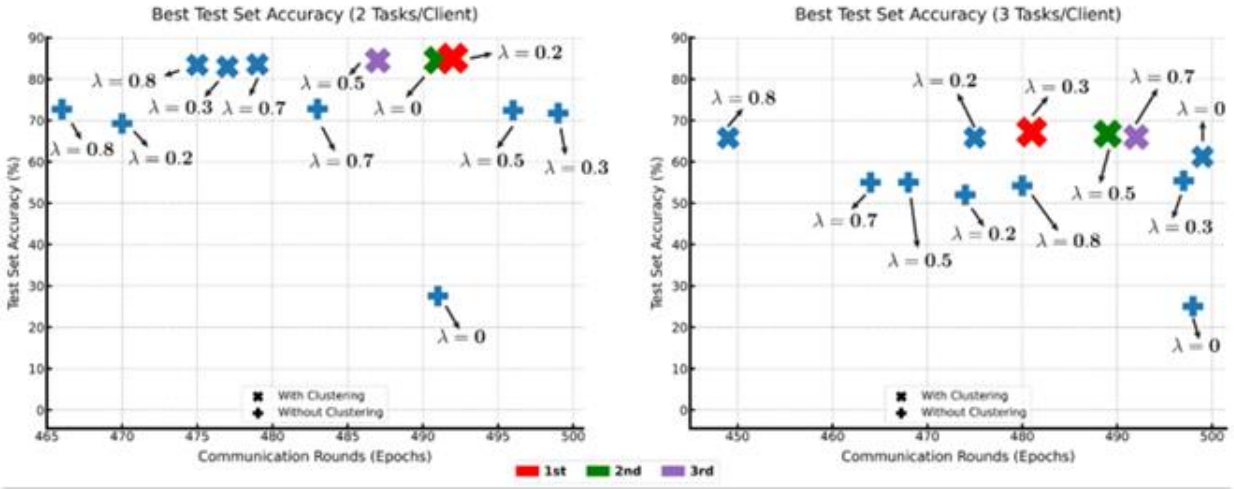


Figure 4. Top test set accuracy for each lambda, clustering or not, along with the training round it occurred.

Acknowledgments

This research has been co-financed by the Digital Transformation of Small and Medium Enterprises Initiative of the European Union (NextGeneration EU) and Greek national funds (Ellada 2.0) through the program “Development of Digital Product and Services”, project acronym “Product Watch” (project code: 104702).

References

[1] Chelsea Finn, P. A. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Retrieved from https://arxiv.org/abs/1703.03400

[2] Farzin Haddadpour, M. M. (2019). On the Convergence of Local Descent Methods in Federated Learning. Retrieved from https://arxiv.org/abs/1910.14425

[3] Felix Sattler, K.-R. M. (2019). Clustered Federated Learning: Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints. Retrieved from https://arxiv.org/abs/1910.01991

[4] Filip Hanzely, P. R. (2020). Federated learning of a mixture of global and local models. Retrieved from https://arxiv.org/abs/2002.05516.

[5] H. Brendan McMahan, E. M. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. Retrieved from https://arxiv.org/abs/1602.05629

[6] Virginia Smith, C.-K. C. (2018). Federated multi-task learning. Retrieved from https://arxiv.org/abs/1705.10467

[7] Yishay Mansour, M. M. (2020). Three Approaches for Personalization with Applications to Federated Learning. Retrieved from https://arxiv.org/abs/2002.10619.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 17–23

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Leveraging Empirical Mode Decomposition for Time Series Anomaly Detection in IoT and Smart City Applications

Ioannis Stasinopoulos[1] and Ilias Theodorakopoulos[1]
[1] *Democritus University of Thrace, Greece.*
ioanstas@ee.duth.gr, *iltheodo@ee.duth.gr*

Abstract

This paper proposes a novel forecasting-based anomaly detection method that leverages Empirical Mode Decomposition (EMD) to break down complex univariate time series into Intrinsic Mode Functions (IMFs). Gated Recurrent Units are tasked with processing IMFs to predict future values. By reconstructing the time series from predicted IMFs, anomalies are detected when deviations between predicted and observed values of the time-series occur. The proposed method is validated on real-world datasets from smart city applications, demonstrating its efficiency in handling noisy data and multiscale seasonal trends while maintaining low computational overhead, making it suitable for deployment in resource-constrained IoT environments.

1 Introduction

Anomaly detection in univariate time series (TS) is a critical task across various domains, including finance, healthcare, and industrial monitoring. The primary approaches to anomaly detection involve forecasting, reconstruction, representation learning, and hybrid models [1]. Forecasting-based techniques predict future data points based on historical trends, flagging anomalies when actual observations deviate significantly from predictions. This method is particularly effective for real-time monitoring applications, such as detecting abnormal traffic in data centers, monitoring urban environments and financial markets or identifying equipment malfunctions in industrial settings. Reconstruction-based techniques, on the other hand, focus on capturing the underlying structure of TS data by reconstructing it from a low-dimensional representation. Anomalies are detected when the reconstruction fails to accurately represent the data. These techniques are especially useful in complex or noisy environments, such as medical diagnostics, where subtle deviations from the norm can indicate significant issues. Representation learning techniques, which involve learning a compact and informative representation of the TS, are adept at capturing intricate patterns in data. They are often employed in applications like fraud detection, where understanding the relationships between transactions is crucial. Hybrid models combine these approaches to leverage the strengths of each, making them effective in scenarios where both temporal and spatial data characteristics are important, such as in the management of multivariate TS in power grids.

In the context of Internet of Things (IoT) and smart city applications like traffic monitoring, the choice of anomaly detection technique must consider the limited computational resources available. Forecasting-based methods present a compelling solution for such applications, since in some configurations such methods can be computationally efficient and operate effectively on devices

with constrained processing capabilities. By focusing on predicting the next data point in a TS, forecasting based models can both promptly identify anomalies with small computational footprint, as also benefit from training without the need of data painstakingly annotated in quest of often rare and indistinct anomalies. The scalability of forecasting-based models also makes them ideal in domains such as smart cities, for widespread use across a city's sensor network. Each sensor can be equipped with a locally trained model that is tailored to the specific traffic patterns of its location, enabling context-aware anomaly detection. This localized approach not only enhances the accuracy of anomaly detection but also ensures that the system remains efficient and responsive, even as the number of deployed sensors increases.2

In this work, we present a lightweight forecasting-based method for anomaly detection in univariate TS, leveraging signal decomposition via Empirical Mode Decomposition (EMD) [2]. The employment of EMD allows the separation of a TS into multiple simpler components, thus enabling improved forecasting with small recurrent neural models in the form of Gated Recurrent Units (GRU) architecture. The rest of the paper is organized as follows: In Section 2 we provide a brief outline of the literature in forecasting-based anomaly detection in TS and comment on the pros and cons of each neural architecture. In section 3 we describe the proposed method for anomaly detection. In section 4 we provide experimental results on two popular datasets from the field of traffic monitoring in urban environments and highlight the efficiency of decomposition-based approaches in forecasting with small computational footprint. Finally, conclusions are drawn in section 5.

2 Related work

Forecasting-based models are fundamental for anomaly detection in univariate TS, wherein a model learns historical patterns to predict future values or sequences. Anomalies are flagged when the deviation between the predicted and actual values exceeds a defined threshold. This approach is particularly effective for real-world anomaly detection, where normal behavior is plentiful, and anomalous instances are rare [3]. By forecasting one step at a time, these models can efficiently identify deviations in both short-term and long-term dependencies within the TS.

One of the primary architectures used for forecasting-based anomaly detection in univariate TS is the Long Short-Term Memory (LSTM) [4] network, a variant of Recurrent Neural Networks (RNNs) designed to capture long-term dependencies in sequential data. For instance, LSTM-AD [5] is specifically designed to handle univariate TS (UTS) without labeled data, leveraging stacked LSTM layers to forecast multiple future time steps, allowing model to capture complex temporal patterns, making it more effective particularly when the distinction between normal and anomalous data depends on long-term behavior. Bontemps et al. [6] introduced a simpler LSTM architecture for collective anomaly detection in UTS, focusing on learning normal TS behavior and detecting collective anomalies by tracking prediction errors that exceed a threshold.

Besides recurrent neural architectures, feedforward topologies such as Convolutional Neural Networks (CNNs) and Transformers have also been proposed for forecasting-based anomaly detection. DeepAnt [7], a CNN-based model, applies a sliding window approach for forecasting in UTS, predicting future values based on historical windows and detecting anomalies by comparing predictions to actual observations. DeepAnt is particularly notable for its robustness to data contamination and its ability to detect small deviations in TS patterns in noisy or sparse datasets. TCN-ms [8] (Multiscale TCN) extends the standard Temporal Convolutional Network (TCNs) [9], by incorporating multiscale dilated convolutions to capture temporal dependencies at various resolutions, making it effective for detecting anomalies in quasi-periodic data, such as medical TS or

industrial sensor data.

SAnD (Self-Attention Network for Anomaly Detection) [10] utilizes transformers' attention mechanisms to detect anomalies in UTS data. Unlike RNN-based models, SAnD processes entire sequences in parallel, capturing both short-term and long-term dependencies simultaneously. Furthermore, the self-attention mechanisms enhances its ability to detect anomalies in clinical and industrial applications where fine-grained analysis is essential.

While feedforward architectures offer robustness to noise and superior multi-scale mechanisms, they often require significant computational resources. An alternative approach to improving forecasting with simpler recurrent models is decomposing the TS into simpler components, where long term trends are removed or disregarded, allowing multiple simpler recurrent models to forecast each component separately. In that spirit, the recently proposed AD-LTI [11] combines Gated Recurrent Unit (GRU) [12] networks with the Prophet model [13] to enhance the detection of seasonal patterns in UTS. GRU, is employed to model temporal dependencies, while the Prophet model captures seasonal trends. A key innovation of AD-LTI is the Local Trend Inconsistency (LTI) metric, which measures the deviation of recent predictions from the expected trend, enabling more reliable anomaly detection, even when historical data includes noise or anomalies.

3 Proposed Method

A popular method for the decomposition of UTS into simple components is Empirical Mode Decomposition (EMD) [2]. EMD effectively and efficiently [14] decomposes a complex, non-linear TS signal into a finite number of simpler components, known as Intrinsic Mode Functions (IMFs). Each IMF captures specific oscillatory modes within the signal, isolating patterns such as noise, short-term fluctuations, and long-term trends. This decomposition simplifies the prediction task by enabling a GRU model to focus on predicting the evolution of less complex, individual IMFs rather than the entire raw signal.

EMD begins by decomposing the input TS $\boldsymbol{x} = \{x_0, x_1, \dots, x_t\}$ into a set of m IMFs, $\boldsymbol{I^1} = [I_0^1, I_1^1, \dots, I_t^1]$ ..... $\boldsymbol{I^m} = [I_0^m, I_1^m, \dots, I_t^m]$. Each IMF represents a distinct component of the original signal, ranging from high-frequency oscillations to slower trends. This separation of timescales allows the forecasting model to handle different temporal characteristics of the data in a structured way. For instance, higher frequency IMFs, which capture short-term fluctuations, can be prioritized for prediction tasks, while lower-frequency IMFs representing long-term trends can be left unmodified, as they provide less information for detecting short-term anomalies.

Figure 1 illustrates the overview of our proposed approach for UTS anomaly detection. In this scheme, each IMF is used as input to a separate GRU model, with the task of predicting the next value in its respective series. The GRU models are trained in an unsupervised manner, using a reconstructive objective to minimize the difference between the predicted and actual IMF values. Once the GRU models have forecasted the next value for each IMF, the predicted IMFs are summed to reconstruct the overall forecasted value of the original TS. This allows the model to combine the predictive power of the GRUs for the relevant IMFs while potentially incorporating unmodified components, such as longterm trends, that are not critical for short-term anomaly detection.

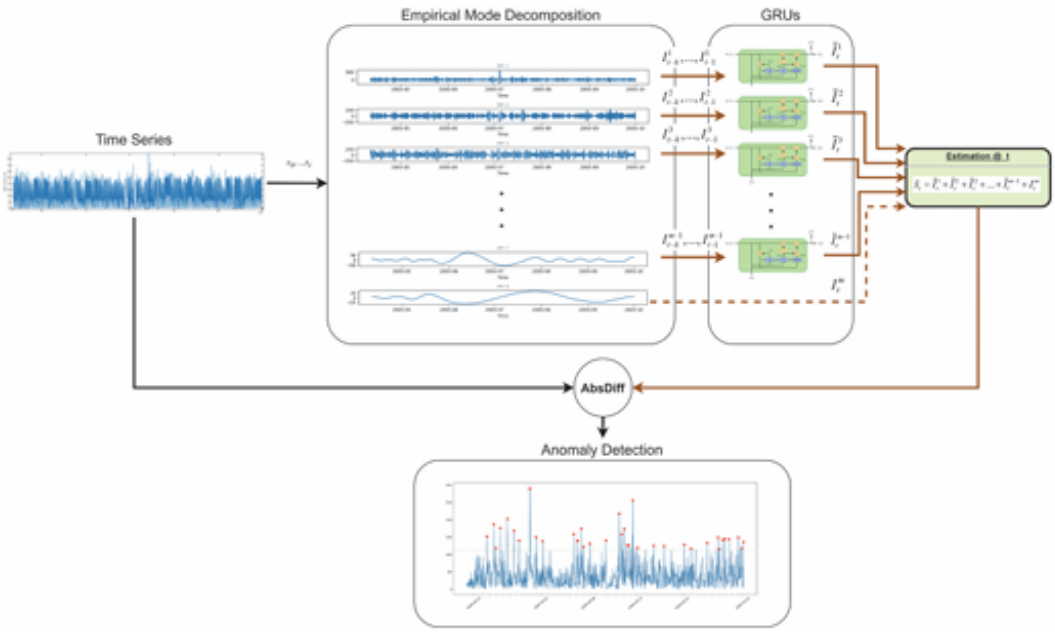


Figure 1: Overview of the proposed forecasting-base anomaly detection scheme

This framework also allows for flexibility in the choice of EMD variants. Ensemble EMD (EEMD) [15] can be used to improve robustness by addressing mode-mixing issues, often observed in standard EMD. Alternatively, in scenarios where current detection of anomalies is critical, online EMD [16] can be employed to ensure that the decomposition process can handle streaming data efficiently. Furthermore, depending on the prediction horizon required, the GRU models can be configured to process either the entire IMF sequence or a fixed-size time window, which captures the necessary temporal context for forecasting.

4 Experimental Evaluation

For the experimental evaluation, we used two real-world datasets, the Dodgers Loop Sensor[5] and Calit2[6] [17], to demonstrate the functionality of the proposed forecasting-based anomaly detection method in smart city applications. The Dodgers dataset consists of vehicle count data collected every 5 minutes over a span of six months from an inductive loop sensor on the Glendale on-ramp to the 101 North freeway in Los Angeles. With approximately 50,000 data points, this dataset focuses on comparing regular traffic patterns to those during baseball games at Dodger Stadium. However, it presents several challenges such as missing values due to sensor malfunctions and the presence of noise and irregularities caused by external factors like weather, holidays, and other special events that introduce significant variations in traffic patterns that are not explicitly marked in the data.

The Calit2 dataset captures building inflows and outflows from the main entrance of the Calit2 Institute at the University of California, Irvine, over a three-month period, recorded every 30 minutes, amounting to roughly 4,000 data points. This dataset monitors human activity entering and exiting the building, reflecting both regular daily patterns and sudden changes during organized events.

Similarly to [11], we converted The Dodgers TS to 1-hour sampling intervals, by summing up consecutive 5-min samples. For the Calit2 dataset we combined the two streams (inflows - outflows)

[5] https://archive.ics.uci.edu/dataset/157/dodgers+loop+sensor
[6] https://archive.ics.uci.edu/dataset/156/calit2+building+people+counts

to a UTS by computing the cumulative sum on people in and out of the building each day. For the Dodgers dataset we employed standard EMD and for Calit2 the EEMD variant. In both cases, 2-layer GRUs were employed with the size of hidden state being 55 for the Dodgers dataset and 32 for the Calit2. GRUs were trained using a 48-hour time window of the IMFs. To test e challenging setting, the Initial 60% of Dodgers dataset and 50% of Calit2 dataset were used to train the respective GRUs with a reconstructive objective, without using the annotation to discard the anomalous events. For both datasets, 40% of the rest of the TS was used for validation and 60% was used for testing utilizing the provided annotations.

Figure 2 (a-b) illustrates the ROC curves derived from Dodgers and Calit2 datasets from the proposed model. For comparison, we evaluated a baseline model using a GRU corresponding to each dataset on the raw TS data prior to decomposition. The goal for both datasets is to detect the presence of the events, so for the Dodgers dataset we consider as true positive an anomaly detected within a 2-hour window prior or after the annotated time of the game, and for Calit2 1 hour prior and after an event. Figure 2 (c)

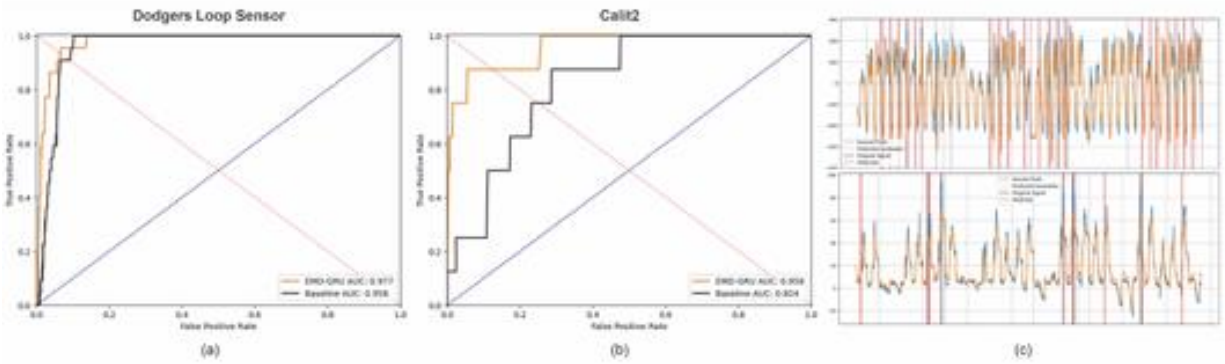


Figure 2: Experimental results. ROC curve for Dodgers (a) and Calit2 (b) datasets for the proposed (orange) and baseline (black) models. (c) Forecasted TS and predicted anomalies detections vs ground truth for Dodgers (top) and Calit2 (bottom) datasets. Blue: Actual TS, Orange: Estimated TS, Red: GT, Dashed: Predicted Anomalies.

illustrates the original TS, the forecasted TS for 1-sample ahead, the predicted anomalies based on the best threshold as defined in the validation set and the annotated anomalies. With obtained AUC of 0.977 for Dodgers and 0.958 for Calit2, it can be easily inferred that the EMD approach has significantly better performance from the baseline approach for both datasets (0.958 and 0.824 resp.), validating out assumption that by employing TS decomposition a lightweight recurrent model can offer remarkable performance in detecting anomalies in complex real-world scenarios.

5 Conclusions

In this study, we introduced a novel forecasting-based anomaly detection method that combines EMD with GRUs for anomaly detection in univariate time series data. By decomposing the time series into simpler IMFs, we enhanced the GRU's ability to predict short-term variations, improving anomaly detection performance in small computational budget. The experimental results on real-world smart city datasets demonstrate that the proposed method effectively detects anomalies, handling both noise and multiscale seasonal trends. In the future we intend to investigate alternative mechanisms of prediction such a Temporal Neural Networks and extend the proposed scheme into multivariant time series.

References

[1] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep Learning for Time Series Anomaly Detection: A Survey," ACM Comput. Surv., Aug. 2024, doi: 10.1145/3691338.

[2] G. Rilling, P. Flandrin, and P. Gonçalves, "On empirical mode decomposition and its algorithms," presented at the Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, 2003.

[3] J. Ma and S. Perkins, "Online novelty detection on temporal sequences," in KDD '03. New York, NY, USA: Association for Computing Machinery, Aug. 2003, pp. 613–618.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[5] P. Malhotra, L. Vig, G. M. Shroff, and P. Agarwal, "Long Short Term Memory Networks for Anomaly Detection in Time Series," presented at the European Symposium on Artificial Neural Networks, 2015.

[6] L. Bontemps, V. L. Cao, J. McDermott, and N.-A. Le-Khac, "Collective Anomaly Detection Based on Long Short-Term Memory Recurrent Neural Networks," in Future Data and Security Engineering, T. K. Dang, R. Wagner, J. Küng, N. Thoai, M. Takizawa, and E. Neuhold, Eds., Cham: Springer International Publishing, 2016, pp. 141–152.

[7] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series," IEEE Access, vol. 7, pp. 1991–2005, 2019,

[8] Y. He and J. Zhao, "Temporal Convolutional Networks for Anomaly Detection in Time Series," J. Phys.: Conf. Ser., vol. 1213, no. 4, p. 042050, Jun. 2019.

[9] M. Thill, W. Konen, and T. Bäck, "Time Series Encodings with Temporal Convolutional Networks," in Bioinspired Optimization Methods and Their Applications, B. Filipič, E. Minisci, and M. Vasile, Eds., Cham: Springer International Publishing, 2020, pp. 161–173.

[10] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: clinical time series analysis using attention models," in AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, Feb. 2018, pp. 4091–4098.

[11] W. Wu et al., "Developing an Unsupervised Real-Time Anomaly Detection Scheme for Time Series With Multi-Seasonality," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 9, pp. 4147–4160, Sep. 2022.

[12] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in Proceedings of SSST-8, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111.

[13] S. J. Taylor and B. Letham, "Forecasting at scale," PeerJ Inc., e3190v2, Sep. 2017.

[14] Y.-H. Wang, C.-H. Yeh, H.-W. V. Young, K. Hu, and M.-T. Lo, "On the computational complexity of the empirical mode decomposition algorithm," Physica A: Statistical Mechanics and its Applications, vol. 400, pp. 159–167, Apr. 2014.

[15] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in 2011 IEEE International Conference

on Acoustics, Speech and Signal Processing (ICASSP), May 2011, pp. 4144–4147.

[16] R. Fontugne, P. Borgnat, and P. Flandrin, "Online Empirical Mode Decomposition," in ICASSP 2017, Mar. 2017, pp. 4306–4310.

[17] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying poisson processes," in KDD '06. New York, NY, USA: Association for Computing Machinery, Aug. 2006, pp. 207–216.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 24-30

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Edge-Optimized NILM: Combining Structured Pruning and Quantization for Energy Disaggregation

Sotirios Athanasoulias[1], Dimitris Karagkounis[2], George Samaras[2], Isidoros Kokos[2]
[1] *National Technical University of Athens, Athens, Greece sotiriosathanasoulias@mail.ntua.gr*
[2] *Intracom Telecom, Athens, Greece*
{dkaragkounis, gsamaras, isik}@intracom-telecom.com

Abstract

Non-Intrusive Load Monitoring (NILM) enables the disaggregation of total energy consumption, measured in a single household, into the individual energy usage of domestic appliances. This paper presents a novel methodology aimed at optimizing NILM applications by integrating structured pruning and quantization techniques to enhance the efficiency and performance of deep neural networks for edge-enabled deployment. Structured pruning is employed to identify and remove non-essential weights on a layer-by-layer basis while preserving critical feature information and maintaining overall model performance. In parallel, quantization is applied further to compress the model in terms of its memory requirements. The proposed approach is evaluated using the Plegma dataset, a novel resource in NILM research that captures local consumption behaviors and device usage pat terns specific to the Mediterranean region. Experimental results demonstrate significant reduction in the size of the model without jeopardizing inference performance, highlighting the potential of edge-based NILM to contribute to energy efficiency and transition in Mediterranean environments.

1 Introduction

Non-Intrusive Load Monitoring (NILM), also known as energy disaggregation, enables the identification of the operational state and energy consumption of individual appliances from aggregated consumption data [4]. NILM empowers consumers by enhancing energy awareness, improving cost management, supporting demand-side flexibility services, and promoting energy efficiency, ultimately aiding the transition to more sustainable energy systems. Advanced metering infrastructure is crucial for NILM applications, providing high-resolution, ideally real time, data streams. With over half of EU households now equipped with smart meters, 13 countries surpassing 80% penetration, and a strategic focus on smart meter deployment both in the EU and globally [1], NILM has the potential to play a pivotal role in advancing innovative energy services.

NILM algorithms span from traditional statistical analysis methods to more recent deep learning approaches, with the latter demonstrating promising outcomes [4], though they come with higher computational demands. These demands pose challenges for deployment in edge environments and limit scalability due to privacy concerns and data transmission issues. NILM is inherently a context-sensitive problem, as appliance energy consumption is influenced by a range of external and internal factors, including weather conditions, seasonal variations, appliances' usage patterns and

technological variations. Although substantial research has been conducted in regions such as Northern Europe, the UK, and the USA [2], studies focusing on the Mediterranean are relatively limited. This region presents distinct environmental conditions and energy use patterns, particularly in relation to appliances like air conditioners and electric boilers, which contribute significantly to household energy consumption and hold potential as sources of demand-side flexibility.

To address these gaps, our work proposes an edge-based NILM approach for performance aware model minimization, facilitating edge deployment for such applications in the domestic sector. This work also provides and evaluates a NILM algorithm in a new context, utilizing the newly published Plegma dataset [3]. This open dataset provides granular measurements from in-home appliances in the Mediterranean region, enriching the availability of open data for energy disaggregation found in the literature and enabling novel energy services that consider the specific behavioral patterns and devices of the region. In summary, the contributions of this work are as follows:

- Introduce an optimized model minimization methodology for edge-NILM applications based on structured pruning and quantization
- Establish a benchmark for NILM algorithms in the Mediterranean context using the Plegma dataset, while offering insights into the impact of region-specific appliance usage patterns on NILM performance and scalability.

The rest of the article is organized as follows: Section 2 presents the related work, Section 3 presents the mathematically formulation of the problem of NILM and describes the proposed edge-optimized methodology, Section 4 discusses the evaluation results, while concluding re marks are presented in Section 5.

2 Related work

Deep learning has seen remarkable success across various fields, including natural language processing, time-series analysis, and computer vision. Recently, several deep learning methods have been developed for NILM, demonstrating superior performance in energy disaggregation tasks. However, despite their effectiveness, these deep learning-based NILM models are hindered by high computational complexity, leading to increased training costs and presenting significant challenges for real-world deployment, particularly in resource-constrained edge environments [8].

To overcome these challenges, recent advancements have driven the integration of NILM and related energy applications into edge devices. Deploying these models on the edge not only reduces computational overhead but also eliminates the need for data transmission between users and a central server, addressing privacy concerns and challenges tied to centralized data processing. The research landscape for NILM on edge devices is expanding rapidly, with various methodologies being explored, including deep learning models optimized for edge deployment, feature extraction techniques, and federated learning approaches [4,5,8].

A common technique for compressing NILM models is pruning, which is a deep learning method used to create more computationally efficient neural networks by eliminating less significant parameters [8]. However, in most studies, pruning is implemented through the use of binary masks, where model weights are set to zero but not physically removed from the network. While this method offers a reasonable approximation of the model's behavior, it does not represent true pruning, as the network's structure remains unchanged.

To address this issue and enable real-world deployment, we implement a comprehensive and fully

automated structured pruning method. This approach explicitly models the dependencies between layers and comprehensively group coupled parameters for pruning [6]. Additionally, we combine this method with dynamic model quantization to further reduce the model size, enhancing the overall efficiency of deploying NILM applications on edge devices.

3 Optimising NILM for edge deployments

3.1. Problem Formulation

Non-intrusive load monitoring (NILM) problem [7], involves estimating the electricity consumption of individual appliance y_i $(i \in [1, M])$, solely based on the aggregated consumption at household level x, at timepoint t within a fixed time window $t \in [1, T]$. A noise term ϵ is also utilized to capture error of measurement instruments and appliances not individually metered. Hence, NILM formula is defined as:

$$x(t) = \sum_{i=1}^{M} y_{i(t)} + \varepsilon(t)$$

Deep learning networks are frequently used to address this problem, as they can analyze time series data and identify appliance signatures. However, the high-dimensional nature and computational demands of these models require the application of compression techniques to enhance efficiency.

3.2. Structured Pruning

Pruning in deep neural networks (DNNs) is a widely used technique aimed at reducing model size and computational complexity by eliminating unnecessary parameters, thereby enhancing the model's efficiency while maintaining acceptable performance. Pruning can target either individual weights or entire structural units within the network [4]. Unstructured pruning involves the removal of individual weights, while structured pruning eliminates entire units, such as filters or channels, based on a predefined criterion, such as the L1 norm. In the case of structured pruning, the sparsity ratio s is defined as the proportion of pruned units (filters, neurons, or channels) relative to the total units. For structured pruning, a channel C_i is pruned if its L1 norm $\left|\left|C_i\right|\right|_1$ falls below a threshold determined by the sparsity ratio s. This can be expressed as $\left|\left|C_i\right|\right|_1 < s$, for all $Ci \in C$. After pruning, the resulting set of active channels is denoted as $C' \in \mathbb{N}^{k'}, where\ k' < k$, is the reduced number of channels.

The structured pruning process reduces both the model size and the computational com plexity. The number of parameters in the network is reduced proportionally to the number of pruned channels, and the computational cost is reduced by eliminating the associated floating-point operations (FLOPs) of the pruned channels.

3.3. Dynamic Quantization

Quantization is a technique that reduces storage and memory requirements by converting model weights from floating-point to lower-precision integers. This process can be described as an irreversible mapping where floating-point weights are discretized into integer bins [9]. Quantization may be performed either post-training, by compressing a pre-trained model, or during training through quantization-aware training. This study focuses on post-training dynamic quantization, which offers flexibility and adaptability to varying input data. It also simplifies the compression process by determining scale and zero-point dynamically during inference, removing the need for a

separate calibration phase.

In this study, dynamic quantization was applied specifically to the linear layers of the model because these layers, compared to convolutional layers, are less sensitive to precision loss and have a more predictable activation range. This makes linear layers better suited for efficient quantization without significant degradation in model accuracy.

4 Experimental Results

4.1. Experimental Setup

The selected model for this study is a sequence-to-sequence 1D Convolutional Neural Network, noted for its effectiveness in edge-NILM tasks [4]. Experiments were conducted using the Plegma dataset [3], featuring electricity consumption data sampled every 10 seconds. The model was trained on data from houses 1-13, with house 2 excluded and used as unseen test data.

To implement the model compression strategy, we utilized DepGraph [6] and the associated Torch-Pruning library. These tools prune substructures based on their interdependencies, resulting in smaller models, unlike other libraries that only apply binary masking without fully removing components from the model. Regarding quantization, we employed the Pytorch quantization library.

4.2. Model Compression & Performance Evaluation Metrics

The model's performance in predicting the energy consumption profiles of the tested appliances was assessed using Mean Absolute Error and Mean Relational Error. F1 Score was used to measure its accuracy in identifying the operational states of appliances (on/off). Additionally, the model's computational complexity was evaluated through Floating Point Operations (FLOPS), the number of trainable parameters, and their storage size in Megabytes (MB). To determine the optimal pruning threshold, we used the metric from [4], selecting the point where the Euclidean distance between the sparsity-F1 pruning ratio curve and the ideal point is minimized.

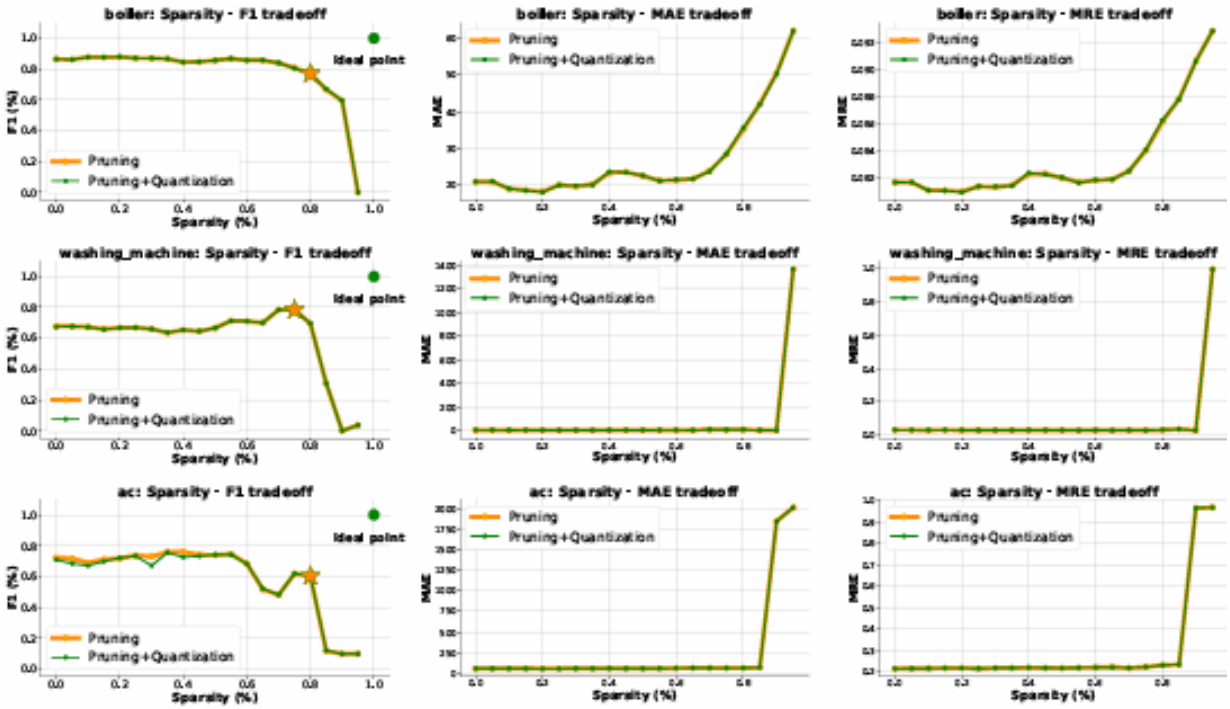


Figure1: Comparison of different pruning thresholds (0-95%) to performance degradation (F1, MAE, MRE) for different devices for structured pruning and structured pruning with dynamic quantization approaches. The optimal pruning thresholds for both approaches are denoted with the star symbol.

4.3. Results Evaluation

Figure 1 presents a comparative analysis between structured pruning and structured pruning combined with dynamic quantization. The evaluation spans various compression levels, with sparsity ratios ranging from 0% to 95%, and uses the selected performance metrics. The resulting 'sparsity vs. performance' curves indicate that both methods demonstrate comparable performance across all pruning thresholds. In most cases, pruning with dynamic quantization achieves nearly identical results to pruning alone, though it occasionally exhibits a slight decrease in performance.

Appliance	Approach	Pruning (%) P_thr = P_opt	Performance metrics		
			F1- Score	MAE	MRE
Boiler	Baseline Model	0	0.864	21.002	0.081
	Structured Pruning	80	0.772	35.067	0.086
	Structured Pruning + Quantization	80	0.768	35.480	0.086
Washing Machine	Baseline Model	0	0.757	7.834	0.029
	Structured Pruning	75	0.806	9.678	0.026
	Structured Pruning + Quantization	75	0.784	9.680	0.026
A/C	Baseline Model	0	0.72	56.687	0.215
	Structured Pruning	80	0.603	63.842	0.231
	Structured Pruning + Quantization	80	0.602	63.855	0.231

Table 1: Comparative performance evaluation results- baseline model vs optimal pruned model vs optimal pruned quantized model.

A closer examination of the performance metrics at the optimal pruning thresholds, as shown in

Table 1, reveals that the optimal pruning thresholds are identical for both compression methods across all tested appliances. The performance metrics confirm the insights drawn from the sparsity diagram, showing that structured pruning with quantization has a minimal impact on the performance of structured pruning alone for all appliances.

Appliance	Approach	Pruning (%) P_thr = P_opt	Compression metrics		
			Model params	FLOPS	MB
Boiler	Baseline Model	0	22147640	2501.32	84.49
	Structured Pruning	80	868835	89.96	3.31
	Structured Pruning + Quantization	80	868835	89.96	0.84
Washing Machine	Baseline Model	0	22147640	2501.32	84.49
	Structured Pruning	75	1417050	154.84	5.41
	Structured Pruning + Quantization	75	1417050	154.84	1.36
A/C	Baseline Model	0	22147640	2501.32	84.49
	Structured Pruning	80	868835	89.96	3.31
	Structured Pruning + Quantization	80	868835	89.96	0.84

Table 2: Comparative compression evaluation results- baseline model vs optimal pruned model vs optimal pruned quantized model.

Although dynamic quantization has minimal impact on the performance of the pruned models, the same cannot be said for compression metrics. As shown in Table 2, quantization significantly reduces the storage requirements of the pruned models by 74.62% for both the boiler and the A/C, and by 74.86% for the washing machine. The FLOPS and the number of trainable parameters, however, remain unchanged, as quantization does not alter these aspects of the model. In conclusion, the proposed structured pruning combined with dynamic quantization proves to be an effective approach, as it maintains model performance while significantly reducing storage requirements. When comparing the baseline model with the optimized approach, we observe a slight performance decline, yet the compression metrics exhibit significant improvements. Specifically, the number of trainable parameters and FLOPS were reduced by up to 96%, and storage requirements saw a reduction of 99%. Despite an average performance drop of approximately 10% across all appliances, there were instances, such as with the washing machine, where the compressed model outperformed the baseline. This improvement is likely due to the overparameterization of the baseline model, allowing the compressed model to generalize more effectively.

5 Conclusions

This study tackles the challenges of resource-intensive deep learning models in NILM, specifically addressing their context-aware nature. Unlike previous research, this work benchmarks appliances such as air conditioners and boilers from the Plegma dataset, which are common in the Mediterranean region. It evaluates the performance of a baseline CNN model, structured pruning, and structured pruning combined with quantization. Results show that all methods effectively disaggregate appliance usage, providing insights into NILM's applicability in this region. Notably, the structured pruning approach reduces the model's MFLOPs by up to 96% and model size by 99%, lowering computational demands, improving feasibility for edge deployment, promoting NILM technologies for wider adoption.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 and Horizon Eu rope research and innovation program under the Marie Sklodowska-Curie and ODEON grant agreements, No 955422 and No 101136128, respectively.

References

[1] Energy Retail and Consumer Protection 2023 Market Monitoring Report, ACER, September 2023.

[2] Georgios-Fotios Angelis, Timplalexis, Stelios Krinidis, Dimosthenis Ioannidis, and Dimitrios Tzo varas. Nilm applications: Literature review of learning approaches, recent developments and chal lenges. Energy and Buildings Volume, 261, 2022.

[3] S. Athanasoulias, F. Guasselli, N. Doulamis, A. Doulamis, N. Ipiotis, A. Katsari, L. Stankovic, and V. Stankovic. The plegma dataset: Domestic appliance-level and aggregate electricity demand with metadata from greece. Scientific Data, vol. 11, no. 1, p. 376, April 2024.

[4] S. Athanasoulias, S. Sykiotis, M. Kaselimi, A. Doulamis, N. Doulamis, and N. Ipiotis. Opt-nilm: An iterative prior-to-full-training pruning approach for cost-effective user side energy disaggregation,. Transactions on Consumer Electronics, vol. 70, no. 1, page 4435–4446, 2024.

[5] Sotirios Athanasoulias, Stavros Sykiotis, Nikos Temenos, Anastasios Doulamis, and Nikolaos Doulamis. A pre-training pruning strategy for enabling lightweight non-intrusive load monitor ing on edge devices. In 2024 ICASSPW, pages 249–253, 2024.

[6] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16091–16101, 2023.

[7] G. W. Hart. Nonintrusive appliance load monitoring. Proceedings of the IEEE, 80, no. 12:pp. 1870–1891, 1992.

[8] Stavros Sykiotis, Sotirios Athanasoulias, Maria Kaselimi, Anastasios Doulamis, Nikolaos Doulamis, Lina Stankovic, and Vladimir Stankovic. Performance-aware nilm model optimization for edge deployment. IEEE Transactions on Green Communications and Networking, 7(3):1434–1446, 2023.

[9] Yuhui Xu, Shuai Zhang, Yingyong Qi, Jiaxian Guo, Weiyao Lin, and Hongkai Xiong. Dnq: Dynamic network quantization. arXiv preprint arXiv:1812.02375, 2018.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 31-38

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Federated Learning for Workload Forecasting as enabler for Network Service Management

Dimitrios Brodimas[1], Dimitrios Kapolos[1], Eleftherios Mylonas[1, 2], Alexios Birbas[1] and Michael Birbas[1]

[1] *Dept. of ECE University of Patras, Patras, Greece*

[2] *Department of Research, Technology and Development, Independent Power Transmission Operator (IPTO) S.A., Athens, Greece*

brodimas@ece.upatras.gr, d.kapolos@ac.upatras.gr, e.mylonas@admie.gr, birbas@ece.upatras.gr, mbirbas@ece.upatras.gr

Abstract

Next generation networks are expected to connect and manage a vast number of heterogeneous devices stretching over diverse and distributed technological domains (e.g., radio, transport, core), by embracing a ubiquitous presence of AI and ML within their operations. This shift imposes a significant challenge for the management and orchestration (MANO) frameworks in order to efficiently handle that great demand for computational power, particularly in the edge of the networks. Therefore, in order to increase the effectiveness of the task scheduling of the next generation networks it is crucial to proactively detect both periodic and non-periodic patterns that could affect the network's decision-making processes. This paper showcases a computational load forecasting method that aims to feed the MANO frameworks with predicted insights of the infrastructure state. The forecasting is performed by employing machine learning techniques such as LSTMs and Bi-LSTMs, by utilizing the federated learning family of algorithms, FedOpt, facilitating the distributed training process. The experimental results show that the proposed federated learning approach provides results of comparable quality to the centralized learning methods, while minimizing the data exchanged between the edge nodes and the cloud.

1 Introduction

The tendency of the sensing layer and subsequently the computation to shift towards the edge, introduce a growing complexity and diversity of the next-generation networks. This transition poses significant challenges for traditional centralized management and orchestration (MANO) methods in their pursuit to achieve high efficiency and sustainability. These tools heavily rely on the insights and patterns of the infrastructure's usage to enable efficient resource management, prompting the exploration of innovative methods in predictive and prescriptive analytics within distributed network environments. The accurate prediction of different metrics such as CPU, memory, and bandwidth is of utmost importance for optimizing resource allocation, improving the network's performance, and enabling proactive decision-making. Traditional methods of microservices scheduling and mapping

mostly rely on monitoring just the current state of the computational resources, or on the best-case scenarios, they engage centralized approaches, leading to increased communication overhead and potential privacy concerns due to the continuous data transmission.

Edge computing allows for distributed computing resources to be located closer to the data sources, enabling real-time processing and analysis. By placing computing power at the network edge, latency can be minimized, and bandwidth demands can be reduced, thereby enhancing system responsiveness and efficiency.

In order to combine the edge computing with machine learning, federated learning (FL) has emerged as a promising solution. FL, a decentralized learning paradigm, enables collaborative model training among edge devices without the need for data transfer to a central server. This approach allows for reducing communication overhead and leveraging the computational power available at the network's edge. It also inherently complies with the data privacy constraint of the operators, who are often cautious with sharing data both for proprietary reasons and customer privacy requirements

The main contributions of this paper are the following:

- an extensive performance analysis of FL algorithms (FedAvg, FedAdam, FedYogi and FedAdagrad) compared to centralized forecasting methods for predicting the computational load of a network spanning across the edge-cloud environment
- the definition of a new metric, which along with other metrics, enables the evaluation of the overall performance of FL algorithms while incorporating the notion of network offloading.

The effectiveness and efficiency of the proposed framework is demonstrated through extensive experiments and evaluations. The results will showcase the benefits of leveraging FL in edge computing for CPU load forecasting, highlighting improvements in prediction accuracy, reduced communication overhead, and enhanced system performance. The findings of this research have the potential to contribute to the optimization of resource management in the distributed computing environments of the next generation networks, opening avenues for proactive decision-making and improved operational efficiency.

The work is structured as follows: Section 2 provides a literature review in the domain of CPU load forecasting and FL. Section 3 presents in detail the FL scheme implemented in this paper's context. In Section 4, the experimental setup is described, along with the results and the performance evaluation. Last, Section 5 comprises the conclusion and future directions for research in this area.

2 Related work

The utilization of CPU workload forecasting techniques to effectively manage and orchestrate microservices and hardware resources across the computing and network continuum has garnered significant interest. Various studies have addressed the challenges of accurately predicting CPU load in dynamic computing environments

Saxena et al. [1] conducted an extensive survey on the implementation of workload forecasting using different machine learning models. Karim et al. [2] introduced a hybrid prediction model called BHyPreC, which combines a gated recurrent unit (GRU) with a bidirectional long short-term memory

(BiLSTM) network to predict CPU usage workload in cloud virtual machines. This paper serves as a foundational basis for our research, as the BHyPreC model is incorporated in this work within the FL framework alongside other models. Subramanya et al. [3] proposed prediction-based NFVI (Network Function Virtualization Infrastructure) scaling solutions in virtualized Mobile Edge Computing (MEC) environments. They applied various neural network (NN) techniques and employed federated learning for workload prediction. Although their work shares similarities with ours in terms of using FL techniques, we differentiate ourselves by studying a wider range of algorithms beyond FedAvg and evaluating the results by incorporating the concept of network offloading.

3 Computational Load Forecasting

3.1. Federated Learning

FL operates as a decentralized approach that facilitates accumulating global knowledge from distributed clients. To initiate the process, a general or pre-trained model is distributed to the clients. Each client then utilizes its local data to train and creates a personalized model. Once the training is completed, the clients transmit their model's hyper-parameters back to the aggregator. The aggregator collects and combines all the updates received from the clients to construct an updated global model [4]. This ensures that the clients are consistently equipped with the latest global knowledge as it emerges.

Assuming that there are K clients participating in the FL procedure, each one equipped with a P_k set of indexes of data points and $n_k = |Pk|$, the optimization problem that needs to be solved is the following:

$$\min_{w \in \mathbb{R}^d} f(w) \tag{1}$$

Where

$$f(w) = \sum_{k=1}^{K} \left(\frac{n_k}{n} 1\right) F_{k(w)} \text{ and } F_k(w) = \left(\frac{1}{n_k}\right) \sum_{i \in P_k} f_i(w) \tag{2}$$

Typically, $f_i(w) = l(x_i, y_i; w)$ which is the loss function of the prediction example (x_i, y_i) made with model parameters w.

For the purposes of this paper, two FL algorithms are going to be studied: FedAvg and FedOpt.

1) *FedAvg:* FedAvg, introduced in [5], is a well-known and widely used FL algorithm. It relies on implementing the Stochastic Gradient Descent (SGD) algorithm to achieve convergence. During each round of training, FedAvg selects a fraction C of the clients and computes the gradient of the loss over the data held by these clients. The model weights w of client k are then updated locally using multiple epochs E of the formula $w^k = w^k - \eta \nabla f(w^k)$ where η is the learning rate.

   After the number of local epochs is completed, the local weights are sent back to the aggregator, which constructs the new model by aggregating the local model weights of the selected clients. The algorithm restarts by returning the updated weights to the clients and selecting new random candidates.

   Although FedAvg has exhibited good performance, recent studies have brought attention to its convergence challenges in certain scenarios [6]. These challenges arise from several factors,

including client drift, i.e., Non-Independent and Identical (Non-IID) data distribution across clients, and lack of adaptivity to the data examined.

2) *FedOpt:* FedOpt is an evolution of FedAvg aiming to solve the above-mentioned drawbacks by introducing the notion of adaptive optimizers at both client and aggregator levels instead of using stochastic gradient descent for both operations. The algorithm's operation is based on the FedAvg algorithm by formulating the update equation on the aggregator's side as in eq. (3).

$$w_{r+1} = \left(\frac{1}{n_k}\right)\sum_{k\in S_r} w_r^k = w_r - \left(\frac{1}{n_k}\right)\sum_{k\in S_r}\left(w_r - w_r^k\right) \quad (3)$$

Considering the differences $\Delta w_r^k = w_r^k - w_r$ and $\Delta_r = {}^1/_{n_k}\sum k\varepsilon S_r\, \Delta_r^k$ the FedAvg can be solved by applying the SGD to the $-\Delta_r$ with learning rate $\eta = 1$. Thus, by adopting SGD for the ClientOpt and Adagrad, Yogi or Adam for the ServerOpt, the FedAdagrad, FedYogi and FedAdam algorithms are generated, respectively. The actual changes are shown in eq. (4) and eq. (5).

$$w_{r,j+1}^k = w_{r,j}^k - \eta\, \nabla l(w_r^k;\, b) \quad (4)$$

$$w_{r+1} = w_r + \eta\, \left(\frac{m_r}{\sqrt{u_r + r}}\right) \quad (5)$$

where $m_r = \beta_1 m_r\text{-}1 + (1 - \beta_1)\, \Delta_r$, while:

$$u_r = u_{r-1} + (\Delta_r)^2 \quad (6)$$

$$u_r = u_{r-1} - (1 - \beta_2)(\Delta_r)^2 \mathrm{sign}(u_{r-1} - \Delta_r)^2) \quad (7)$$

$$u_r = \beta_2 u_{r-1} + (1 - \beta_2)(\Delta_r)^2 \quad (8)$$

By selecting one of the eq. (6), (7) and (8), we derive the algorithms of Adagrad, Yogi, and Adam, respectively. The parameter τ controls the algorithms' degree of adaptivity, with smaller values of τ representing higher degrees of adaptivity, while β_1 and $\beta 2$ are the decay rates of the average of the gradients ($\beta_1, \beta_1 \in \{0, 1\}$).

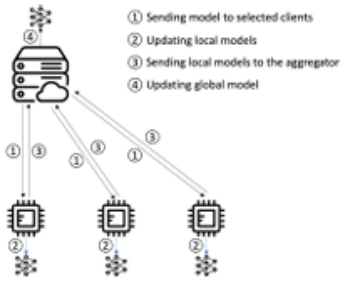


Figure1: Federated learning scheme

3.2. Losses and Metrics

During the whole building, training and testing procedure of both centralized and FL models, the Huber loss is less sensitive to outliers compared to the squared error loss, as it behaves like the absolute error loss for large errors (greater than delta) and like the squared error loss for small errors (less than or equal to delta). This makes it a robust alternative for regression problems where the presence of outliers or noisy data can affect the regression capabilities of the model. The evaluation metrics that

were used in both centralized and FL models were the Mean Squared Error (MSE), the Mean Absolute Error (MAE) and the Coefficient of Determination (R2). Finally, in this paper, an overall criterion is proposed in order to evaluate the models more adequately. As the MAE metric offers the most objective view of the models' performance, it is combined with the communication cost CC of the learning procedure to compose the new criterion ϕ. The mathematical formulation of this criterion is:

$$\phi = MAE \cdot CC \tag{9}$$

The communication cost is defined as the burden that the learning procedure imposes on the communication network for the required data to travel from their source up to the point of processing. It is considered that the *CC* for the centralized models is equal to 1 because all the data travel to the central node. For the FL part, the *CC* is thought as the size of the parameters exchanged between the server and the clients in bytes, multiplied by two (to include both steps 1 and 3 of Figure 1) and divided by the total amount of bytes of the dataset. Thus, the communication cost *CC* is calculated as:

$$CC = \begin{cases} 1, \ centralized\ learning \\ \frac{1}{D}\sum_{r=1}^{R}\sum_{k=1}^{S} 2P, \ federated\ learning \end{cases} \tag{10}$$

Where *D* is the total size of the data set, *P* is the number of parameters exchanged in bytes. Finally, *R* is the total number of rounds, *S* is the random set of clients in each round and *K* the total amount of clients.

4 Experimental Results

4.1. Experimental Setup

The experiments were conducted in a cloud-native compliant lab environment where K3s [7], a lightweight distribution of Kubernetes, was employed to unify the underlying computing infrastructure. This infrastructure is constituted by"

- one server equipped with an Intel® Core™ i9-12900F processor, an NVIDIA RTX A5000, 64GB RAM and 1TB disk capacity.
- one worker computer equipped with Apple M1 processor with integrated GPU, 16GB RAM and 256GB disk capacity.
- four Raspberry Pi 4 equipped with ARM processor, 4GB RAM and 64GB disk capacity.

To leverage the computational power of the graphics card and expedite time-consuming procedures, the centralized models were trained and tested on the server. In the case of FL models, all available nodes were utilized, with the server acting as the aggregator. The rest of the infrastructure hosts the twenty-two clients that are engaged in the experiment as microservices.

The dataset used for the purposes of this paper is the one provided by Bitbrains in [8]. This real-life dataset concerns traces originating from VMs that are connected to a fast storage area network (SAN) with storage devices located over a distributed data center. This dataset is in csv format for each VM, providing eleven exogenous values of CPU, memory, disk and bandwidth. The column "CPU Usage Percentage" is the selected input to obtain the needed information and perform the forecasting algorithms described in this paper. In terms of preprocessing, the data were checked for missing

values, abnormal values (e.g., negative values or values surpassing 100%) and outliers. No scaling action was performed as the data is already in the percentage scale. The initial dataset was split into training, testing, and validation sets, with the size of 70%, 20% and 10% of the total data, respectively.

4.2. Performance Evaluation and Discussion

In terms of the centralized models, all of them are closely aligned with the patterns of the dataset's timeseries. The bar chart in Figure 2 demonstrates that linear regression achieves the lowest MSE on the testing data, while BiLSTM performs best in terms of the validation data. On the other hand, BiLSTM attains the optimal MAE values for both testing and validation data. The R^2 metric, scaled up by a factor of 10 for better visualization in the figure, exhibits minimal differences among the models, with BiLSTM delivering the most favorable result. Subsequently, BiLSTM seems to have the best overall performance while utilizing a centralized structure with access to all the data, exhibiting a mean absolute deviation of 1.41%. Shifting towards the federated deployment of machine learning models showcased in Table 1 and Table 2, better results are evident with linear regression (mentioned as LR in the tables and figures), particularly in terms of MSE. Among the NNs, BiLSTM showcases the best performance. The inferior performance of the NNs can be justified by the constrained access to data, as each client model is exclusively trained on the timeseries of the single client. Consequently, this limitation hinders the generalization of their knowledge. When considering the FL algorithms, FedAvg consistently delivers superior outcomes in both MAE and MSE, regardless of the deployed model. Among the FedOpt family of algorithms, FedYogi exhibits a more favorable performance compared to FedAdam and FedAdagrad. It is important to note that results in this phase of deployment may vary due to the stochastic client selection during distributed training and evaluation sessions.

The predictive ability of the FL approach is always poorer than centralized learning, but this needs to be anticipated as the training is performed in particular set of data for each client. This is why a notable mention needs to be made for the combination of linear regression with FedAdam, which yields the best-distributed results that are comparable to those of centralized learning. The last mean of evaluation that is employed is ϕ. All the FL algorithms explored in this paper have the same amount of exchanged parameters in each step, thus the communication cost CC is only affected by the machine learning model parameters. For example, linear regression in the FL environment exchanges 2 ⋅ 10 learning rounds ⋅ 22 clients ⋅ 72 bytes of parameters ÷6095296 bytes of the dataset ≈ 0.005.

As seen in Table 3, the substantial difference in the order of magnitude of data exchanged between the centralized and FL models, along with the relatively smaller variation observed in terms of MAE, contributes to the favorable evaluation of the metric ϕ in the context of the FL scheme. This disparity highlights the efficient utilization of resources and the ability to achieve comparable levels of predictive accuracy in FL despite the reduced communication requirements. Consequently, the consideration of ϕ in FL compensates for the accuracy benefits of traditional centralized approaches (also exhibiting a lack of privacy).

5 Conclusions

The experimental results demonstrate that the FL approach achieves accurate and robust CPU load forecasting in a network service environment for efficient network management, approaching the results of traditional centralized methods. Metric ϕ proves that this deployment offers better results

for the overall computational and network environment, than the centralized learning, even if the forecasting accuracy is mildly poorer. This solution could be extended by adopting of visual transformers for the learning part, such as the one in [9], as they can discover hidden patterns in the time series with less data available in comparison with the NNs used in this paper. In addition the exploration of pure networking data as input for the algorithms, such as bandwidth or latency, could be of great interest in order to provide a holistic future view of the underlying network infrastructure to the MANO frameworks.

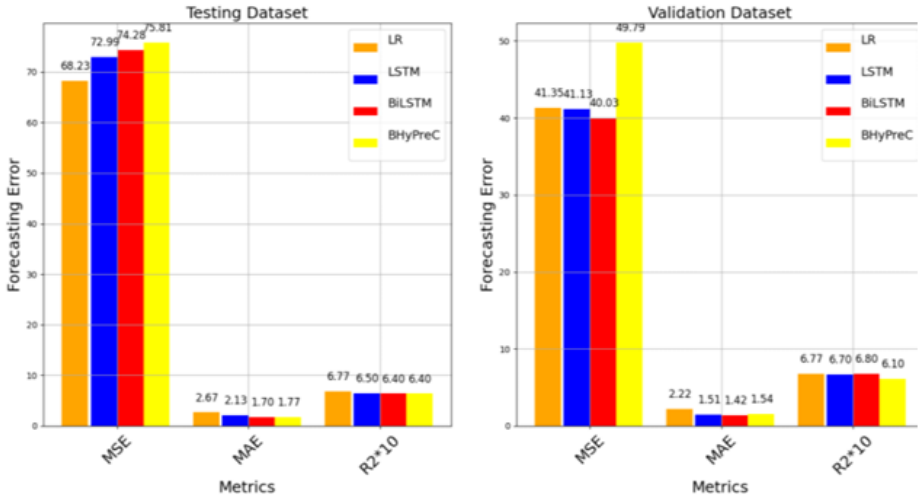


Figure 2: Centralized models analysis

	FedAvg	FedAdagrad	FedYogi	FedAdam	Mean
LR	4.93	5.37	4.95	2.53	4.72
LSTM	4.64	6.10	7.44	5.95	6.03
BiLSTM	4.82	5.96	5.88	6.76	5.86
BHyPreC	6.01	6.51	8.13	11.84	8.12
Mean	5.10	6.88	5.99	6.77	

Table 1: Federated Learning MAE

	FedAvg	FedAdagrad	FedYogi	FedAdam	Mean
LR	66.94	116.35	105.08	48.18	93.31
LSTM	182.93	273.01	323.54	247.02	175.74
BiLSTM	172.62	194.57	106.59	207.78	170.29
BHyPreC	139.25	154.40	328.26	233.80	213.93
Mean	140.43	184.19	144.15	184.58	

Table 2: Federated Learning MSE

	Centralized			Federated		
	CC	MAE	ϕ	CC	MAE	ϕ
Lin. Regression	1	2.22	2.22	0.005	4.72	0.115
LSTM	1	1.51	1.51	0.007	6.03	0.273
BiLSTM	1	1.42	1.42	0.007	5.86	0.257
BHyPreC	1	1.54	1.54	0.021	8.12	1.409

Table 3: Metric ϕ Comparison

Acknowledgements

This work has been partially supported by the EU Horizon INCODE project Programming Platform for Intelligent Collaborative Deployments over Heterogeneous Edge-IoT Environments (HORIZON-CL4-2022-DATA-01-03 / 101093069).

References

[1] D. Saxena, J. Kumar, A. K. Singh, and S. Schmid, "Performance analysis of machine learning centered workload prediction models for cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 4, pp. 1313–1330, 2023.

[2] M. E. Karim, M. M. S. Maswood, S. Das, and A. G. Alharbi, "Bhyprec: A novel bi-lstm based hybrid recurrent neural network model to predict the cpu workload of cloud virtual machine," IEEE Access, vol. 9, pp. 131 476–131 495, 2021.

[3] T. Subramanya and R. Riggio, "Centralized and federated learning for predictive vnf autoscaling in multi-domain 5g networks and beyond," IEEE Transactions on Network and Service Management, vol. 18, no. 1, pp. 63–78, 2021.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020

[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2023.

[6] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," 2021.

[7] R. by SUSE. Lightweight kubernetes. [Online]. Available: https://k3s.io/

[8] S. Shen, V. Van Beek, and A. Iosup, "Statistical characterization of business-critical workloads hosted in cloud datacenters," in 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2015, pp. 465–474.

[9] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing crossdimension dependency for multivariate time series forecasting," in The Eleventh International Conference on Learning Representations, 2023. [Online]. Available: https://openreview.net/forum?id=vSVLM2j9eie

EMERGING TECH CONFERENCE 2024
Edge Intelligence

Papers

Session 1.2 | Internet-of-Things & AI

Session Chairs: **Gregory DOUMENIS & Sotiris BANTAS**

Short-term weather forecasting in maritime environments exploiting a ML-based model at the Edge
Vasiliki Naskari, Ioannis Masklavanos, Stavros Adam and Gregory Doumenis

Edge deep learning for low capabilities devices
Fotis Filippou, Fotis Foukalas and Theodoros Tsiftsis

Comprehensive comparison of YOLO-based object detection in edge applications: A use case on free parking spots
Vasilios Karvelas and Christoforos Kachris

Computing the Cleanness of the Photovoltaic (PV) Panels
Alexandros Tsagkaropoulos, Christoforos Vasilakis, Georgios Venitourakis, Tzouma Amrou, Georgios Konstantoulakis, Panagiotis Golemis and Dionysios Reisis

HETiA
EXPANDING THE DIGITAL FRONTIERS

UNIVERSITY OF THESSALY
FOUNDED 1984

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 39-45

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Short-term weather forecasting in maritime environments exploiting a ML-based model at the Edge

Vasiliki Naskari[1‡‡], Ioannis Masklavanos, Stavros Adam and Gregory Doumenis
[1] *Department of Informatics and Telecommunications, University of Ioannina, Arta, Greece*
naskvasil@uoi.gr

Abstract

This study deals with short-term weather forecasting in maritime environments using machine learning. Data from autonomous ship sensors, such as, temperature, humidity, wind speed, electrostatic field, and sea conditions, were used to enhance forecast accuracy. A Radial Basis Function (RBF) model was developed to tackle nonlinear structure of the data. Data preprocessing and model development were performed using MATLAB, while model evaluation was conducted in Python. The dynamic modeling approach that was adopted demonstrated significant accuracy improvements over static models, indicating the potential for enhancing maritime safety and operational efficiency with automated, energy-efficient systems.

1 Introduction

1.1. Background

Marine weather forecasting is crucial for safeguarding seafarers and optimizing maritime operations. Accurate forecasts prevent accidents, guide navigation, and conserve shipping resources. Traditionally, expert forecasters manually compile detailed bulletins on wind, sea states, visibility, and significant weather events, usually covering large areas. Data science and machine learning advancements now enable potentially more accurate, automated forecasting processes. Such processes utilize analysis of data from sensors on ships, capturing real-time metrics like pressure, temperature, humidity, wind speed, and sea conditions (K.U. Jaseena, 2022), (EFOS, n.d.). Integrating these sensors with machine learning algorithms intends to boost the creation of automated weather prediction systems which process data with high accuracy and minimal delay (Aakash Parmar, 2017). Weather forecasting for maritime operations faces challenges from the unpredictable nature of wind and waves. Researchers and practitioners have developed various methods to tackle these challenges, categorized into physical, statistical, and machine learning approaches. Physical methods use meteorological data and physical laws to model weather conditions, often applying empirical models for wave and wind predictions (Foley Aoife M, 2012) (Kazeminezhad M.H., 2005). Statistical methods analyze historical data to find patterns, effective for trends but limited by environmental unpredictability (Zuluaga Carlos D., 2015) (Lydia M, 2016) (Poggi P., 2003). Machine

‡‡‡‡ Corresponding author

learning approaches while adopting the statistical processing view, however, they process large datasets to uncover complex, dynamic relationships, offering flexibility and superior accuracy over traditional methods (Mengning Wu, 2019). By integrating these diverse approaches, researchers strive to enhance forecast reliability and maritime safety (Özger Mehmet, 2007) (Chang G.W., 2017) (Berbić Jadran, 2017).

1.2. Problem Statement

Our aim is to develop a system able to provide short-term local weather predictions (nowcasts), with emphasis on abrupt/adverse phenomena. We differentiate from similar studies, by adding the Static Atmospheric Electric Field measurement in our dataset. Additionally, the study seeks to improve the entire data pipeline, from collection to analysis. By implementing a tailored data-cleaning technique and developing an advanced machine learning algorithm, we aim to enhance the accuracy and reliability of weather predictions, crucial for safe maritime operations.

2 Methodology

2.1. Data Collection

The initial phase of our analysis involved data cleaning to ensure the reliability and accuracy of our data, which was sourced from the Azure Databricks platform (Microsoft, n.d.). This platform provided a scalable environment essential for handling large datasets. We focused on key meteorological variables critical to our predictive model, including atmospheric pressure, humidity, temperature, electrostatic field, wind speed, weather conditions, and timestamps. During data cleaning, we checked each data point for validity and consistency, removing inconsistencies or irrelevant information that could affect our analysis. Measurement errors or outliers were corrected to maintain the dataset's integrity. For gaps in data, we used linear interpolation for short interruptions and removed any segments where data was missing for longer than five minutes to preserve dataset coherence. Categorical data were converted into numerical formats to facilitate detailed analysis.

Subsequently, we conducted a correlation analysis to identify relationships between the variables using Pearson's and Spearman's correlation coefficients. This step aimed to detect both linear and non linear relationships within the data. The initial correlation results were weak, prompting further investigations using dynamic time warping, which confirmed the presence of non-linear relationships critical for developing models to accurately predict complex weather patterns. The data, summarized in Table 1, include:

Parameter	Unit	Measurement Range
Static Atmospheric Electric Field	KV/m	-20 - 20
Air Humidity	%	0 - 100
Air Temperature	°C	-30 - 60
Air Pressure	hPa	800 – 1100
Wind True Speed	m/sec	0 - 20
Weather Condition	-	Discrete Categories
Timestamp	-	Continuous Time

Table 1: Timeseries characteristics

This structured approach ensures an analysis framework, leveraging both traditional statistical techniques and advanced machine learning in a MATLAB environment to enhance the accuracy and depth of our weather forecasting models.

2.2. Systemic approach

To address the challenges of analyzing time series data and predicting the evolution of phenomena captured by these series, various techniques from classical statistics, such as linear and non-linear regression, have been proposed. However, our initial evaluations using these methods did not yield satisfactory results. In contrast, the field of machine learning offers numerous effective approaches for predicting system behavior, particularly when non-linear relationships exist between input data and the phenomena under study. Techniques such as Artificial Neural Networks, Bayesian Models, kernel machines, and Gaussian processes have shown particular promise. These advanced techniques, regardless of their origin— statistical learning or biological learning—have demonstrated exceptional performance in handling data analysis problems, especially those where data is infected with noise, errors, or missing values. Bayesian methods, although effective in many applications, did not seem adequate for our case due to unknown a priori distributions of input data.

Neural networks have been successfully applied to various problems, including pattern recognition, classification, regression, and dynamic system modeling. Our literature review on best practices for using neural networks revealed that radial basis function networks (RBF networks) effectively meet our requirements. These networks are feed-forward architectures similar to Multi layer Perceptrons (MLP) but with the added capability of processing inputs based on kernel functions.

2.3. Experimental Setup

Upon completing the data pre-processing stage, the next phase involved selecting the appropriate approach for applying the predictive algorithm to forecast weather conditions in maritime contexts. Two approaches were considered: a static approach and a dynamic approach. In the static approach, the predictive model is trained once using the available data. This method offers several advantages such as computational efficiency, as it minimizes the need for repeated model adjustments, thereby conserving resources and time. Suitability for scenarios where the underlying data distribution remains relatively stable over time, reducing the need for frequent updates. However, the static approach has notable limitations. It is highly sensitive to concept drift, where the underlying data distribution changes over time, rendering the model outdated and less accurate. This limitation prevents the model from adapting to new patterns and trends that may emerge in the data, thus compromising its robustness in dynamic environments. The inability to respond to evolving information significantly impairs the effectiveness of static models, especially in areas characterized by constant change.

Conversely, the dynamic approach involves periodic retraining of the predictive model to incorporate new data. This methodology's primary strength lies in its adaptability to changing data distributions and evolving patterns. By regularly updating the model with new information, the dynamic approach ensures that the predictive model remains aligned with the underlying data, thereby enhancing its predictive accuracy.

3 Case Study: Weather Condition Forecast

3.1. Introduction to the Case Study

Accurate weather forecasts are essential for maritime operations, as they directly affect navigation safety, route planning, and overall efficiency. The "weather condition" metric was selected for prediction due to its comprehensive nature, which helps in making more informed maritime decisions. This case study focuses on demonstrating how dynamic predictive models can be used to forecast weather conditions at sea, comparing their accuracy to traditional methods, and exploring how automated forecasts can improve maritime safety and operational efficiency.

The analysis method used is highly flexible and can be applied to any subset of weather data. The process involved a rolling window approach, where each window contained five minutes of data. We calculated key signal statistics of the electrostatic field for each window, including arithmetic mean, sign of the mean, signal energy, variance, skewness and curvature. These statistics were then combined with correlated variables from other weather data.

To prepare the data for forecasting, we shifted the historical weather condition data by 60 positions (equivalent to 60 minutes). This transformation helps train the model to predict future weather conditions, adjusted to the time interval defined by the shift. A representation of this process is illustrated in Figure 1.

		Time				
		n	n+1	n+2	n+3	...
Air Relative Humidity		.	.	.	.	.
Air Temperature		.	.	.	.	.
Air Pressure		.	.	.	.	.
Wind True Speed		.	.	.	.	.
Static Atmospheric Electric Field Statistics	Power	.	.	.	.	.
	Mean	.	.	.	.	.
	Sign	.	.	.	.	.
	Variance	.	.	.	.	.
	Skewness	.	.	.	.	.
	Curvature	.	.	.	.	.
Weather Condition		.	.	.	.	.
Shifted Weather Condition		Weather Condition(n+60)	Weather Condition(n+61)	Weather Condition(n+62)	Weather Condition(n+63)	

Figure1: Preprocessed Data Table

Given the challenges discussed in Section 2.2, along with the test results and data characteristics, we chose to develop a Radial Basis Function (RBF) model with a Kernel filter. This model was selected for its ability to handle non-linear relationships and provide accurate predictions, making it ideal for forecasting weather conditions in dynamic maritime environments.

4 Results

In the static approach, we forecasted the weather condition metric 60 minutes into the future. This method yielded an accuracy of **57.32%.** While this accuracy highlights the model's ability to provide general trends, it lacks the precision of dynamic models, as reflected by the comparative results.

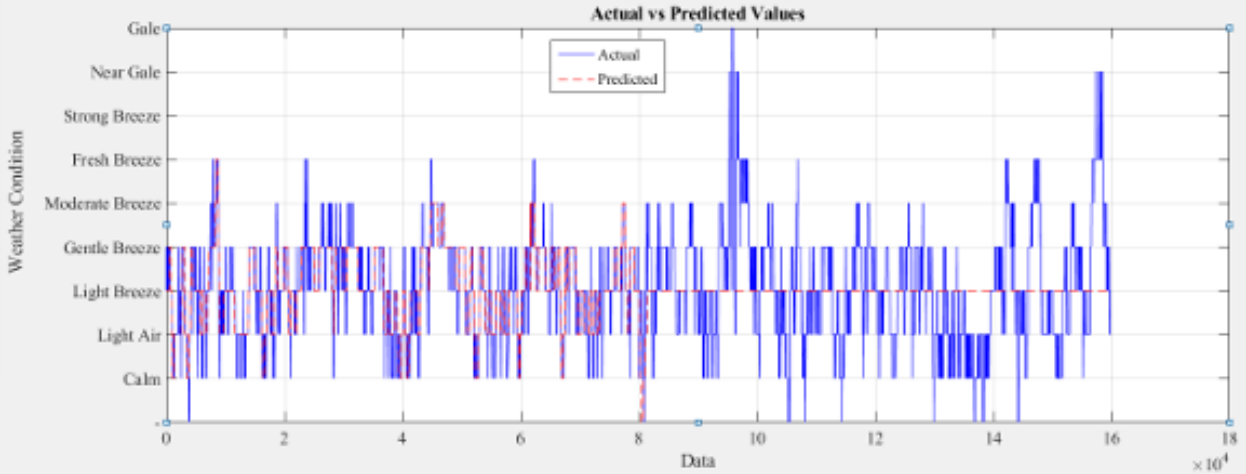


Figure2: Accuracy Results for The Static Model

1st dynamic approach (Prediction 60min Forward for a 20min Forecast Period)

In this dynamic approach, we predicted the weather condition 60 minutes ahead, focusing on a forecast window of 20 minutes. This method significantly improved prediction accuracy, achieving **76.98%.** The model was better at capturing fluctuations in the weather conditions over shorter forecast periods.

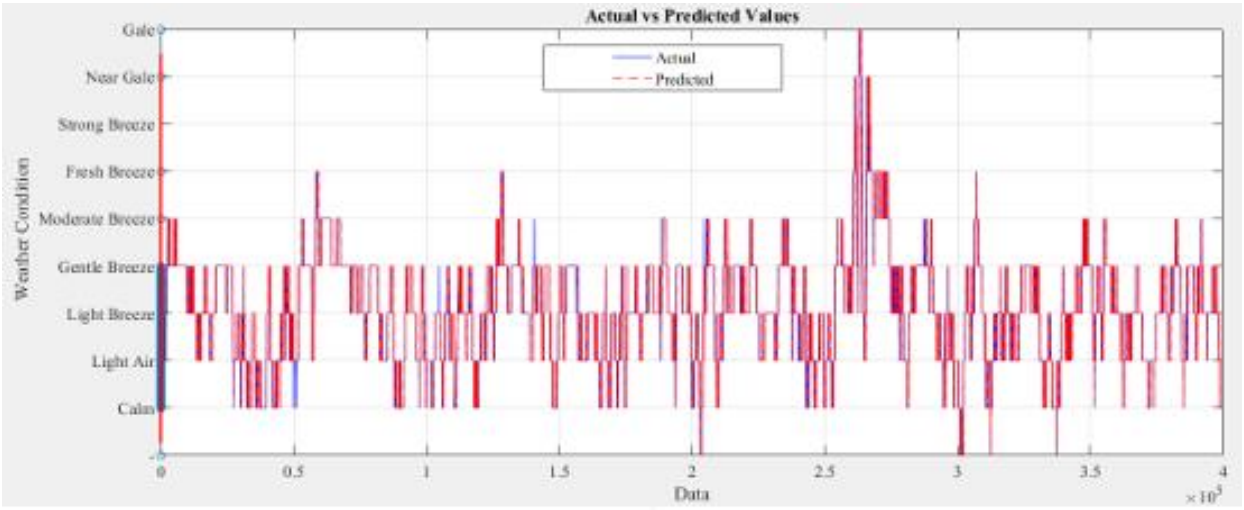


Figure3: Accuracy results for the 1st dynamic approximation model

2nd dynamic approach (Prediction 60min Forward for a 15min Forecast Period)

Further refining the forecast window to 15 minutes led to the highest accuracy of **83.17%.** This model demonstrated a strong capability in accurately forecasting short-term weather conditions, making it

the most effective approach in our study.

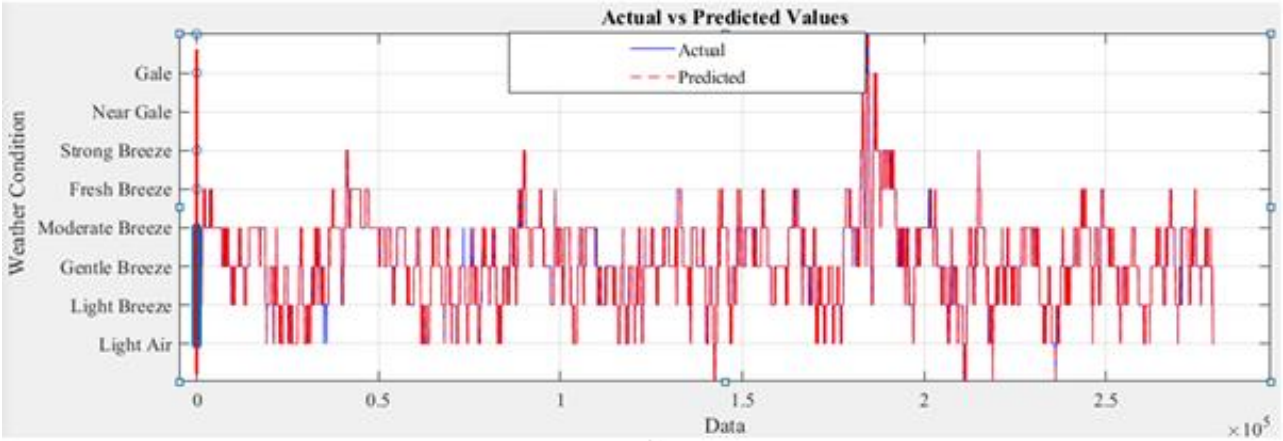


Figure4: Accuracy results for the 2nd dynamic approximation model

The model was tested in Python, and the following metrics were measured: accuracy, mean squared error (MSE), root mean squared error (RMSE), and correlation coefficient. Below is a comparative table of the results for the static approach and the two dynamic models, with the dynamic models showing considerable improvements across all evaluation metrics.

Approach	**Static**	**20min (1st approach)**	**dynamic**	**15min (2nd approach)**	**dynamic**
Accuracy (%)	49.4972		76.98		83.17
MSE	1.1145		0.2555		0.1912
RMSE	1.0557		0.5054		0.4372
Cor. Coefficient	0.493648		0.9014		0.9276

Table 2: Performance evaluation

4.1. Discussion

The study revealed that dynamic predictive models, which incorporate periodic retraining, significantly outperform static models in both accuracy and adaptability. Their ability to integrate new data ensures precise predictions, even in changing weather conditions, making them highly effective for maritime operations. Additionally, the low energy requirements of these models make them ideal for use in resource-limited maritime environments. This energy efficiency allows continuous operation without straining ship power systems. Accurate weather forecasting not only enhances safety by providing timely warnings but also optimizes route planning, reducing fuel consumption and operational costs. The combination of these factors supports more efficient, resilient, and cost effective maritime logistics.

5 Conclusions

This paper presents an application of dynamic, non-linear predictive modeling on a weather dataset employing (static) atmospheric electric field measurements. The application of RBF based predictive models showed notable improvements particularly in terms of accuracy and adaptability. The

execution of the model is particularly efficient in terms of time highlighting its potential for integration into energy-efficient forecasting systems. Using more than 18 months of historical data, the system exhibited over 83% nowcast accuracy. However, a key limitation is the model's inability to accurately track fast and abrupt weather changes.

Thus, having the dynamic model perform well in the short term/low variance conditions, future research should explore hybrid models that incorporate large historic datasets and real-time data assimilation for greater accuracy. Our work could set the stage for advancing towards hybrid predictive models, leveraging the density of measuring stations via federated learning. The system architecture, designed to support both edge and cloud computing, enhances scalability and real-time forecasting capabilities. This flexibility will further strengthen the accuracy and efficiency of maritime weather forecasting, paving the way for more robust and reliable systems in the future.

References

[1] Aakash Parmar, K. M. (2017). Machine Learning Techniques For Rainfall Prediction: A Review. International Conference on Innovations in information Embedded and Communication Systems (ICIIECS).

[2] Berbicć Jadran, O. E. (2017, July). Application of neural networks and support vector machine for significant wave height prediction. Oceanologia, pp. 331-349.

[3] Chang G.W., L. H. (2017). An improved neural network-based approach for short-term wind speed and power forecast. Renewable Energy, pp. 301-311.

[4] EFOS. (n.d.). Retrieved from https://efos.site/

[5] Foley Aoife M, L. P. (2012, January). Current methods and advances in forecasting of wind power generation. Renewable Energy, pp. 1-8.

[6] K.U. Jaseena, B. C. (2022, June). Deterministic weather forecasting models based on intelligent predictors: A survey. Journal of King Saud University - Computer and Information Sciences, pp. 3393-3412.

[7] Kazeminezhad M.H., E.-S. A. (2005, October). Application of fuzzy inference system in the prediction of wave parameters. Ocean Engineering, pp. 1709-1725.

[8] Lydia M, S. K. (2016, March 15). Linear and non-linear autoregressive models for short-term wind speed forecasting. Energy Conversion and Management, pp. 115-124.

[9] Mengning Wu, C. S. (2019). Prediction of short-term wind and wave conditions for marine operations using a multi-step-ahead decomposition-ANFIS model and quantification of its uncertainty. Ocean Engineering.

[10] Microsoft. (n.d.). Azure Databricks. Retrieved from https://azure.microsoft.com/en-us/products/databricks.

[11] Özger Mehmet, Ş. Z. (2007, December). Prediction of wave parameters by using fuzzy logic approach. Ocean Engineering, pp. 460-469.

[12] Poggi P., M. M. (2003, December). Forecasting and simulating wind speed in Corsica by using an autoregressive model. Energy Conversion and Management, pp. 3177-3196.

[13] Zuluaga Carlos D., Á. M. (2015, October). Short-term wind speed prediction based on robust Kalman filtering: An experimental comparison. Applied Energy, pp. 321-330.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Page 46 – 52

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Edge deep learning for low capabilities devices

Fotis Filippou[1], Fotis Foukalas[1] and Theodoros Tsiftsis[1]
[1] *Department of Informatics and Telecommunications, Lamia, University of Thessaly*
fotisfilippoy@gmail.com, ffoukalas@uth.gr, tsiftsis@uth.gr

Abstract

Nowadays, Machine Learning (ML) is being used to construct many applications in domains such us object detection, image classification, speech to text etc. Deep Neural Networks (DNNs) are the core of Machine Learning as they offer remarkable accuracy and performance across various tasks. Despite their powerful capabilities, DNNs often require substantial computational resources, which can be challenging to manage, especially when deploying them on edge devices. So, these models have to be optimized before being deployed to these devices. Optimizing a model means making it smaller and more efficient without losing too much performance. Even though techniques like pruning reduce the number of parameters, the goal is to keep accuracy and speed as close as possible to the original. We are going to present a hybrid solution combining two techniques, pruning and quantization. Pruning is the process of eliminating inessential weights and connections in order to reduce the model size. Once the unnecessary parameters are removed, the model is quantized by converting the weights of the remaining parameters from 32 floating point precision to half. We verify and validate the performance of this hybrid approach using the COCO dataset (contains 80 classes) and the pre-trained YOLOv8 model. At the final stage, the hybrid model is deployed on an edge device, the NVIDIA Jetson Nano (4GB).

1 Introduction

In the last decade, Machine Learning has been the core for our everyday life. DNNs have a big impact on the performance in many applications such as computer vision, natural language processing (NLP), speech recognition etc. (1) (2). Despite of this enormous success of DNNs there are several problems that are being considered about the deployment of such models in computation-constrained environments (mostly on edge devices). These problems are usually created by the large model size and the high computational cost which requires in most cases an edge device with a “strong” GPU in order to avoid these problems (2). So, it is very crucial to compress these neural networks in order to reduce their size and computational demands, making them more suitable for deployment in edge environments.

Accomplishing these compression goals, requires some techniques that are being used more and more in the last years and expound a big interest in the ML community. Based on their properties, these techniques are split to four categories (3):

1 pruning and quantization

2 low-rank factorization
3 transferred convolutional filters
4 knowledge distillation

Parameter pruning and quantization techniques focus on identifying and eliminating redundant or non-essential model parameters. Low-rank factorization methods leverage matrix and tensor decomposition to approximate the most informative parameters in deep neural networks (DNNs). Transferred or compact convolutional filter approaches design specialized convolutional filters to reduce the parameter space, thereby conserving storage and computational resources. Knowledge distillation methods involve training a smaller, more efficient neural network to replicate the output of a larger model. (3)

In this paper, we are planning to examine a hybrid solution for the YOLO model using a hybrid approach that includes pruning and quantization. This is done in order to reduce the model size and number of parameters without compromising the model performance (4). A similar work has been done from the authors of paper (4) where they have applied this hybrid approach in 3 DNNs (ResNet-56, ResNet-110 and GoogLeNet), that are used for image classification. On an average, the authors have achieved to reduce the number of flops and parameters by 40% and simultaneously loosing not more than 2% in accuracy of the model. The difference in our work is that we use this hybrid approach on YOLO model which is not an image classification model but an object detection model.

This paper is organized into 4 sections. The first one is the introduction, that is already described. The section 2 focuses on the previous works on model compression. Section 3 presents the whole pipeline (Figure 1) that has been implemented, started from training till deploying. Last, Section 4 includes the conclusions of the method that has been followed and future work.

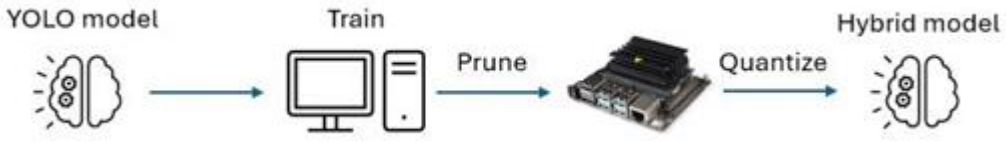


Figure 1: Pipeline of the hybrid technique

2 Related Work

The authors of paper (5) have investigated of complex DL models on optimizing their functionality on an embedded device, particularly on the NVIDIA Jetson Nano. They have measured the performance of image classification and video action detection models based on the inference speed and their experimental results have shown that the optimized models exhibit on average 16.11% speed improvement over their original edition. More specifically, they convert a PyTorch model into a TensorRT engine within 2 steps. Initially, the PyTorch model is being exported into ONNX (Open Neural Network Exchange) format. Next, from the ONNX file, the TensorRT engine is built using the TensorRT Python API. This involves several stages, including creating a network definition, importing the model through the ONNX parser, and building the TensorRT engine with a builder. Moreover, the inference process with the TensorRT engine file includes 6 steps: (5)

1. creation of an inference execution context
2. memory allocation for input and output on the CUDA device
3. input data is transferred from the host into the input memory allocated on the CUDA device

4. TensorRT engine performs inference using the asynchronous execute API
5. the output is transferred back into the host memory
6. the stream used for data transfers and inference execution is synchronized to ensure the completion of all operations

On another paper (6), the authors have focused on Deep Learning Model Optimization (DLMO). They present a usage strategy of DLMO based on the performance evaluation through light convolution, quantization, pruning techniques and knowledge distillation, known to be excellent in reducing memory size and operation delay with a minimal accuracy drop and through experiments on image classification they derive possible and optimal strategies to apply deep learning into edge devices. More specifically, they explore strategies for utilizing DLMO by evaluating its performance across various combinations of Lightweight Convolution, Quantization, and Pruning techniques. They use the Canadian Institute For Advanced Research 10 (CIFAR10) and CIFAR100 datasets for the whole research. Also, they use VGGNet and ResNet as baseline networks and for comparison the lightweight MobileNet v1, v2 and v3 networks. So, in different use cases they evaluate the performances of several quantization techniques, which comprise Quantization Aware Training (QAT) and subtypes of Post Training Quantization (PTQ), i.e., Baseline Quantization (BLQ), Full Integer Quantization (FIQ) and Float 16 Quantization (F16). For the pruning technique, the performance improvement is analyzed by applying the training method to the basic Convolution Neural Network (CNN) and lightweight CNN technologies. By the finish of the whole process and the necessary tests, they conclude that quantization was the most effective in compressing the model size, but it introduced significant delays in data conversion. Also, the pruning technique was excellent in all aspects of model compression, accuracy loss minimization and delay minimization. Additionally, it is recommended to train the deep learning models using knowledge distillation, as this approach can enhance accuracy without causing any additional increases in latency or model size. Last, by classifying Artificial Internet of Things services according to three performance factors (accuracy, size, and delay), they identified the optimal combinations of DLMO techniques for different scenarios.

3 Pipeline

Every day, we watch more and more applications that are using DNNs to have a big impact on our life. Besides the use of DNNs in large computer systems, DNNs are used on many low performance edge devices such as smartphones. So, these models have to be compressed in order to be capable of running is such devices. The model that we have focused on for this paper is the YOLO and more specifically the version 8 of it. YOLOv8's architecture is an evolution of previous YOLO models, utilizing a convolutional neural network divided into two main parts: the backbone and the head. The backbone is based on a modified version of the CSPDarknet53 architecture, consisting of 53 convolutional layers enhanced with cross-stage partial connections. The head comprises multiple convolutional layers followed by fully connected layers responsible for predicting bounding boxes, objectness scores, and class probabilities. Notably, YOLOv8 integrates a self-attention mechanism in the head of the network and a feature pyramid network for multi-scaled object detection, enabling it to focus on various parts of an image and detect objects of different sizes and scales (7).

We adopt the Ultralytics framework (8) in order to download YOLOv8 and make the necessary actions in order to compress it. It's also important to refer that YOLOv8 has 5 editions that differ on accuracy (mAP50-95) and speed. mAP50-95 is the average of the mean average precision calculated at varying

intersection over union (IoU) thresholds, ranging from 0.50 to 0.95 and it's a measure of the model's accuracy considering only the "easy" detections (8). For the whole pipeline we are going to use the COCO dataset which contains 80 pre-trained classes. We aim to test all of the above models with this dataset of how the perform before and after compression on an edge device, the Nvidia Jetson Nano (4GB). First of all, we have to train our model by loading each yolov8 pre-trained model and using the COCO dataset as parameter for the data. Also, we use 300 epochs for training for best results. Moreover, we apply parameters such as patience in order to prevent overfitting. The fact that training process requires high computational cost makes us clear that we should utilize a powerful GPU for the whole process in order to avoid extended training time.

3.1. Pruning

After the training process is completed, we have a ready to deploy model. So, after moving it to Jetson Nano we observe that the speed of inference is too slow and necessitates the use of model compression techniques. So, we will try to make the model faster and appropriate for the Jetson Nano by following a hybrid method that combines both pruning and quantization.

Model pruning refers to the act of removing unimportant parameters from a deep learning neural network model to reduce the model size and enable more efficient model inference. Generally, only the weights of the parameters are pruned, leaving the biases untouched (9). So, after identifying and eliminating unnecessary parameters, we take a lighter model that requires less computational power. It is particularly useful for deploying models on devices with limited resources such as Nvidia Jetson Nano. We can take more information about pruning and understand it better from Figure 2.

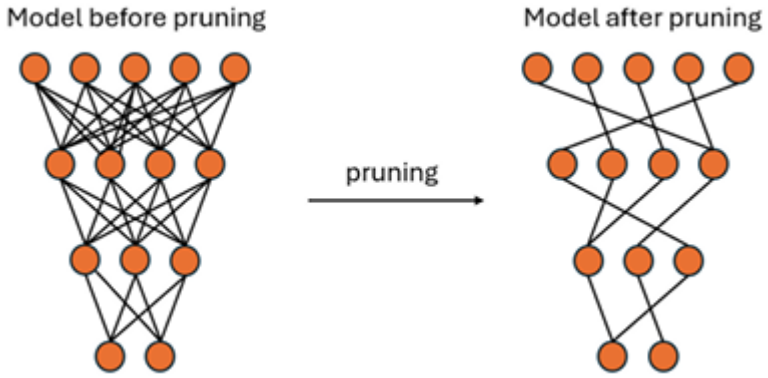


Figure 2: Pruning Process

We have created a custom function in order to prune our model. By this function, we can prune our model by an amount p (0<p

1 For each layer of the model, we check its type and if the layer is Conv2d we continue to step 3

2 Compute the L1-norm (unstructured) of the weight matrix of the layer (W) by the following equation:

$$||W_1|| = \sum_{i,j}|w_{i,j}| \qquad Equation\ 1$$

3 Sort the elements of W by their absolute values

4 Set the smallest weights to zero, based on the computed mask

5 Remove any applied masks to finalize the pruning process

6 Return the pruned model

3.2. Quantization

The next step of our hybrid technique, after pruning is done, is to quantize the pruned model. Quantization converts the model's weights and activations from high precision (like 32-bit floats) to lower precision (like 8-bit integers). By reducing the model size, it speeds up inference and loses a small percent in accuracy (8). We can see an example of Quantization in below figure.

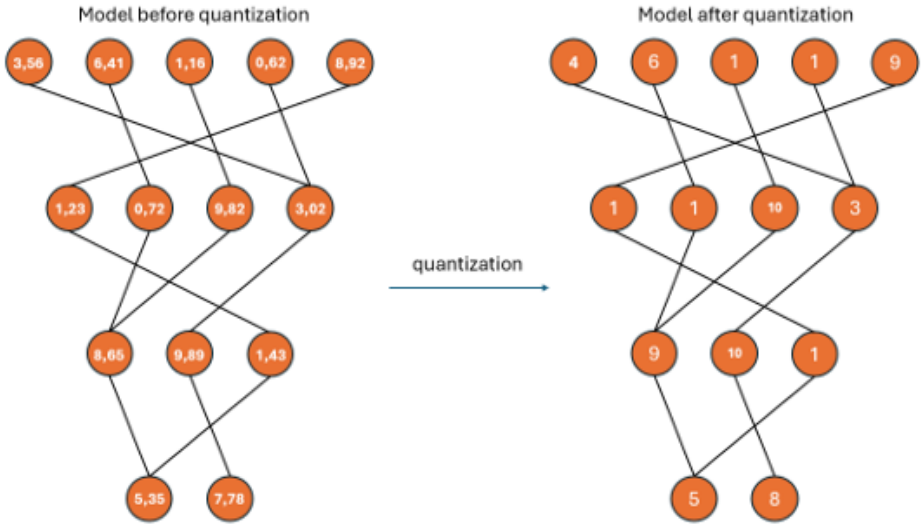


Figure 3: Quantization process

In order to implement quantization, we move our pruned model from the computer where it was trained and pruned to the Nvidia Jetson Nano. This is necessary because quantization needs to be performed on the edge device due to framework incompatibilities between different devices. The framework we use is TensorRT, which is specifically optimized for Nvidia devices like the Jetson Nano. TensorRT provides high-performance inference by leveraging the GPU capabilities of the Jetson Nano, and it supports various optimizations including precision calibration and layer fusion During the quantization process, the model is first converted to an ONNX file. Next, TensorRT generates a .engine file, which is the quantized model ready for deployment. There are several parameters that we can change on quantization like converting FP32 to FP16 or to INT8 which leads to a reduce on model size, lower power consumption and fast inference (8). In the figures we can see the results of mAP50-95 on figure 4 and of inference speed per frame (ms) on figure 5 for yolov8n, s and m versions before and after implementing the hybrid technique (5% pruned and quantized from FP32 to FP16) for Nvidia Jetson Nano.

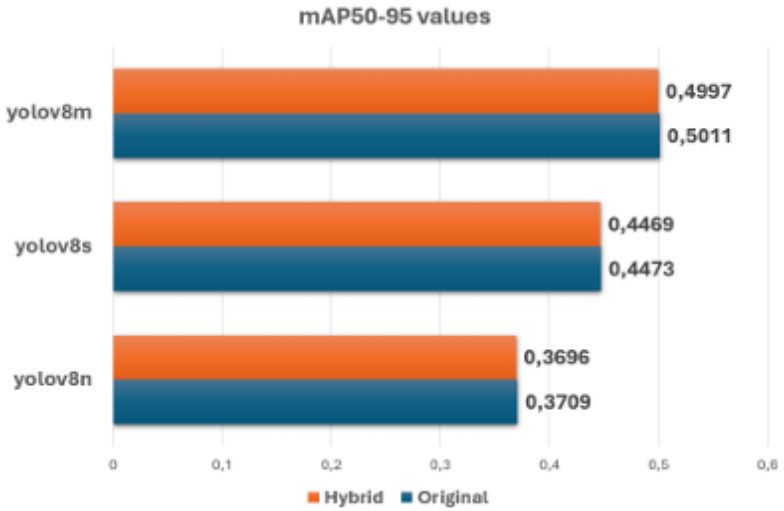


Figure 4: mAP50-95 for different version of YOLOv8 models on Jetson Nano

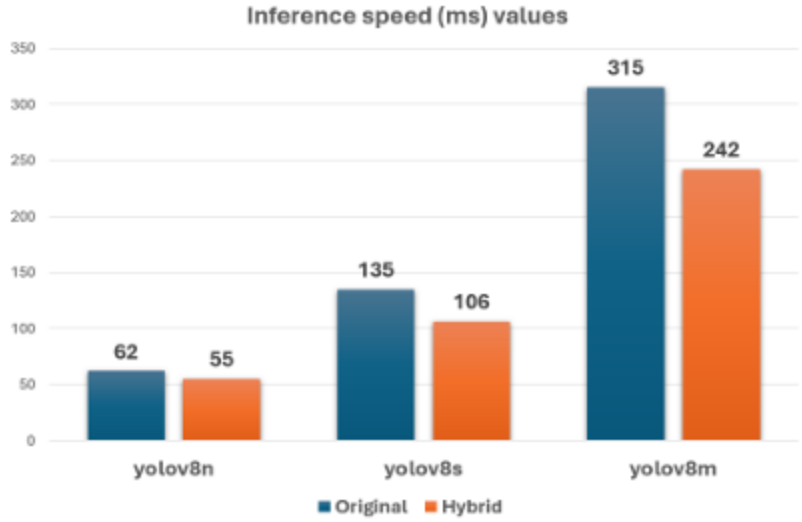


Figure 5: Inference speed (ms) for different version of YOLOv8 models on Jetson Nano

4 Conclusions and Future Work

By following this hybrid technique for different versions of YOLOv8 we can extract many conclusions based on results from figures 4 and 5. First, we can see a drop in mAP50-95 value by less than 0.5% which is normal to happen, due to pruning and quantization. Moreover, we observe that the inference speed is about 1.2X faster than original and by this fact the hybrid model can be easily deployed to edge devices. We should take in mind that we have converted FP32 to FP16(half) during pruning process whereas if we convert FP32 to INT8 we will see a bigger drop in mAP50-95 and a faster inference speed. Also, we have made all the tests for YOLOv8 models with COCO dataset but the whole pipeline can be applied to any PyTorch model that uses a pre-trained version of YOLO in order to be trained. This generalization is crucial for deploying models in diverse real-world scenarios where the ability to adapt to different data and tasks is essential. As edge computing becomes increasingly

prevalent, the importance of optimizing models for deployment in such environments cannot be overstated. The hybrid technique discussed here provides a robust framework for doing so, ensuring that models remain both accurate and efficient. As AI continues to expand into new domains and applications, techniques like these will be essential for making advanced models accessible and practical in a wide range of real-world scenarios.

For future work, advanced pruning strategies, like structured pruning or those guided by neural architecture search (NAS), could be investigated to remove redundant parameters more effectively, resulting in more compact models without significant accuracy loss. Also, another area for future research is dynamic quantization, where the model could adjust its precision based on the input complexity or available computational resources, thereby optimizing speed and efficiency. Expanding the evaluation of this technique across diverse datasets, especially those relevant to specific application domains like medical imaging or autonomous vehicles, will be crucial in assessing its generalizability.

Acknowledgment

This work has been funded by the Greek funded project "Connected smart cities for Greece 2.0 program" with code TAEDR-0536642.

References

[1] Kulkarni, Uday and Hosamani, Abhishek S and Masur, Abhishek S and Hegde, Shashank and Vernekar, Ganesh R and Siri Chandana, K. A Survey on Quantization Methods for Optimization of Deep Neural Networks. 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS). 2022, pp. 827-834.

[2] Peng, Hanyu and Wu, Jiaxiang and Zhang, Zhiwei and Chen, Shifeng and Zhang, Hai-Tao. Deep Network Quantization via Error Compensation. IEEE Transactions on Neural Networks and Learning Systems. 2022, pp. 4960-4970.

[3] Zhang, Yu Cheng and Duo Wang and Pan Zhou and Tao. A Survey of Model Compression and Acceleration for Deep Neural Networks. IEEE Signal Processing Magazine. 2020.

[4] Kulkarni, Narayan and Singh, Nidhi and Joshi, Yamini and Hasabi, Nikhil and Meena, S M and Kulkarni, Uday and Gurlahosur, Sunil V. Hybrid Optimization for DNN Model Compression and Inference Acceleration. 2022 2nd International Conference on Intelligent Technologies (CONIT). 2022, pp. 1-8.

[5] Tushar Prasanna Swaminathan, Christopher Silver, Thangarajah Akilan. Benchmarking Deep Learning Models on NVIDIA Jetson Nano for Real-Time Systems: An Empirical Investigation. June 26, 2024.

[6] Lee, Hyungkeuk and Lee, NamKyung and Lee, Sungjin. A Method of Deep Learning Model Optimization for Image Classification on Edge Device. Sensors. 2022.

[7] Modelbit. [Online] https://www.modelbit.com/model-hub/yolo-v8-model-guide.

[8] Qiu, Glenn Jocher and Ayush Chaurasia and Jing. Ultralytics YOLOv8. [Online] 2023. https://github.com/ultralytics/ultralytics.

[9] Neo, Marcus. Datature. [Online] February 29, 2024. https://www.datature.io/blog/a-comprehensive guide-to-neural-network-model-pruning.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 53 – 58

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Comprehensive comparison of YOLO-based object detection in edge applications: A use case on free parking spots

Vasileios Karvelas and Christoforos Kachris
University of West Attica Department of Electrical and Electronics Engineering Athens, Greece

Abstract

When it comes to edge applications that require object detection based on computer vision, there are several models that can be used. However there is not an easy way to identify what is the best model for each application. In this paper we present a com prehensive comparison of several YOLO-based models for object detection targeting edge applications using single board computers (Raspberry pi). We compare several versions of YOLO (YOLOv5, v8 and v10) in terms of accuracy, inference time and training time. For the specific comparison we evaluate the performance for an edge application where the system has to identify the empty parking spots in a parking lot. The comparison allows developers to select the best model based on the application requirements.

1 Introduction

Object detection is a fundamental task in computer vision that involves identifying and locating objects within an image or video. Unlike image classification, which simply categorizes an image as a whole, object detection not only determines the presence of objects but also provides their precise positions through bounding boxes. This capability is crucial for a wide range of applications, from autonomous driving to real-time video surveillance.

Among various approaches to object detection, single shot models have gained significant attention due to their efficiency and speed. Unlike multi-stage detectors that rely on multiple passes over the image, single shot models such as SSD (Single Shot MultiBox Detector) [3] and YOLO (You Only Look Once) [6] predict bounding boxes and class probabilities in a single forward pass through the network. This allows for real-time performance, making them ideal for time-sensitive tasks.

YOLO-based models, in particular, is a widely-used model in the domain of real-time object detection. Starting from the original YOLO model in 2015 [6], which introduced the concept of treating detection as a regression problem, these models have evolved through several iterations, each improving in accuracy, speed, and adaptability.

The idea is to segment the image into a grid and predict the existence of a bounding box for each of the classes we are considering. When it comes to labeling the data, a grid cell is labeled to contain an object only if the center of the box is in it. If the grid cell contains a center, the ”objectness” is labeled 1 and 0 otherwise. The model will try to predict the probability that a grid cell contains a center. If it contains a center or multiple ones, each class related to those centers will be labeled with 1s.

The success of YOLO-based models lies in their ability to balance computational efficiency with detection precision, pushing the boundaries of what is achievable in real-time object detection.

In this paper we present a comparison of several YOLO versions in terms of accuracy, inference througput and training time for a simple application ported to a Raspberry Pi board for edge applications. Specifically we compare several models for an application that need to identify the empty spots in a parking lot. First we train the model using a large dataset of parking areas and then we evaluate the performance of the models on inference in terms of accuracy, latency and training time.

2 Related work

There are several papers that have been presented looking to compare object detection models. For example, in [4] there is a comparison on the detection accuracy and speed measurements of several state-of-the-art models such as RetinaNet, GHM Faster R-CNN, Grid R-CNN, Double Head R-CNN , and Cascade R-CNN-for the task of object detection in commercial EO satellite imagery.

In [1] a comparison of YOLO-based models has been performed. However, the specific paper only compares the different versions of the YOLOv5 model and does not examine later YOLO models.

In [2], they compare several models for computer vision but the main focus is on privacy exploring the impact of data anonymization on deep learning models.

Naftali et al. [5] have also presented a comparison for object detection algorithms for street-level objects. This paper compares various one-stage detector algorithms such as SSD MobileNetv2 FPN-lite, YOLOv3, YOLOv4, YOLOv5l, and YOLOv5s for street-level object detection within real-time images. However it does not compare the latest versions of YOLO model.

As YOLO-based object detection has emerged as one of the most promising model for edge applications, we compare the performance of several newer versions in terms of accuracy, inference latency and training time for edge applications and specifically for a widely-used typical Raspberry pi 4 board.

3 Comparison of the models

For the comparison, we selected a typical edge application that tries to identify the number of empty spots in a parking lot. The system takes as input the feed from a camera in a parking lot and tries to identify the empty parking spots. The training of the model has been performed using a dataset of 1600 images of parking lots. Out of the 1600 images, 1200 of them (75% of the images) were used for the training part, 200 images were used for the validation and 200 images for the testing.

Since the comparison is targeting edge applications that are using low-power processors, only the smaller versions of the YOLO models have been studied (nano and small versions) for the comparison. Also we evaluated the most widely used versions of YOLO models and specifically Yolov5, Yolov8 and Yolov10.

- **YOLOv5:** A highly popular version of YOLO (You Only Look Once) developed by Ultralytics, optimized for speed and accuracy in object detection, widely used in various computer vision applications.

- **YOLOv8:** The latest official release from Ultralytics, YOLOv8 builds upon previous versions with improved architecture, better performance, and a user-friendly interface for training, deployment, and inference.

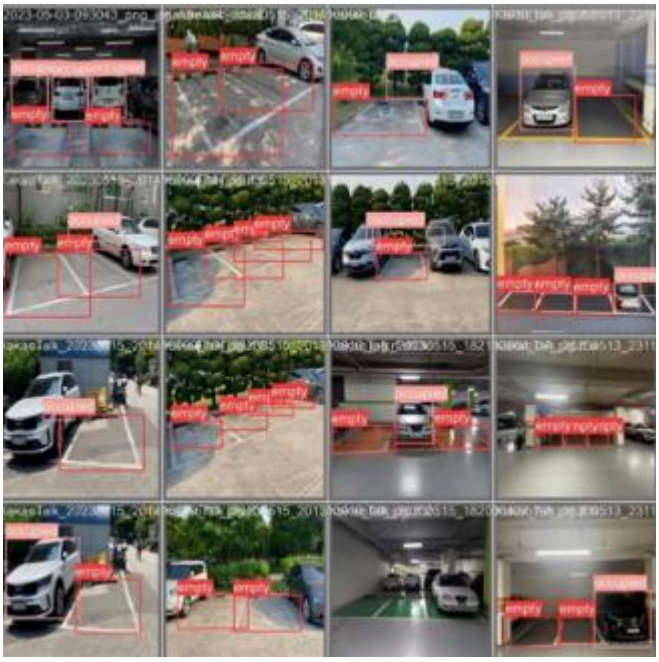


Figure 1: Identifying the empty spots in a parking lot using YOLOv5n

Model	**mAP50**	**Precision**	**Recall**
YOLOv5n	0.971	0.976	0.946
YOLOv5s	0.971	0.976	0.954
YOLOv8n	0.977	0.969	0.963
YOLOv8s	0.976	0.976	0.958
YOLOv10n	0.970	0.973	0.952
YOLOv10s	0.971	0.974	0.959

Table 1: Comparison of mAP, accuracy and recall for various models

- **YOLOv10:** YOLOv10 builds upon the strengths of its predecessors, addressing the limi tations in post-processing and model architecture. YOLOv10 allows for better adaptation in various real-world scenarios, from lightweight models for mobile applications to larger models for high-performance tasks.

Figure 1 shows a case where the YOLOv5n has been used to identify the empty parking spots after the training of the system.

The training of the models has been performed using Google Colab. An Intel CPU clocked at 2.2 GHz has been used coupled with a Tesla K80 GPU with 12 GB DDR5 RAM.

Table 1 shows the accuracy of the training that has been achieved for several versions of the YOLO models. As is it shown for the specific application on edge, the accuracy is quite high for each version. This may be due to the fact that the images form the camera are quite clear and there are not many

objects that move very fast.

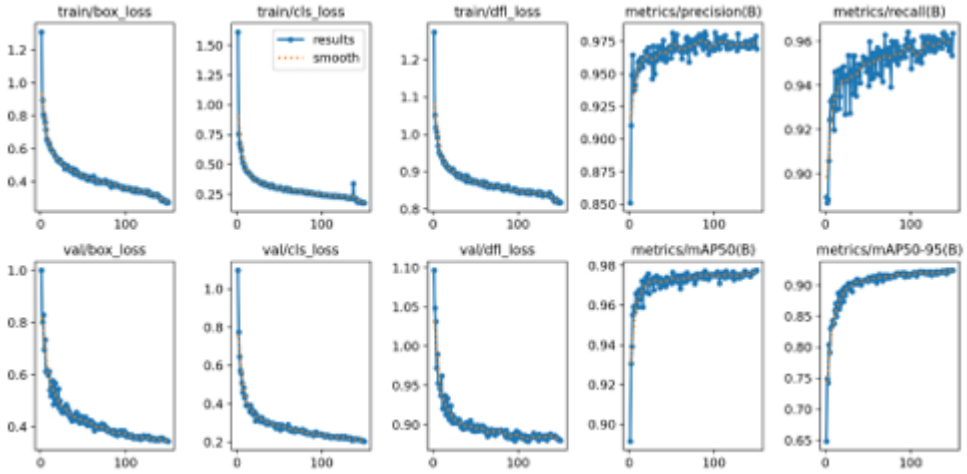


Figure 2: Results of the training using YOLOv8n

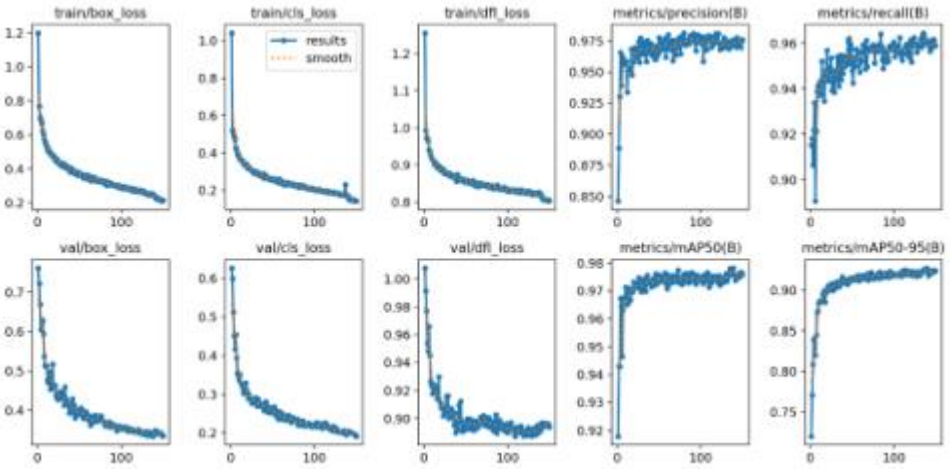


Figure 3: Results of the training using YOLOv8s

Figure 2 and Figure 3 show the results of the training for each model.

Figure 4 shows the training time for each model. As it is shown the latest version of the model require almost 40% higher training time compared to the early versions of the models.

Figure 5 shows the model size for each model. As it is shown the latest versions of the model usually are much larger than the early versions of the model. The most efficient one in terms of model size is the YOLOv5n that uses only 3.73MB and it makes it easier to fit into embedded systems with limited memory and storage.

Table 2 shows the comparison of the YOLO version in terms of the inference time, the pre processing and the post-processing. As it is shown the earlier versions of the model offer much lower inference latency compared to the latest versions. YOLOv5n takes only 584 ms for the inference on a raspberry pi 4 board while YOLOv10s takes around 2200ms for the inference. For the specific application of

identification of empty spots the requirements in terms of inference is around 1 fps.

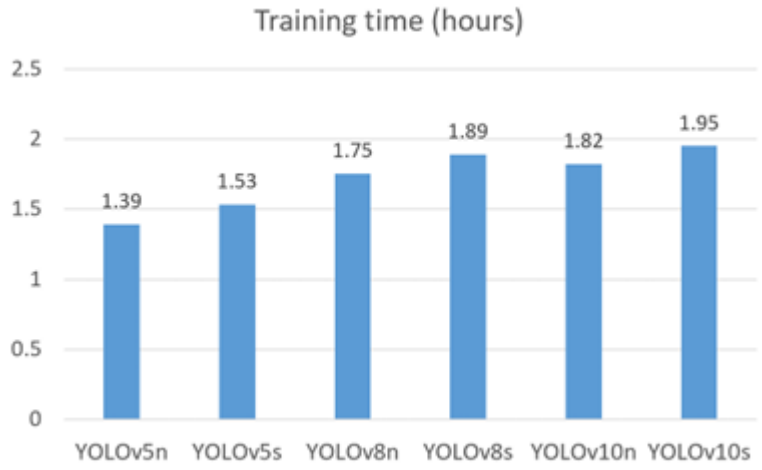


Figure 4: Comparison in terms of training time for the specific dataset

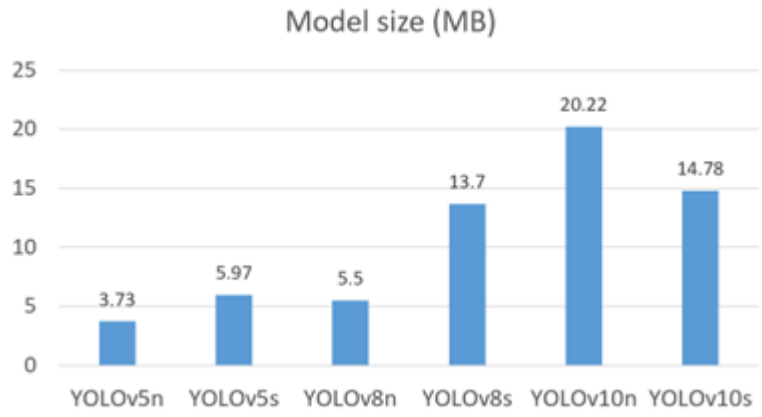


Figure 5: Comparison of the size model for the specific dataset

That means that a raspberry pi 4 could support up to 2 cameras using a YOLOv5n model.

Model	**Postprocess (ms)**	**Inference (ms)**	**Preprocess (ms)**
YOLOv5n	3.8	584	34
YOLOv5s	5.5	1405	34.3
YOLOv8n	3.2	932	15.3
YOLOv8s	3.3	1672	5.8
YOLOv10n	2.6	1075	9.6
YOLOv10s	8.1	2201	8.1

Table 2: Comparison of pre-process, inference time and post-process for various models

4 Conclusions

Based on these results it is shown that for edge applications targeting single board computers like

Raspberry pi, earlier versions of the YOLO models can offer very good accuracy, low inference latency and small model size. For applications where better accuracy is required, latest version of the YOLOmodel such as YOLO 10could offer better results but would require more powerful processors.

5 Acknowledgments

Based on these results it is shown that for edge applications targeting single board computers like Raspberry pi, earlier versions of the YOLO models can offer very good accuracy, low inference latency and small model size. For applications where better accuracy is required, latest version of the YOLOmodel such as YOLO 10could offer better results but would require more powerful processors.

References

[1] Youngkwang Kim, Woochan Kim, Jungwoo Yoon, Sangkug Chung, and Daegeun Kim. Deep learning- based multiple droplet contamination detector for vision systems using a you only look once deep learning. 02 2024.

[2] Jun Lee and Su You. Balancing privacy and accuracy: Exploring the impact of data anonymization on deep learning models in computer vision. IEEE Access, PP:1–1, 01 2024.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector, page 21–37. Springer International Publishing, 2016.

[4] D. Mishra, K. Rout, S. Mishra, and S. Salkuti. Various object detection algorithms and their comparison. Indonesian Journal of Electrical Engineering and Computer Science, 29:330, 2022.

[5] Martinus Grady Naftali, Jason Sebastian Sulistyawan, and Kelvin Julian. Comparison of object detection algorithms for street-level objects, 2022.

[6] Joseph Redmon, S. Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2015.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 59 – 65

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Computing the Cleanness of the Photovoltaic (PV) Panels

A. Tsagkaropoulos[a], C. Vasilakis[a], G. Venitourakis[a], P. Tz. Amrou[b],
G. Konstantoulakis[b], P. Golemis[b], D. Reisis[a]*
[a] *Computer Center, Physics Dpt, National & Kapodistrian University of Athens, Greece*
[b] *InAccess Networks, 12 Sorou Str, 15125, Athens, Greece*
* *Corresponding author. Email address: dreisis@phys.uoa.gr*

Abstract

Among the major factors that may significantly affect the power production of a photovoltaic (PV) power plant is the performance of the PV panels. Their performance depends on their maintenance and especially on keeping their surface clean to allow the panels to absorb the maximum power of the solar irradiance. To improve the maintenance process especially in remote locations, the current work focuses on an edge computing system using a camera and a Neural Network (NN) to decide the cleanness of the panels. The proposed system determines whether the panel is either clean or not. The proposed system utilizes a proprietary dataset. The execution of the inference on a Raspberry Pi Zero 2W validates the design, the edge computing performance and the accuracy.

Keywords: Neural Network, Photovoltaic (PV) panels, Edge Computing.

1 Introduction

A significant fraction of the energy produced by the renewable energy sources comes from the photovoltaic (PV) power plants. The PV parks' designers utilize the irradiance and count upon clear skies or reduced cloud coverage to maximize the efficiency of the installation. The efficiency depends also on the cleanness of the PV panels, which are cleaned on a three or six months schedule while in between the panels may pile up either dirt or particles such as dust, water drops, snow and sand. Engineers focused on estimating the performance losses in these cases. The authors of [1] have conducted experiments which showed that these effects may reduce the efficiency of the panel up to 85%. Since such phenomena will appear in more than one panels the deterioration can affect the entire park performance and the solution is to notify the park management group so that a cleaning process will restore the levels of power productivity. A system with the ability of estimating whether the panels are soiled, will provide an effective solution especially for the PV parks in remote locations.

Aiming at an effective solution to the problem of alerting the PV park management when the PV panel(s) is(are) no longer clean, this work proposes an edge computing system using machine learning on the PV panel images and deciding the panels cleanness. The motivation for this work came from the last phase of the Archon project [2] forecasting the irradiance and the power of a PV park. During the testing

the Archon's research and development team noticed that rain drops and dirt could affect the novel park's power forecasting system. Consequently, they noted that the forecasting system assumes clean PV panels and that plant's manager has to be aware and take care of this issue. The literature study showed the significance of the problem [1] as well as the complete absence of approaches and/or techniques addressing the problem of categorizing the panel's status. There is only one paper [3] referring to binary classification of solar panels to dusty or clean, employing a CNN (Convolutional Neural Network) named SolNet.

Given the lack of approaches and datasets for solving the problem and our target for a solution effective with respect to the accuracy and efficient with respect to computing resources first, our work develops a relatively small proprietary dataset. Second, it considers a variety of neural networks (NNs) in an effort to identify the extremely advantageous of these models while these perform on the limited size dataset. Considering that a significantly larger dataset will be exploited in the following research stages, this paper selects nineteen (19) neural networks to work on the current dataset, it compares their performance and presents the results of this early research stage. The paper is organized with the following section describing the technique. Section 3 presents the accuracy results achieved by the performance of the nineteen (19) models and Section 4 discusses the following steps of this research.

2 The NN Based Technique

The current section gives the details of the architecture starting with the dataset, then with the details of the training process, the data augmentation, the optimizers and the schedulers.

2.1. The Dataset

In the first stage of the current work, we chose as the first source for the dataset the deep-solar-eye dataset [4]. The considerably small variation of conditions though led to overfitting and it has triggered the need of another source of training data. Compared to the latter the second dataset is very small, and it constructed by gathering images from arbitrary Google sites. These images are explored to evaluate the model in actual clean or soiled PV panel cases. The train/test split was 80-20. For the continuation of this work the research team will install a network of cameras in available PV plants and it will develop an appropriate dataset expected to include a wide variety of images and soiling cases.

2.2. Training

For this research, we consider two training schemes. A first approach follows the standard training, validation and test split with a 72:8:20 % split for each subset correspondingly. A second, more sophisticated scheme utilizes the k-fold cross validation [5] method. The latter first divides the training dataset to k subsets called folds. The training scheme starts by training the model k times per epoch, each time excluding one of the k folds and using it as a validation subset. Each epoch corresponds to $k-1$ training iterations over the training subset and one iteration over the training subset, mainly used for debugging purposes. Upon the completion of each epoch we validate the model in the validation subset. The application of the latter training scheme to the models improves the prevention of overfitting and presents an overall faster convergence [6]. Both schemes also utilize transfer learning, since all the models used in this study are well-known models and they are pretrained on the ImageNet dataset [7]. All layers except for the output layer of the pretrained models are frozen and therefore not

trained. The design of the output layer realizes the functionality of a linear layer of two output features fed to a SoftMax activation function, which are rendered as the two probabilities for soiled and clean panels. The transfer learning technique allows for rapid training of the models, while keeping the high accuracy of the pretrained model on the ImageNet dataset. The loss function befitting to the problem of the binary classification of images in two classes is the Cross-Entropy Loss function, as the models output two logits, one for each class of soiled and clean panels.

The limited size of the dataset and the small variations among the input data has led us to experiment with data augmentation. Augmenting the input data is common practice for such occasions, while incorporating those methods will increase the gains with respect to training efficiency and accuracy and will reduce the probability of overfitting occurrence. These methods include resizing, cropping, random flipping and rotating, scaling and normalizing the RGB values with the same mean and standard deviation of the ImageNet training dataset.

2.3. Loss Function & Class Probabilities

The quantification of classifier layer's results is accomplished by using the Cross Entropy Loss function. A preprocessing step involves the application of the softmax function to the model's results to enhance the distinctiveness of the probabilities. The Cross Entropy Loss function accepts these probabilities and it computes a value of ***loss*** with respect to the reference level. This loss value gives the difference from the computation result to the previously known result: either clean or soiled PV panel. The main reason for this approach is that in the future work it may be explored to analyze the class "soiled" into multiple classes. At this early stage the entire architecture is designed for binary classification, mainly due to the limited dataset and consequently, the model outputs two logits.

2.4. Data Augmentation

We incorporate data augmentation techniques to lessen the effect of the dataset's limiting factors mentioned in Section 2.1. The small size of the dataset decreases the training efficiency and results in overfitting of the model on the training data. Thus, the dataset was enriched by methods concerning random image rotations, flipping and cropping. At the same time, we apply the novel TrivialAugment **[8]** transformation to the images, which also affects the saturation, the contrast and the color profile. We note here that, in many cases, the pretraining of the models uses normalization and scaling of the RGB values. For these models we have to use the normalization and the scaling of the images with the values that have been used during the pretraining. The frame sizes of model's inputs are aggregated in the Results Section in **Table 1.**

2.5. Optimizers & Schedulers

We train models by using one of the following three optimizers. The first optimizer is the SGD (Stochastic Gradient Descent) [9], which updates the model's weights by taking into account the gradient of the loss function with respect to the weights. This is the most trivial optimizer that we have used in our experiments. The hyperparameters corresponding to the SGD were: 0.001/0.01 for the learning rate and, optionally, 0.9 for the momentum. The second optimizer that we applied is the RMSProp, which improves the weight change stability. Its hyperparameters were: 0.01 for the learning rate, 0.99 for the alpha and 1e-08 for the epsilon. The third optimizer is the Adam algorithm [10] that

facilitates the optimization of the models' weights. The Adam's hyperparameters were: 0.001 for learning rate, (0.9, 0.999) for betas and 1e-08 for epsilon.

To further enhance the efficiency and consistency of the training procedure we added to the process the use of one (out of two) scheduler. A scheduler reduces the learning rate whenever the advancement in training stalls. We specify the loss as the metric that evaluates the learning progress. The first scheduler is the StepLR, which decreases the learning rate of the optimizer by a specified amount every constant number of epochs. The ReduceLROnPlateau is the second scheduler: its application will reduce the learning rate by a specified amount if the loss value reduction stalls for a "patience" amount of epochs. To explain the need for a scheduler when the training fails to result in loss reduction: the idea is to let the process follow incrementally small advancements and hence, find local minima in multidimensional spaces.

3 The Results Using Different Models

This section presents the accuracy results of the technique with the use of nineteen different models in the same tests. In the following table the first column shows the nineteen models that we used. These are sorted with respect to the accuracy results that they achieved during the tests, which are listed in the second column. The third column presents the top 1 accuracy that each model achieved with the ImageNet dataset. For sake of results completeness the fourth column displays the frame size that we used with each model. The fifth column gives the number of parameters of each model.

The first in the list, namely the ShuffleNet_V2_X1_5, shows to be the most efficient model with respect to both the accuracy results as well as the number of parameters needed. This high ranking though may be due to the small dataset used. Moreover, the models ranked from 2 to 4 present an efficiency relatively close to the ShuffleNet_V2_X1_5. Furthermore, taking into account that the execution time of the entire panel-cleanness application is practically indifferent to the application (it can be quite slow), the number of parameters plays a significant role to the cost of the edge device. Hence, the low cost implementation of these four models (the upper four in the rank) is a fact that makes them an attractive solution for edge applications.

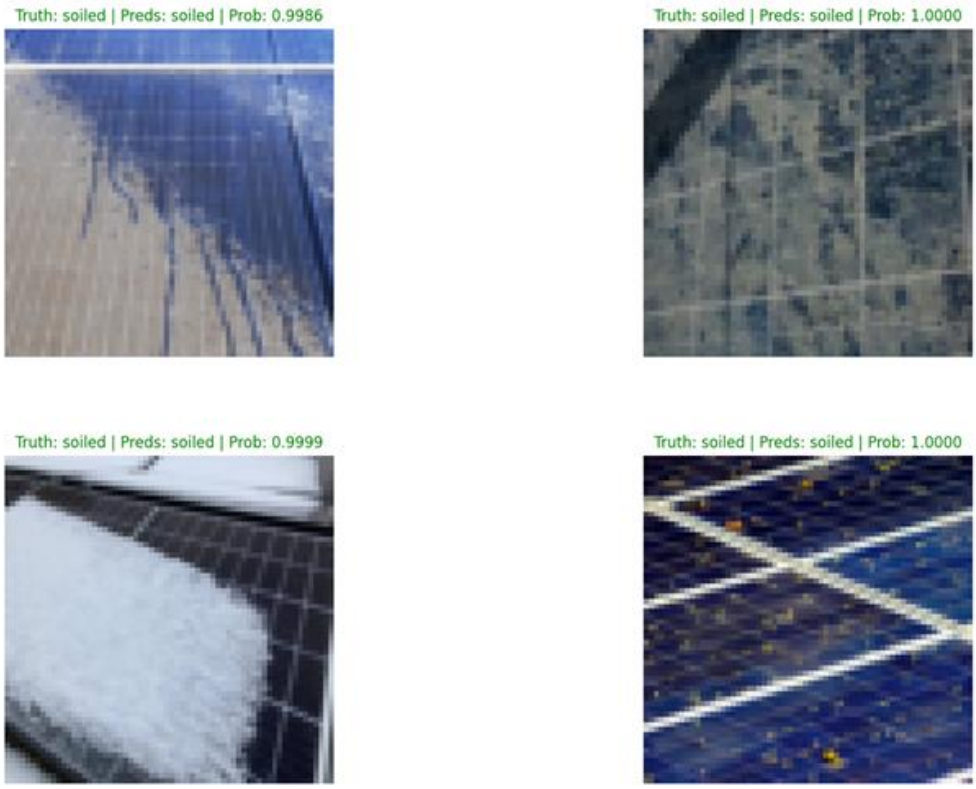
Table 1: Different NN models Ranked with respect to their Accuracy Results

Model Name	Test Accuracy (%)	Acc@1 – ImageNet (%)	Frame Size	Params
ShuffleNet_V2_X1_5 [11]	93.97	72.996	224	3.5M
ResNet34 [12]	91.00	73.314	224	21.8M
ShuffleNet_V2_X2_0 [11]	90.36	76.23	224	7.4M
EfficientNet_B2 [13]	89.30	80.608	288	9.1M
EfficientNet_V2_M [14]	88.89	85.112	480	54.1M
EfficientNet_V2_L [14]	88.50	85.808	480	118.5M
MobileNet_V3_Large [15]	88.45	75.274	224	5.5M
EfficientNet_B7 [13]	88.30	84.122	600	66.3M

MobileNet_V3_Small [15]	87.65	67.668	224	2.5M
EfficientNet_B7 [13]	88.30	84.122	600	66.3M
MobileNet_V3_Small [15]	87.65	67.668	224	2.5M
EfficientNet_B0 [13]	87.34	77.692	224	5.3M
EfficientNet_B1 [13]	86.90	79.838	240	7.8M
ResNet50 [12]	86.02	76.130	232	25.6M
ResNet18 [12]	85.79	69.758	224	11.7M
EfficientNet_B3 [13]	85.77	82.008	300	12.2M
MobileNet_V2 [16]	85.51	71.878	224	3.5M
EfficientNet_B6 [13]	83.73	84.008	528	43.0M
EfficientNet_V2_S [14]	83.78	84.228	384	21.5M
ShuffleNet_V2_X1_0 [11]	81.80	69.362	224	2.3M
ShuffleNet_V2_X0_5 [11]	79.57	60.552	224	1.4M

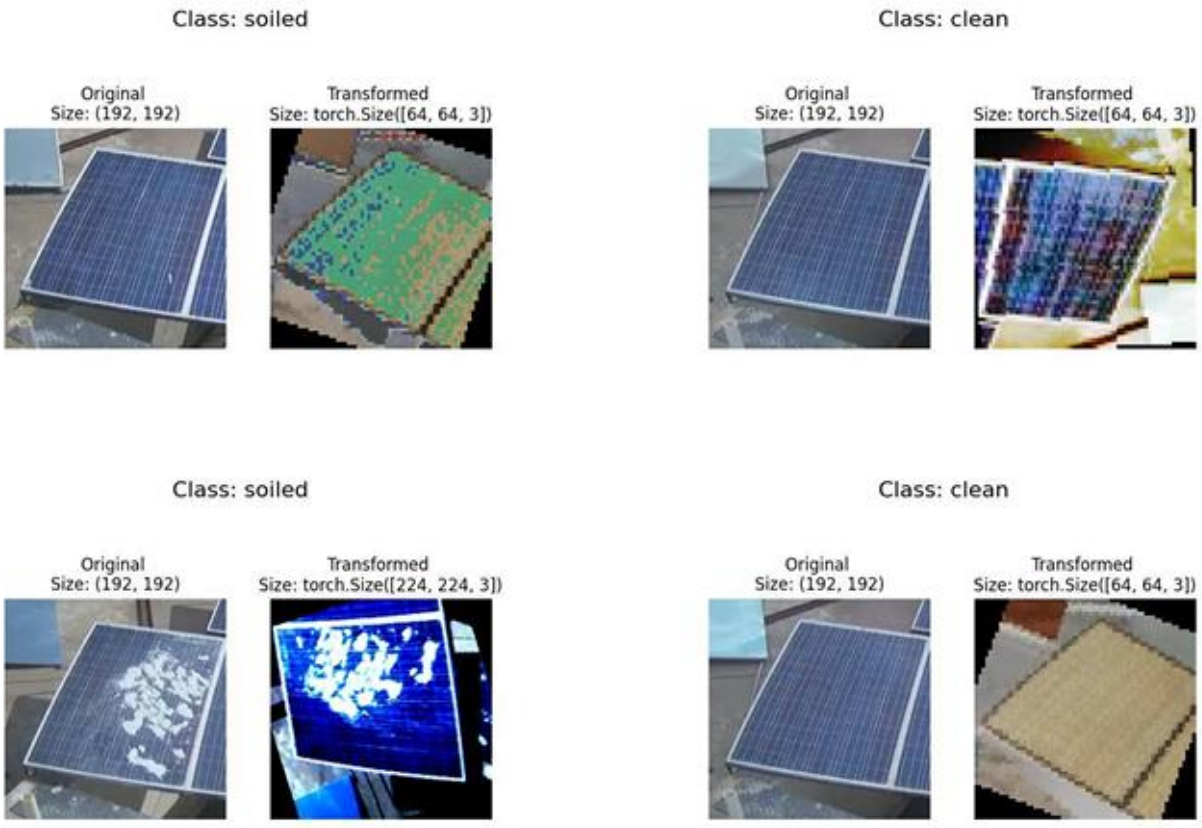
We include two images presenting the results from the inference. Figure 1 depicts these examples.

Figure1: Results from Inference



Furthermore, we give an example of the data augmentation, which is applied only during the training process. The images show the two different classes in two different cases. Figure 2 depicts these examples.

Figure2: Results from Data Augmentation



4 Future Work

This paper has presented an edge computing system that decides whether a (or as many as the camera locates) PV panel is clean or soiled. The research of the system is ongoing and currently we focus mainly on the lack of a dataset and the creation of a new dataset with a very large number of images. For this initial stage we have employed nineteen different models that have been tested to sort them according to their efficiency with respect to the accuracy results and the number of parameters.

Our future work will focus on developing a dataset by installing cameras at PV plants, which will be located at various locations. The variety of locations will cover a multiple of soiling cases. These include dirt, dust, pollen, bird droppings, snow or ice, desert sand, etc. Moreover, we will improve the architecture to let the user choose the level of cleanness of the panels. This can be achieved by enhancing the proposed system so that it will guide the user during the choice of the cleanness threshold.

References

[1] S. Shaharin, K. S. Atul, M. M. Mior, Mokhtar and A. B.-R. Mohammed, "Influence of Dirt Accumulation on Performance of PV Panels," Energy Procedia, vol. 50, pp. 50-56, 2014.

[2] "Archon," In Access Nteworks, 2022. [Online]. Available: http://archonproject.eu/.

[3] M. S. H. Onim, Z. M. M. Sakif, A. Ahnaf, A. Kabir, A. K. Azad, A. M. T. Oo, R. Afreen, S. T. Hridy, M. Hossain, T. Jabid and M. S. Ali, "SolNet: A Convolutional Neural Network for Detecting Dust on Solar Panels," MDPI Energies, vol. 16, 2023.

[4] S. Mehta, A. P. Azad, S. A. Chemmengath, V. Raykar and S. Kalyanaraman, "DeepSolarEye: Power Loss Prediction and Weakly Supervised Soiling Localization via Fully Convolutional Networks for Solar Panels," Arxiv, 2018.

[5] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," Journal of the Royal Statistical Society. Series B (Methodological), vol. 36, no. 2, pp. 111-147, 1974.

[6] D. S. Soper, " Greed Is Good: Rapid Hyperparameter Optimization and Model Selection Using Greedy k-Fold Cross Validation," MDPI Electronics, 2021.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[8] S. G. Müller and F. Hutter, "TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation," 2021.

[9] Sutskever, Ilya, J. Martens, G. Dahl and G. Hinton, "On the importance of initialization and momentum in deep learning," in Proceedings of the 30th International Conference on Machine Learning, 2013.

[10] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2017.

[11] N. Ma, X. Zhang, H.-T. Zheng and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," 2018.

[12] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2015.

[13] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2020.

[14] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," 2021.

[15] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le and H. Adam, "Searching for MobileNetV3," 2019.

[16] S. Mark, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2019.

EMERGING TECH CONFERENCE 2024
Edge Intelligence

Papers

Session 1.3 | Field Devices

Session Chairs: Thomas NOULIS & Fotis PLESSAS

Session Presentation: George KERAMIDAS

Using AMR Sensors for Surface Residual Stress Determination in Steel Constructs

Spyros Angelopoulos, Gregory Doumenis, Aphrodite Ktena and Evangelos Hristoforou

A novel approach on continual operation of compromised ECU functions: REWIRE Perspective

Athanasios Athanasiadis, Christoforos Koutsianoudis, Konstantinos Lamaris and Tilemachos Matiakis

A Low Power FPGA Implementation of LDPC Encoder for Space Application

Christos Sidiras and Vasilis Pavlidis

Evaluation and DSP Benchmarking of the European Radiation-Hardened NG-ULTRA FPGA

Anastasios Xynos, George Lentaris and Dimitrios Soudris

Autonomous Multi Source Energy Harvesting Multi-Sensor

Ioannis Masklavanos, Theodoros Georgiadis, Vasiliki Naskari, Fotios Vartziotis and Gregory Doumenis

HETiA
EXPANDING THE DIGITAL FRONTIERS

UNIVERSITY OF THESSALY FOUNDED 1984

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Page 66 – 73

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Using AMR Sensors for Surface Residual Stress Determination in Steel Constructs

Spyros Angelopoulos[1], Gregory Doumenis[2], Aphrodite Ktena[3] and Evangelos Hristoforou[1]*
[1] *Laboratory of Electronic Sensors, School of Electrical & Computer Engineering, NTUA, Greece*
[2]*Autonomous Systems Laboratory, School of Informatics and Telecommunications, Univ. of Ioannina, Arta, Greece*
[3]*National Kapodistrian University of Athens*
**Corresponding author's Email: hristoforou@ece.ntua.gr*

Abstract

In this paper, a non-destructive methodology designed for quantification of residual stresses in magnetic steels is presented. A portable anisotropic magnetoresistance (AMR) sensor has been developed for specimen inspection capable of determining the distribution of magnetic permeability tensor in a contactless manner. Correlating the AMR response with residual stresses, the localized stress tensor has been determined. This method holds significant promise for enhancing the reliability and safety of critical components in various ship-environment applications and beyond.

Keywords: Non-destructive testing, Anisotropic magnetoresistance, Residual stress, Magnetic permeability, Fatigue assessment, Finite element analysis

1 Introduction

Residual stresses and/or strain gradient monitoring is an important parameter to determine and monitor the structural and conditional health of steels and steel structures. Mechanical and thermal fatigue during manufacturing process or during steel operation introduce dislocations, dislogation forests and finally nano-craks and crack in the steel. Beyond the yield point, plastic deformation includes already nano- and micro- cracks that are formed in order to achieve stress relief. These cracks propagate in certain directions leading to possible steel failure. Therefore, it is vital to monitor the evolution of residual stress or strain gradients over time -even in the elastic region-, to observe the tendency to reach the yield point and therefore generate a crack. Additionally, a sharp spatial gradient in residual strains or stresses before yield can result in a crack initiation, while a uniform distribution of higher level of residual stresses may not lead to a fracture [1].

The current industrial methods of non-destructive stress monitoring are the strain gauge [2], as well as the hole drill method [3], both used for local surface monitoring, having several operating limitations. Neither method can provide distribution stress monitoring. Recently, non-linear acoustics have been used for residual stress monitoring, allowing for distribution stress measurements [4].

As steel materials and structures exhibit magnetic properties, magnetic methods have also been proposed and investigated as an alternative or complementary to the above sensing methods [5]. The

magnetization process and resulting macroscopic magnetic properties depend on the crystalline structure as well as on the microstructure at the grain level [6]. Residual strains affect the anisotropy profile and residual stresses act as effective magnetostatic fields on the spatial variation of the magnetization. Hence, microstructural changes related to residual stresses may be detected through the monitoring of macroscopic magnetic parameters obtained from hysteresis loop measurements, such as the differential magnetic permeability [7], from magnetic Barkhausen noise (MBN) measurements [8] and from magnetoacoustic waves [9, 10].

The magnetic stress calibration (**MASC**) curve principle has been proposed to quantitatively link residual stresses, determined through X-ray diffraction measurements in the Bragg Brentano arrangement, to the differential permeability or the MBN rms voltage [11]. It has been found that a unique MASC curve can be obtained for each given steel grade: it is of sigmoidal shape with stress along the horizontal axis and the magnetic property on the vertical axis. The MASC curve can be used to convert a measured value of differential permeability to residual stress. Obtaining MASC curves is a laborious process, which is perfomed only once for a given steel grade. However, this process led to the following observation: normalizing the residual stress σ (X-axis) against the yield point of the given steel and of the magnetic permeability μ (Y-axis) against the maximum value of the differential permeability, all MASC curves collapse into one single curve, called the universal MASC curve [11].

In this paper, we introduce the anisotropic magnetoresistance (AMR) measurement in order to determine residual stresses on the surface of magnetic steels. Such sensor has some advantages, namely low power consumption, contactless measurement up to 10 mm far from the surface of the under test steel, and ability of moving along the under test steel to determine gradient effects.

2 Flux leakage without surface crakcs in the presence of residual stresses

The novelty of this study lies in measuring the flux leakage in the absence of cracks, but in the presence of residual stresses, on the surface of magnetic steel specimens. The idealization presented assumes only surface residual stresses, i.e. for depths in the order of 10 – 100 μm, without any loss of the generality. Any residual stresses are expected to result in an increase or decrease of the differential magnetic permeability of the material as follows:

- If the material is positively magnetostrictive[§§] at a given direction, then any tensile residual stresses will result in an increase of the differential magnetic permeability and a decrease of the coercivity field Hc[***], while any compressive stresses at a given direction, will result in a decrease of the differential magnetic permeability and an increase of the coercive field Hc.
- If the material is negatively magnetostrictive at a given direction, then any tensile residual stresses will result in a decrease of the differential magnetic permeability and an increase of the coercivity field Hc, while compressive stresses will result in an increase of the differential magnetic permeability and a decrease of the coercive field Hc.

Residual stresses can be assumed to operate like an effective magnetic field which is added or subtracted on the biasing field. If in one axis e.g. the X-axis, the magnetic material is positively magnetostrictive, then it will be negatively magnetostrictive in its perpendicular counterpart e.g. the

[§§] Magnetostrictive material is defined the magnetic material changing its dimensions under the influence of magnetic field (in the order of part per million – ppm). Positive magnetostrictive material is the magnetic material with increasing dimensions along the applied field, where the opposite occurs for a negative magnetostrictive material.

[***] Coercive field is the biasing field in which the macroscopic magnetization of the magnetic material is null.

Y axis. Notwithstanding this, the same principles apply for positive and negative magnetostriction materials concerning bulk residual stresses, i.e. including residual stresses in the Z-axis. All these are valid under the assumption of absence of cracks in the material.

By moving an AMR sensor along the X-axis of a positively magnetostrictive material will result in a decrease of the magnetic flux in the X & Z axes and an increase in the Y-axis, in case of tensile residual stresses. The opposite is expected for compressive stresses. Moving the AMR sensor along the X-axis of a the same (positive magnetostrictive material) under compressive residual stresses will result in an increase of the magnetic flux in the X & Z axes and a decrease in the Y-axis. The opposite is expected for the case of negative magnetostrictive material in each direction. Tensile stress will result in an increase of X & Z field component, as well as in a reduction of Y-axis field component. The opposite holds for compressive stresses. This behavior is illustrated in Fig 4.

In contrast to surface defects which lead to magnetic lines distortion of the order of 100-200μT, the magnetic field distortion due to residual stresses is much smaller, notably in the range of nT to a few μT. The use of AMR sensors because of their sensitivity to the order of nT and their capability to display 3-dimensional field measurements makes them the most suitable option for such measurements.

The problem is the knowledge of the positive or negative magnetostrictive character of an under test steel specimen. However, this is not of paramount importance, since the spatial gradient of the residual stresses is the decisive factor for fatigue failure assessments and for making predictions on the remaining lifetime of a steel. However, the determination of the positive or negative character of magnetostriction can be precisely determined by observing the evolution of the response of an array of AMR sensors over time, using AI-ML algorithms and protocols. This is work in progress.

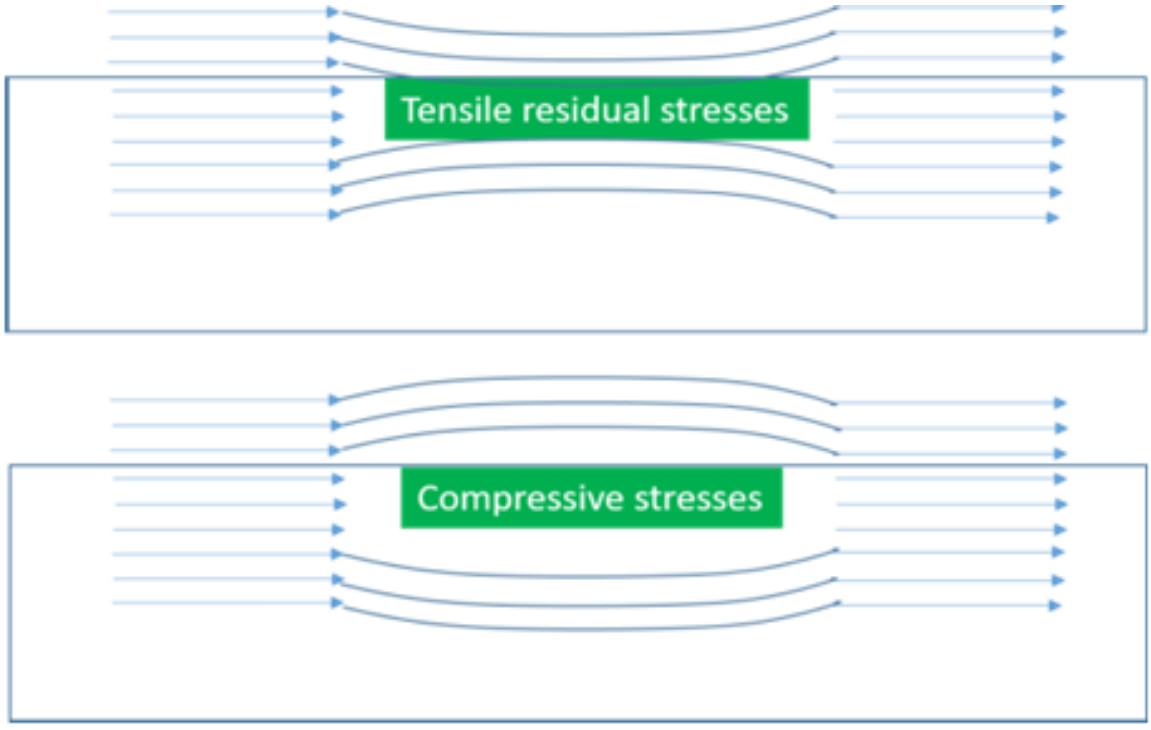


Figure 4: Change of the orientation of magnetic lines in a positive magnetostrictive direction under the influence of tensile and compressive stresses.

An important issue is the correlation of the AMR output (in V or T) with residual stresses (MPa), offering the prediction of location and time a steel failure may happen. Fundamental to this is the spatial and periodic monitoring of the differential magnetic permeability as presented in the following sections.

3 AMR response with and without biasing field

By an electromagnetic yoke it is possible to compare the magnetic flux leakage, as measured by the AMR sensor in the X, Y and Z axis, with the presence or absence of the additional biasing field. Under zero biasing field, the specimen is magnetized by only the Earth's magnetic field. Under the influence of additional biasing field, the flux leakage screens additional information about the status of residual stresses in the steel specimen. Until it is driven to saturation, the magnetic flux leakage will increase or decrease as a function of the residual stresses with respect to the positive or negative character of the magnetostriction in each axis. After exceeding the saturation level, any additional biasing field becomes insignificant within the context of this study. This is because the additional magnetic lines are not trapped by the material. In that case the AMR sensor's response is not dependent on residual stresses, but only on the additional biasing field applied.

Figure 10: Electric steel samples after mild (top) and heavy (bottom) induction heating.

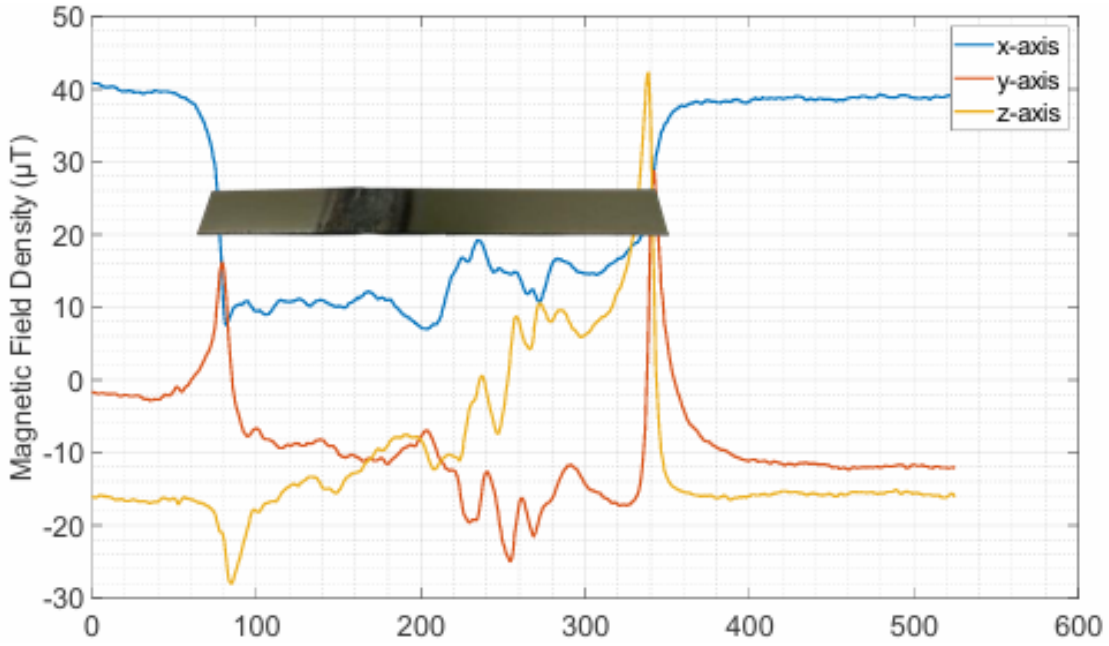


Figure 5: 3D AMR response without the biasing DC magnetic yoke (steel element in physical dimensions with respect to distance measurements).

To qualitatively highlight the differences between the two methods, a specimen in which residual stresses had been intentionally introduced by localized RF heating and consequent quenching was used. The AMR sensor's response, along the path length assumed for the test specimen, without using an additional biasing field, is illustrated in Fig 5. Respectively, the AMR sensor's output along the same path on the same specimen, using a biasing field introduced by a DC magnetic yoke (Fig.1) is illustrated in Fig 6. It is apparent that the sensitivity of the AMR sensor is better in the case of not using a biasing field, as it allows for the identification of several small-scale stress fields along the test path, which in the case of using the biasing field go unnoticed. This happens because magnetic domain walls overpass easily the small residual stress fields in case of elevated biasing field, while Earth's biasing field allows only for small Barkhausen jumps[†††] for small stress fields. Thus, the characterization of the steel in the absence of biasing field, for the case of AMR sensors, is preferable, allowing for a more detailed mapping of stress field.

It is worth noting that the large terminal field changes, at about 100 mm and 350 mm, refer to the boundaries of the steel specimen, allowing for knowing where the steel specimen is, while all the rest of signals inside the envelope of the outer large signals illustrate smaller and larger stress fields.

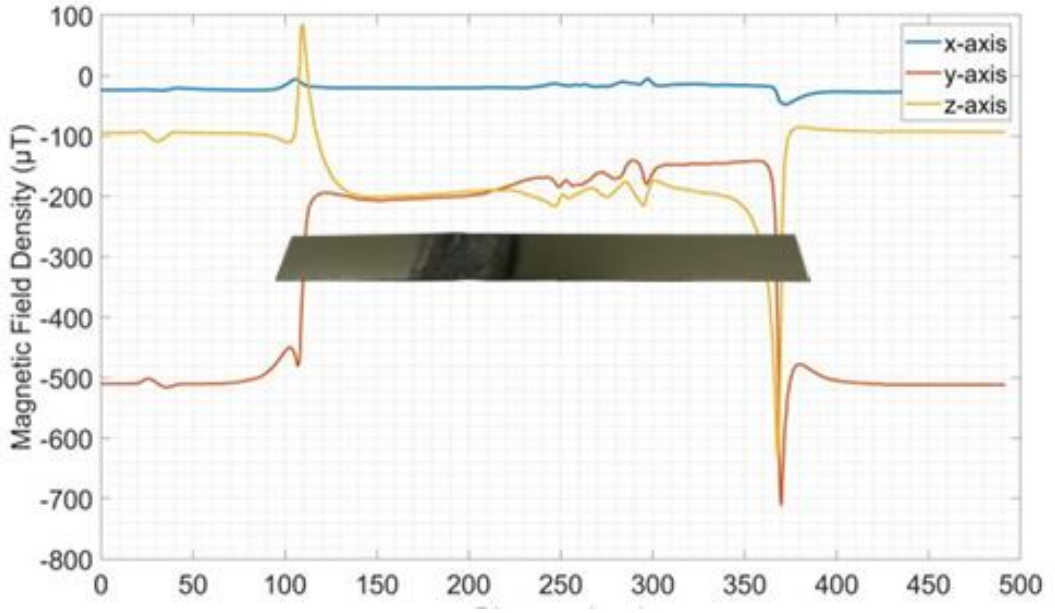


Figure 6: 3D AMR response with the biasing DC magnetic yoke. (steel element in physical dimensions with respect to distance measurements)

[†††] Barkhausen jump is the sudden overpassing of a stress field, after energy accumulation, resulting in hysteresis effects in steels.

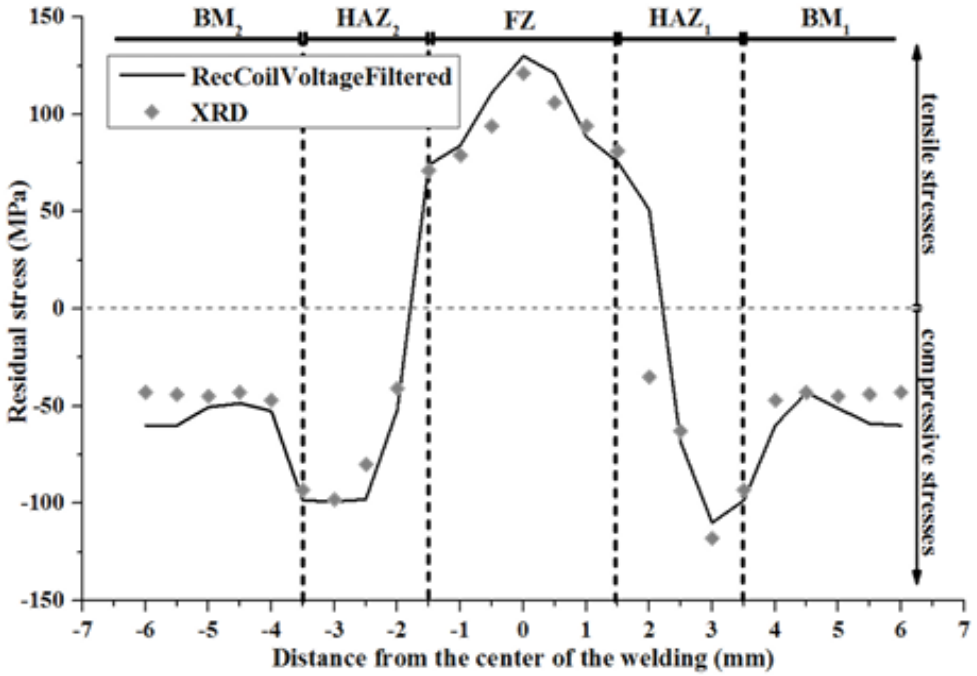


Figure 7: Correlation of residual stresses (dot points) with local differential magnetic permeability (continuous line, considered in arbitrary units) [15].

The AMR sensor's response indicates an almost linear dependence of residual stresses with the magnetic permeability.

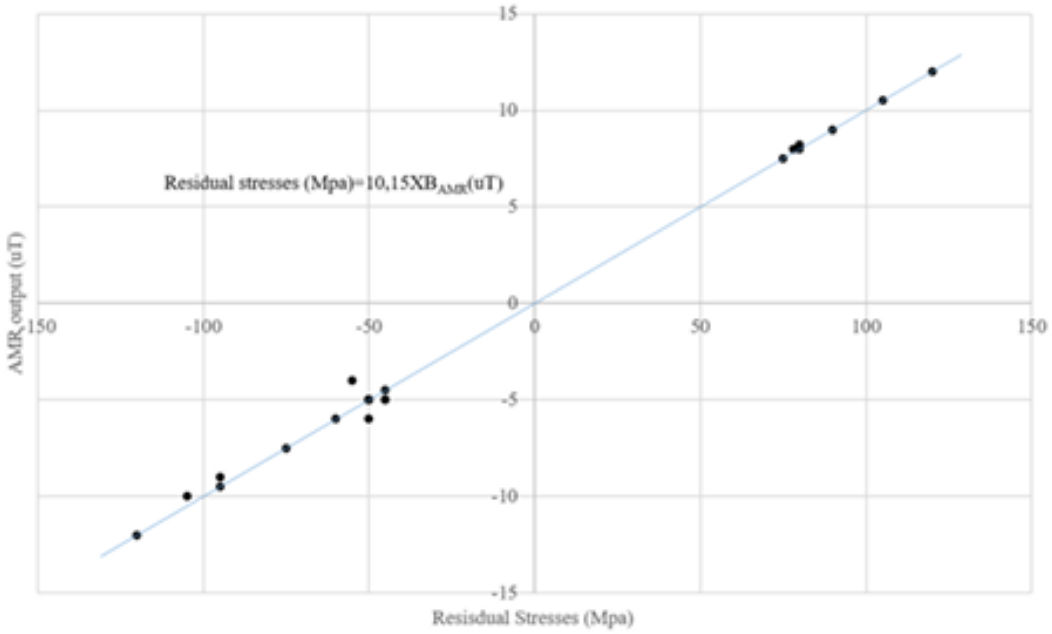


Figure 8: Dependence of residual stresses (MPa) on the AMR response (uT).

4 Conclusions

A new method for the determination of the magnetic stress calibration (MASC) curve is proposed in this paper, related to an advanced method of inducing residual stresses in magnetic steels, following localized induction heating and consequent quenching, allowing for MASC curve, suffering from much less uncertainties and having larger span with respect to stresses. The use of an AMR sensor appears to be promising for a distant detection of the localized residual stresses in different types of magnetic steels including 42CrMo4 steels

References

[1] R.W. Cahn, P. Haasen, Physical Metallurgy, 4th Edition, 1996, eBook ISBN: 9780080538945

[2] Atzori, B., and Meneghetti, G., Fatigue strength of fillet welded structural steels: finite elements, strain gauges and reality, International Journal of Fatigue, 23.8, 713-721, 2001

[3] Beaney, E. M., Accurate measurement of residual stress on any steel using the centre hole method, Strain, 12(3), 99-106, 1976

[4] Nagy, P. B., Fatigue damage assessment by nonlinear ultrasonic materials characterization, Ultrasonics, 36(1-5), 375-381, 1998

[5] Kikuchi, H., Henmi, Y., Liu, T., Ara, K., Kamada, Y., Kobayashi, S., & Takahashi, S., The relation between AC permeability and dislocation density and grain size in pure iron, International Journal of Applied Electromagnetics and Mechanics, 25(1-4), 341-346, 2007.

[6] Hauser, H., Energetic model of ferromagnetic hysteresis: Isotropic magnetization, Journal of Applied Physics, 96(5), 2753-2767, 2004.

[7] Jiles D.C., Review of Magnetic Methods for Non-Destructive Evaluation, NDT International, 23 (2) , 2112−2115, 1987

[8] Stupakov, O., Perevertov, O., Tomáš, I., & Skrbek, B., Evaluation of surface decarburization depth by magnetic Barkhausen noise technique, Journal of magnetism and magnetic materials, 323(12), 1692-1697, 2011.

[9] Augustyniak, B., Chmielewski, M., Piotrowski, L., & Kowalewski, Z. (2008). Comparison of properties of magnetoacoustic emission and mechanical Barkhausen effects for P91 steel after plastic flow and creep. IEEE Transactions on magnetics, 44(11), 3273-3276.

[10] Hristoforou, E., Magnetostrictive delay lines: Engineering theory and sensing applications, Measurement Science and Technology 14(2), pp. R15-R47,2003

[11] Hristoforou, E., Vourna, P., Ktena, A., Svec, P., On the Universality of the Dependence of Magnetic Parameters on Residual Stresses in Steels, IEEE Transactions on Magnetics 52(5),7362189, 2016

[12] Ege, Y., Coramik, M., A new measurement system using magnetic flux leakage method in pipeline inspection, Measurement: Journal of the International Measurement Confederation, 123, pp. 163–174, 2018

[13] https://www.ansys.com/news-center/press-releases/09-18-18-ansys-19-2-delivers-faster-problem-solving capabilities-across-entire-portfolio.

[14] Li, C.-C., Leslie, W.C., Effects of dynamic strain aging on the subsequent mechanical properties of carbon steels, Metallurgical Transactions A, 9(12), pp. 1765–1775, 1978

[15] Vourna, P., Ktena, A., Tsakiridis, P.E., Hristoforou, E., A novel approach of accurately evaluating

residual stress and microstructure of welded electrical steels, NDT and E International, 71, pp. 33-42, 2015

[16] Vourna, P., Hervoches, C., Vrana, M., Ktena, A., Hristoforou, E., Correlation of magnetic properties and residual stress distribution monitored by X-ray and neutron diffraction in welded AISI 1008 steel sheets, IEEE Transactions on Magnetics, 51(1),7029219, 2015

[17] Mangiorou, E., Damatopoulou, T.V., Angelopoulos, S., Ktena, A., Hristoforou, E., Revisiting the universality law in magnetically detected residual stresses in steels, AIP Advances, 14(2),025126, 2024

[18] Rosenthal, D., The Theory of Moving Sources of Heat and Its Application to Metal Treatments, Journal of Fluids Engineering, Transactions of the ASME, 68(8), pp. 849–865, 1946.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 74 – 80

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

A novel approach on continual operation of compromised ECU functions: REWIRE Perspective

Athanasios Athanasiadis[1], Christoforos Koutsianoudis[1], Konstantinos Lamaris[1] and Tilemachos Matiakis[1]
[1] *KENOTOM P.C., Kalamaria-Thessaloniki, Greece*
a.athanasiadis@kenotom.com, c.koutsianoudis@kenotom.com, k.lamaris@kenotom.com, t.matiakis@kenotom.com

Abstract

As the complexity of automotive Electronic Control Units (ECU) constantly increases with functionalities such as Autonomous Driving (AD), Advanced Driving Assistance Systems (ADAS), and Augmented Reality (AR), new trends emerge in automotive ECU network topologies and software engineering. Centralized ECU architectures are gaining traction, consolidating more vehicle functions into fewer ECUs to reduce network load and improve efficiency, in comparison to traditional de-centralized ECU architectures. However, this shift introduces new challenges, particularly in terms of security and the continual operation of critical functions. A major challenge is that a potential security breach and the subsequent ECU compromise would impact numerous vehicle functionalities. Therefore, ensuring the reliability and security of these systems requires innovative solutions. This paper proposes an approach that addresses these challenges by presenting the concept of "vehicle functions migration" to alternative or auxiliary ECUs, enhancing the overall system's robustness and security. Finally, the implementation and demonstration of this approach within the EU funded project "REWIRE" is presented.

1 Introduction

Traditional automotive electrical/electronic (E/E) architectures primarily utilize a "decentralized" approach, where each specific vehicular function is managed by an individual ECU, connected through a common bus network (e.g. CAN, CAN-FD, FlexRay, LIN). This method has been prevalent for many years and is widely implemented in most current production vehicles. Decentralized vehicular E/E architectures offer several advantages, including:

- Separation of Concerns: They allow specific functions to be managed by individual ECUs, simplifying verification processes.
- Ease of Replacement: Lightweight ECUs can be easily replaced when damaged.
- Simplicity in Integration: With limited functionality per ECU, integrating them into a network is straightforward.

However, this decentralized approach presents several drawbacks and challenges, particularly in terms of scalability and communication [1].

In response to these challenges, “centralized” ECU network architectures are gaining significant traction, with many OEMs and Tier 1 suppliers exploring these solutions. Unlike the decentralized approach, centralized “Domain” automotive architectures group more vehicle functions or even entire vehicle domains (Powertrain, Chassis, etc.) into more powerful and less in number ECUs (Figure 1 ”Domain centralized architecture”) , while a fully centralized “Zonal" architecture (Figure 1 ”Vehicle centralized E/E architecture”) would only have a single central computing platform to handle the majority of vehicle’s functions, and multiple simplistic control units that would only serve as “smart actuators” or gateway units.

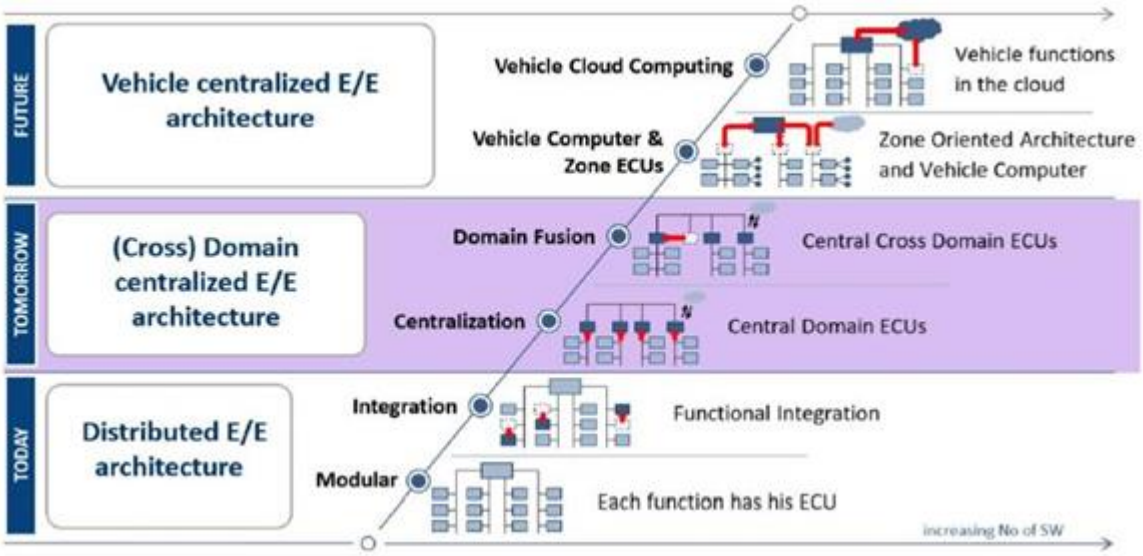


Figure 1: Possible evolution of vehicular E/E architectures. Figure adopted by [5]

Overall, centralized architectures signify an architectural shift more tailored to the current automotive needs, presenting various benefits:

- Enhanced scalability through the centralization of functions, allowing for: easier management of vehicle functions, easier software updates, while reducing complexity and maintenance costs.
- A smaller number of ECUs reduces needed network relations and improves communication efficiency.
- Provides a better suited platform for latest automotive trends, such as ADAS, Over-The Air (OTA) updates and vehicle connectivity.
- Standardization of Zonal/Domain ECUs (e.g. with respect to a standard communication interfaces)

However, centralized ECU architectures introduce new challenges. A potential security breach could compromise a Domain or Zonal ECU, which on centralized architectures manages numerous vehicle functionalities. For example, on the Domain architecture the Functional Domain Controller would group large portions of the functionalities of a given domain (e.g. Powertrain, Chassis, etc.) with the lower-level ECUs within that network acting as ‘smart sensors/actuators”. As a result, fail-safe mechanisms and rollback measures become increasingly necessary in order to maintain at least

portions of the functionality in the case of a compromise.

This paper presents the novel concept of maintaining the operation of critical functions through "vehicle function migration" to alternative, auxiliary ECUs. This concept will be explored within the automotive use case, undertaken by KENOTOM, within the REWIRE project. REWIRE aims to develop a novel security and trust assessment framework, for Next-Generation connected "Systems-of Systems" (SoS) covering the strict security, safety, and resilience requirements during the entire lifecycle of a Cyber Physical System (CPS). The REWIRE framework will enable real-time protection through the implementation of a continuous security improvement process covering [2]:

- the design phase based on overarching system requirements
- the runtime phase covering the operation, update re-configuration, and even decommissioning of a compromised device.

2 Compromised ECU migration

In order to maintain the compromised ECU functionality, migration to a neighboring or back-up ECU will guarantee the reliability and the continuity of safety-critical services. Within the REWIRE project, the migration will be implemented and demonstrated in the following discrete steps:

- Device attestation, attack and subsequent compromise detection. Within the REWIRE ecosystem, physical, software, side-channel, and denial-of-service attackers are listed as examined cases that can compromise the security guarantees of the system.
- Extraction of the latest safe state of the ECU and the relevant functions to be migrated.
- Deployment of functions on other predefined ECUs that can handle the workload.

Expanding more on the REWIRE approach on the migration of functions, on the following exemplary setup Figure 2, Zonal/Domain Controllers are serving the role of collecting and processing data from a given sub-network of low-level ECUs. These ECUs can act as lower level, simplified "smart actuators" for example VCM (Vehicle Control Module) is responsible for operating the vehicle's engine, or collecting brake pedal position, with the ADAS ECU used to perform lower-level vehicle motion control (e.g. steering or acceleration). Following that setup, higher level vehicle decisions, like motion or driving strategies (e.g. ECO or SPORT modes, V2X communication, etc.), could be performed on a higher level, "master ECU", that would serve as a central High-Performance Computer (HPC), or even as a server to off-load the computational tasks on the cloud. Zonal/Domain Controller 1 (ZCU1) is the main on-board REWIRE edge device that will include the REWIRE security artefacts, along with the exemplary automotive demo application. ZCU2 is the second REWIRE-specific board that demonstrates migration of functions, an identical with ZCU1, GENESYS2 board. Migration could be facilitated either via a direct connection of ZCU1 and ZCU2 or through ZCU2 connection to higher level ECU systems (e.g. Central HPC), through the **REWIRE Trusted Execution Environment (TEE)**[‡‡‡] of each ZCU, by establishing a security key between the TEEs. Security of the migration will be independent of how secure the physical network is, since REWIRE TEE will undertake this task.

[‡‡‡] A **Trusted Execution Environment (TEE)** is a secure area of a main processor. It helps the code and data loaded inside it be protected with respect to confidentiality and integrity. Data confidentiality prevents unauthorized entities from outside the TEE from reading data, while code integrity prevents code in the TEE from being replaced or modified by unauthorized entities [4].

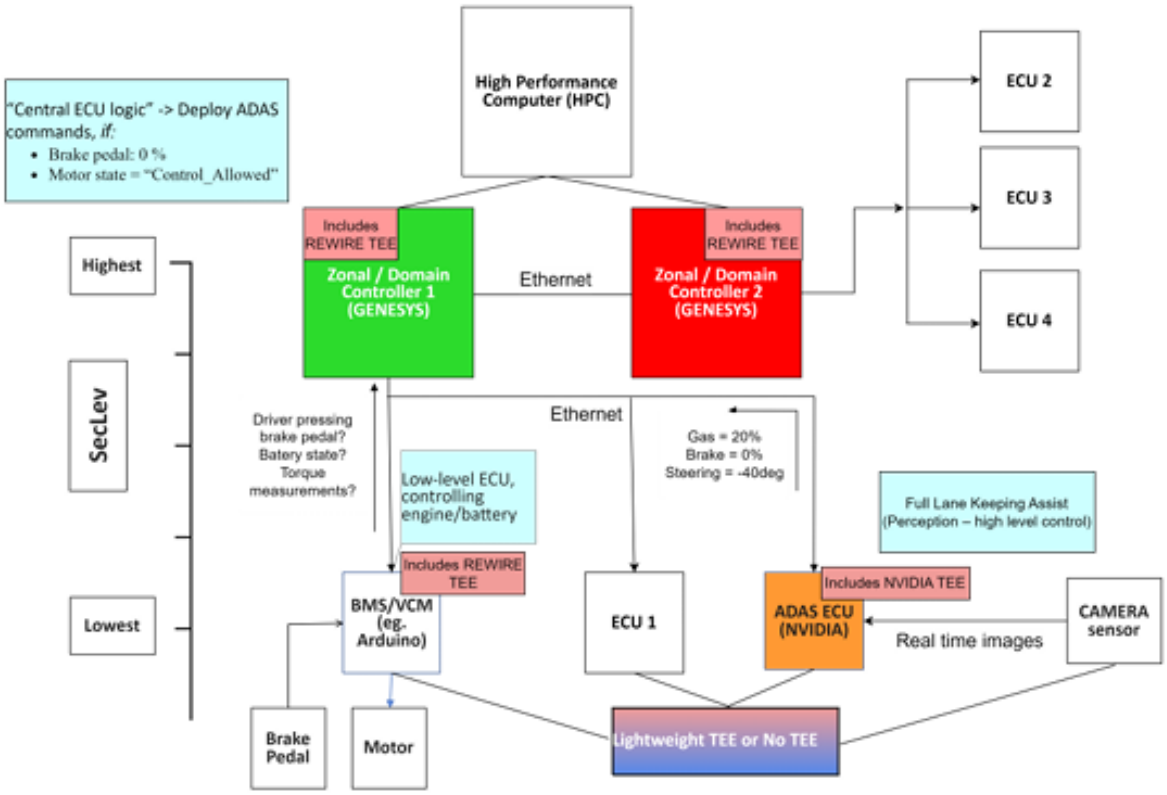


Figure 2: Exemplary vehicle network including automotive use case demo (example 1)

A potential security attack on a given Zonal Control Unit (ZCU), for example Figure 2, Zonal/Domain Controller 1, could compromise a significant portion of the underline sub-system, posing increased problems in comparison to a decentralized architecture, where each function would be grouped to a dedicated ECU (e.g. Motor, Transmission, ABS Control Units, etc.). To compensate that, ZCU1 functionalities could be migrated to a neighboring ZCU2, or to the "centralized", High Performance Computer as well, to either maintain some vehicle functionality or to handle the deployment of a fail-safe mechanisms on a clean and secure environment. Current vehicles do not have such a concept in any kind of form, since typically, if a specific ECU is compromised (e.g., system malfunction or security breach), safety SW intervenes to drive the system to a "minimum" safe state. Nevertheless, ECU safety SW interventions could be also compromised at a successful or extended security breach.

In Figure 3, a sketch of the demo setup along with its components is presented for the actual "test bed" demonstrator of REWIRE.

This demo setup could be thought of as part of a hypothetical vehicle-internal architecture as depicted in Figure 2. In this architecture, ZCU2 would be primarily responsible for a different domain (e.g. ECU2, ECU3, ECU4) than that of ZCU1 and would undertake the migration load. In Figure 2, some exemplary signals of a theoretical ADAS application are also depicted.

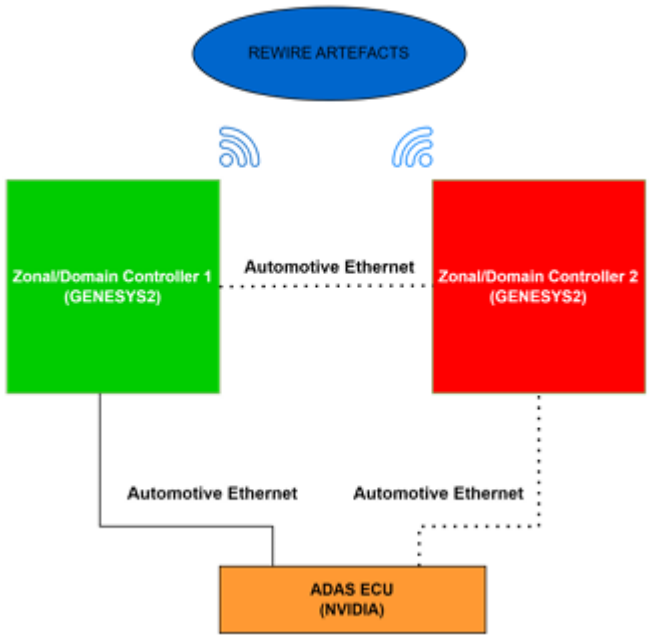


Figure 3: Automotive Use Case Setup

Should there be no possibility to migrate the compromised vehicle functions, the concept still holds validity for the deployment of fail-safe mechanisms. In this case, although no functionality migration is possible, transferring and deploying the compromised SW on a neighboring controller would allow the deployment of the fail-safe mechanisms of the attacked system itself (driver warnings, emergency braking, V2X communication, etc.), but on a "clean" and secure environment. Such an exemplary case is depicted below on Figure 4.

On this setup, an exemplary V2X communication is depicted, with the High-Performance Computer (HPC) receiving V2X data (adversary vehicle speeds, positions, etc.) and transmitting them to ZCU1 (Zonal Controller 1) for subsequent usage at the ADAS ECU. Similarly, ZCU1 handles transmission of speed, motion, etc. data of the vehicle itself to the HPC for informing adversary vehicles. Here, we depict a potential successful compromise on ZCU1 or/and a portion of the underlined ECUs ("X" crossed-section area). Such a case could happen for instance through a successful manipulation of network messages on VCM or ZCU1 to constantly accelerate the vehicle. Upon detecting such an abnormality, a function migration to the neighboring ZCU2 is deployed and due to the inability of gaining access to the engine itself, an appropriate message is transmitted to the adversary vehicles, to notify for the malfunction (e.g. "Vehicle compromised. Yield crossing priority, if needed").

Finally, the REWIRE project, with the automotive application SW and test setup for each board already implemented, excitedly enters the next demonstration phase where these concepts will be integrated with REWIRE security framework and artifacts.

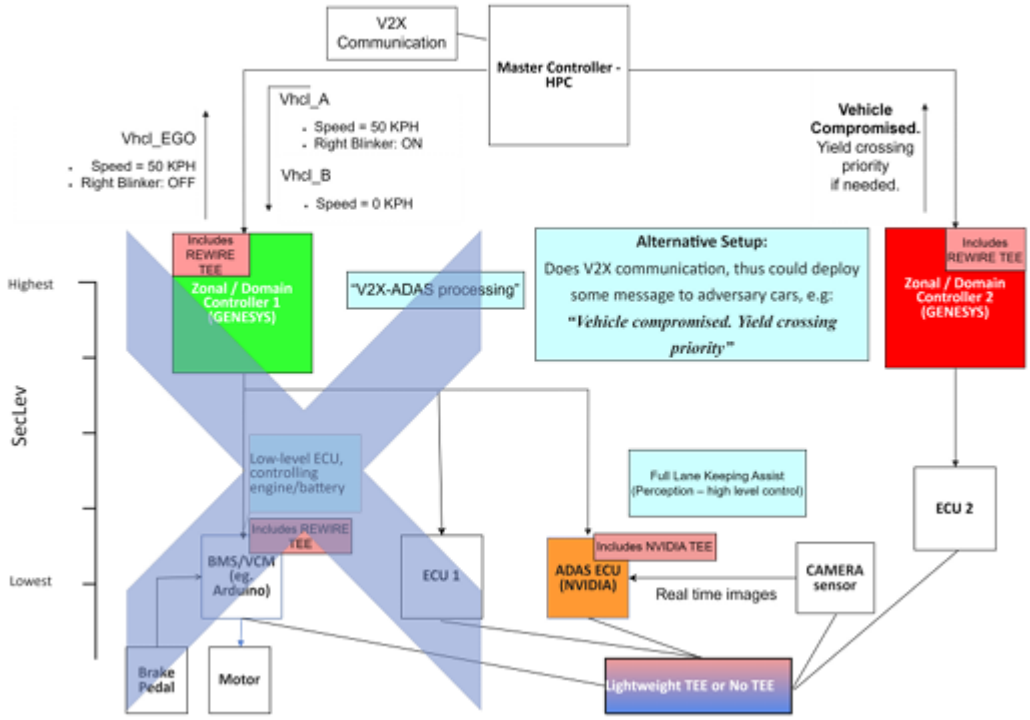


Figure 4: Exemplary vehicle network including automotive use case demo (example 2)

3 Conclusion

In conclusion, the evolution of automotive electrical/electronic (E/E) architectures from decentralized to centralized systems presents both opportunities and challenges. While centralized architectures offer benefits such as reduced network load and fewer ECUs, they also present security challenges. In particular, the potential compromise of a given ECU that manages numerous vehicle functionalities, underscores the necessity for robust fail-safe mechanisms and rollback measures.

This paper presented the innovative concept of "vehicle function migration" to alternative ECUs as a means to ensure the continual operation of critical functions. By leveraging REWIRE's functionalities, this approach enhances the overall reliability, robustness, and security of the vehicle's E/E architecture.

References

[1] V. Bandur, G. Selim, V. Pantelic and M. Lawford, "Making the case for centralized automotive E/E architectures," pp. 1230-1245, 2021.

[2] "REWIRE_D2.1-REWIRE-Operational-landscape-requirements-and-Reference Architecture-Initial-Version-V1.0," [Online]. Available: https://www.rewire-he.eu/wp content/uploads/2024/01/REWIRE_D2.1-REWIRE-Operational-landscape-requirements-and Reference-Architecture-Initial-Version-V1.0.pdf.

[3] "E/E architectures: Optimize their development for automotive projects," [Online]. Available:

https://www.techdesignforums.com/blog/2021/11/17/balancing-the-requirements-of-e-e-architectures-for-automotive-design/

[4] Wikipedia, "Trusted Execution Environment," [Online]. Available: https://en.wikipedia.org/wiki/Trusted_execution_environment.

[5] V. M. Navale, K. Williams, A. Lagospiris, M. Schaffert and M.-A. Schweiker, ""(R)evolution of E/E architectures," SAE International Journal of Passenger Cars-Electronic and Electrical Systems, vol. 8, no. 2, p. 282–288, 2015.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Page 81 – 84

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

A Low Power FPGA Implementation of LDPC Encoder for Space Applications

Christos Sidiras and Vasilis F. Pavlidis
Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece
chrisidi@ece.auth.gr, vpavlid@ece.auth.gr

Abstract

The emergence of 5G, 6G technologies and the expanding Internet of Things (IoT) necessitate advanced data encoding techniques, with Low-Density Parity-Check (LDPC) codes being particularly notable for the high error correction efficiency and low computational complexity. An LDPC encoder with a 4/5 data rate, as specified in CCSDS standards is developed in this work, tailored for transmitting scientific data from a nano-satellite to a ground station. Implemented in SystemVerilog and designed for FPGA deployment, the encoder employs a streaming architecture and the AXI4 Stream protocol for communication interfaces. The resulting solution aims lower power consumption while demonstrating good resource utilization and throughput performance.

Keywords: LDPC code, space communication, quasi-cyclic LDPC encoder, FPGA implementation, power reduction.

1 Introduction

The advent of 5G and 6G technologies and the increasing proliferation of IoT, require all types of telecommunication modulation techniques and data encoding, thereby escalating the demand for efficient data encoding methods. One of these methods is LDPC codes for which there is extensive literature, covering various aspects of their design, implementation, and optimization. LDPC codes exhibit their advantages over other code classes, and are, therefore, integrated in many communication standards, such as 10GBase-T Ethernet (IEEE 802.3an), Digital Video Broadcasting (DVB-S2) and standard for Wireless LANs (IEEE 802.11n). These codes are known for their ability to perform near the Shannon limit, providing excellent error correction capabilities with relatively low complexity [1]. As a result, LDPC codes are highly favored in applications that require high reliability, particularly in space applications. The implementation presented in this paper is part of the telecommunication subsystem of the Acubesat satellite developed as part of the European Space Agency's "Fly Your Satellite! 3" program.

2 Low Density Parity Check codes

LDPC codes are a type of block error-correcting codes introduced by Robert Gallager in the early 1960s. These codes are characterized by a sparse (low percentage of “1”) parity-check matrix H, which represents a set of parity-check equations that valid codewords must satisfy, thereby ensuring

$H * c^T = 0 \ (mod\ 2)$ for any valid codeword c. The generator matrix G, derived from H, is used to produce the codewords, such that $c = u \cdot G$, where u is the input message. Regular LDPC codes are defined by the parameters (w_r, w_c), where w_r is the number of ones in each row and w_c is the number of ones in each column of the parity-check matrix H [**2**].

The sparsity of the parity-check matrix H is essential for facilitating efficient encoding and decoding processes. Iterative decoding algorithms, such as belief propagation, leverage this sparsity to identify and correct errors with minimal computational complexity. Deriving the generator matrix G from H is a common method for encoding LDPC codes. Alternative encoding methods [**3**] include systematic encoding techniques that directly construct codewords to satisfy the parity-check equations

3 LDPC Encoder Design

The object is to design a robust and fully functional LDPC encoder with a data rate of 4/5, as defined in CCSDS standards [**4**]. The encoder exclusively provides the scientific data of the transfer from a nano-satellite experiment to the ground station. The intellectual property core (IP) to be extracted follows a streaming architecture, making use of the AMBA AXI4-Stream protocol [**5**] for its interfaces. Developed in System Verilog, the encoder's IP core is intended for deployment on a Field Programmable Gate Array (FPGA).

3.1. LDPC Design Specifications

Space communication protocol recommends using two classes of LDPC codes. One class is AR4JA codes which offers a variety of block lengths (1,024, 4,096 and 16,384 bits) and rates (1/2, 2/3, 4/5) and exhibit good performance in terms of $Eb/N0$, emphasizing reliable communication in noisy environments. The other class is $C2$ code, which is a regular (8,176, 7,156) LDPC code that achieves high transmission speeds.

Parity check matrix of the selected AR4JA code with 4,096 block size and 4/5 rate is a quasi cyclic matrix consists of 512 × 512 sparse submatrices, such as identity, zero or permutation matrices all strictly implied by CCSDS standard. Generator matrix G produced by the aforementioned parity check matrix has the form of $G = [I_{4096}\ W]$, where I_{4096} is the square identity matrix of dimension equal to 4,096 and W is a dense 4,096 × 1,024 matrix of circulants each one of size 128 × 128. This type of LDPC codes can be assumed to be systematic, meaning that the first 4,096 bits of information encoded are the same as the input bits while next 1,024 bits are the parity ones. Apart from block length and information rate, detailed specifications are listed in Table 1.

3.2. Design Process

The IP core must have a specific set of interface signals in order to be compliant with AXI4 Stream protocol. A slave interface accepting the message and a master interface exporting the codeword are designed and only the signals that attribute to the handshake are used, as depicted in Figure 1(A). The provided Figure 1(B) represents a Recursive Convolutional Encoder (RCE) recommended by the standard. A detailed description of the encoder operation follows:

- First row bits of generator matrix, stored in memory, are loaded in registers from right to left. These bits are multiplied with the first single bit of the message and the result is stored in registers inside RCEs.
- The second bit of the message enters sequentially the encoder and is multiplied with the first-

row bits of generator matrix. The result is added exclusively with the previously stored one. This process is repeated for 128 cycles and for eight consecutive times.

- Every 128 passing cycles, the next row circulants of the generator matrix are loaded.
- When all 4,096 bits of the message have been processed, the encoder proceeds to the parity phase. The memory is disabled, demuxes change their output and the encoder starts executing a right shift operation for 1,024 cycles extracting calculated parity bits.

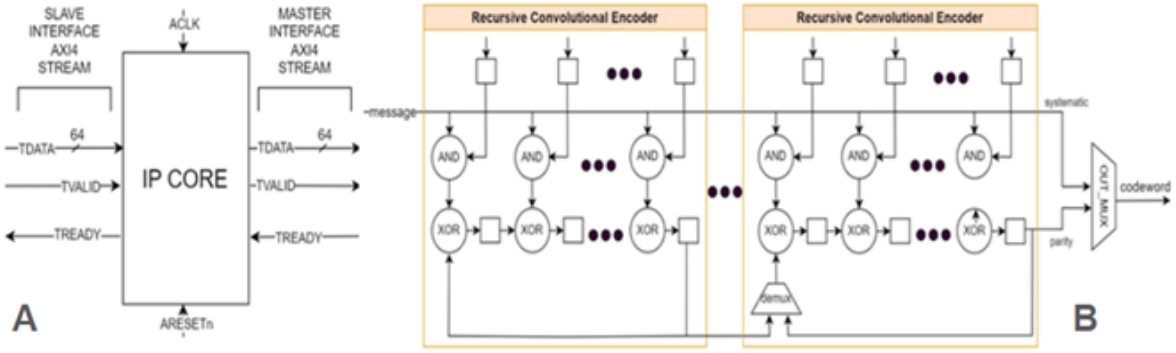


Figure 1: LDPC encoder (A) interface block diagram of the encoder, (B) internal architecture

The three subsystems (AXI-slave, encoder, AXI-master) within the IP core are almost independent and can work simultaneously, thus accelerating the encoding process. Rows of generator matrix are stored in a block memory due to the high efficiency in managing large data sets and the ability to provide rapid access times. Additionally, Block RAM (BRAM) of FPGAs offers increased parallelism and lower power, both of which are vital for the specific design. The FSM controls the different phases of the encoder and communicate with the FSMs of the interfaces.

Further modifications are exploited to reduce the power consumed by the circuit. Memory, registers, and FIFOs are disabled between long periods of inactivity. FIFOs of interfaces are implemented with BRAM to save more slice registers and power and all designs with BRAM operate with gated clocks. The latter technique enhances significantly (up to 80%) power savings since memory of this implementation occupies most resources on FPGA, as can be also seen in Table 1.

3.3. Encoder Verification

The primary functional verification of the LDPC encoder is conducted using a C++ script. This script demands a report file in .txt format, where details, such as seed, index of test cases, and captured input messages and output codewords, are documented. These report files hold information generated by either simulated testbenches or during measurements on the FPGA. The C++ script scans these report files for the relevant messages. For each message, the script triggers another C++ software program that simulates a software version of the LDPC encoder, and compares the resulting SW-generated codeword with the corresponding HW-generated codeword. This process is iterated for each tested message and stores the comparison results.

Verification aims to achieve approximately 100% code and functional coverage using the Xcelium and Integrated Metrics Center (IMC) tools by Cadence. Code coverage encompasses several aspects: block, expression, toggle and FSM. Functional coverage, on the other hand, is divided into data

oriented, performed through covergroups that monitor input and output data and control-oriented verification, carried out using SV Assertions that mostly track AXI4-Stream transactions.

4 Evaluation and Integration on FPGA

The proposed LDPC encoder is also integrated on the XC7Z020CLG400-1 FPGA of family Zynq-7000 of AMD Company [6]. The encoder has been enriched with Processing System (PS), AXI Direct Memory Access (DMA), interconnects and Integrated Logic Analyzer (ILA) to monitor waveforms. Vivado 2023.2 and Vitis 2023.2 are used to generate the encoder bitstream and to assist bit transactions between Software and Hardware, respectively. The implementation results of the encoder (not including the remainder design) are reported in Table 1. Power consumption is estimated by the provided tools, however, power measurements on board include additional components, such as voltage regulators and active peripherals, leading to deviations and therefore are omitted. Regarding throughput, estimations indicate the maximum achievable performance, whereas real measurements reflect the actual performance tested with a 10 ns period cycle.

Logic Utilization	**Used**	**Available**	**Utilization**	**[7] Xilinx XC6VLX240T rate=1/2, length=2048**	**[8] Kintex UltraScale KCU105**	**Specifica -tions**
Number of Slice Registers	1249	106400	1.17%	3317	3670	< 2500
Number of Slice LUTs	1228	53200	2.31%	2211	6770	< 5500
Number of Block RAM	16.5	140	11.79%	32	1	-
Power consumption	**Estimation**			**Estimation**	**Measurement**	
Power [mW]	117			N/A	1371	< 2500
Performance	**Estimation**	**Measurement**		**Estimation**	**Estimation**	
Throughput [Mbps]	136	99.3		180	30000	> 0.5

Table 1: Design specifications and implementation results

5 Conclusion

In this paper, the implementation of an LDPC encoder, fully applicable to CCSDS standard for deep space applications and to AXI4-Stream communication protocol is presented. The circuit exploits the structure of generator matrix of quasi-cyclic LDPC code and aims for low power, while in parallel, performance and resource utilization objectives are satisfied. Furthermore, the proposed circuit is verified 100% with commercial tools and implemented on a space-grade FPGA.

References

[1] T.K. Moon, Error Correction Coding Mathematical Methods and Algorithms. Utah , USA, 2005.

[2] S. J. Johnson, Low-density parity-check codes. Cambridge, USA: Cambridge University Press, 2009.

[3] W. C. Huffman and V. Pless, Fundamentals of Error-Correcting Codes. Cambridge University Press, 2003.

[4] Consultative Committee for Space Data Systems, "TM synchronization and channel coding recommended standard," Standard 131.0-B-5, 2017.

[5] [ARM Ltd., "AMBA® 4 AXI4-Stream Protocol," Specification Version 1.0, 2010.

[6] AMD, "Zynq-7000 SoC Data Sheet ," Overview DS190, 2023.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 85 – 90

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Evaluation and DSP Benchmarking of the European Radiation-Hardened NG-ULTRA FPGA

Anastasios Xynos[1], George Lentaris[2], and Dimitrios Soudris[1]
[1]*National Technical University of Athens, Athens, Greece*
[2]*University of West Attica, Athens, Greece*
axynos@microlab.ntua.gr, glentaris@uniwa.gr, dsoudris@microlab.ntua.gr

Abstract

The advent of advanced space applications has transformed the conventional computing systems used in space missions. Prioritizing reliability alongside enhanced performance-per-Watt, radiation-hardened FPGAs have become the preferred choice. This paper focuses on evaluating NanoXplore's NG-ULTRA, the first European radiation-hardened SoC FPGA, for its capability to accelerate high-performance DSP algorithms in space applications. The proposed development and testing methodologies aim to deliver efficient implementations while also assessing the new NG-ULTRA hardware features. The findings indicate that NG-ULTRA offers competitive resource utilization and performance, making it a highly promising option for European space missions.

1 Introduction

The space industry is continually evaluating advanced embedded platforms to handle the growing data and computational demands of modern space applications. FPGAs have emerged as a key solution for on-board data processing [1], offering excellent performance-per-Watt in tasks like Vision-Based Navigation (VBN), Earth Observation (EO), and Satellite Communications (SatCom) [2]. In environments requiring resistance to radiation, thermal fluctuations, and vibrations, radiation-hardened FPGAs are preferred over traditional CPU-based processors like the RAD750 and AT697F due to their higher reliability and processing efficiency.

The high-density space-grade FPGA market is largely controlled by two U.S. companies: AMD/Xilinx, offering devices such as Virtex-4QV, Virtex-5QV, and RT Kintex US, and Microchip/Microsemi, with their RTAX, RTG4, and RT PolarFire models. Notably, only two space grade SoC FPGAs are currently available: AMD's XQR Versal and Microchip's RT PolarFire SoC. Recently, NanoXplore, a European company, has entered the scene with its BRAVE line of radiation hardened FPGAs, some of which have already been deployed in space missions [3], [4]. The newest addition, NG-ULTRA, is Europe’s first 28-nm radiation-hardened SoC FPGA[5]. This FPGA integrates a quad-core ARM-R52@600MHz processor and a variety of programmable logic resources, making it suitable for space missions.

NanoXplore's radiation-hardened FPGAs have attracted interest from both industry and academia ([6], [7], [8], [9], [10], [11], [12]), with ESA testing and evaluating these devices. We developed a methodology to support and test the development of these FPGAs, evaluating both the hardware and

the software tools. Our approach systematically explores tool-level functionality and HDL circuits to achieve efficient implementations on NG-ULTRA, with a focus on DSP algorithms for computer vision and signal processing in space applications.

Our work focuses on a new methodology for supporting and testing space-grade FPGAs like NG ULTRA, addressing all the hardware advancements. By developing and testing DSP algorithms from computer vision and signal processing for space applications, we demonstrate the viability of NG ULTRA for use in space. This paper highlights an enhanced testing methodology based on works of previous projects [6], [7], compares NG-ULTRA's performance to established FPGAs, and reports benchmarking results, showing that NG-ULTRA provides competitive resource utilization and sufficient performance for radiation-hardened environments.

2 Testing Methodology & Development

The proposed methodology is built around NanoXplore's programmable logic toolchain, utilizing the Impulse software for synthesis and place-and-route operations. The synthesis methodology, depicted in Fig. 1, involves three primary stages: parametric configuration of the DSP kernels, exploration of the tool's settings, and an HDL-level deep dive into the DSP kernels. Initially, the configuration phase focuses on fine-tuning algorithmic parameters like input size, input partitioning, convolution dimensions, bit-width of the datapath, and the degree of parallelization, all tailored to match the architecture and resources of the target BRAVE FPGA. Once configured, the synthesis is performed using third-party vendor tools to generate baseline results for comparison.

In the tool-level exploration stage, we first perform an initial synthesis with the default settings to identify any inherent issues. Following this, we systematically explore all available Impulse synthesis settings, applying them both individually and in various combinations, then assessing the resulting synthesis netlists. During this phase, we investigate settings affecting mapping effort, register duplication, arithmetic, logic, and memory targets, DSP usage, and FSM encoding style. If the tool produces errors or unexpected results, alternative settings or changes in the HDL code are employed to resolve these issues. The resource usage of the most optimized, error-free netlist is then compared with the results obtained from third-party FPGA tools. If any major discrepancies are found, adjustments are made either by tweaking the tool settings or by proceeding to the next stage.

The HDL-level exploration phase involves breaking the DSP kernel into smaller subcomponents for individual testing through a recursive process. This detailed examination helps in identifying optimization challenges or errors that may be difficult to detect when analyzing the kernel as a whole. In this phase, we employ template-based coding practices, using NanoXplore's HDL templates to implement the kernel's building blocks. Each HDL block undergoes post-synthesis simulation to ensure accuracy. Once all issues related to errors or optimizations are resolved, we return to the tool-level exploration to evaluate the entire kernel once again. If no viable solution is obtained after these explorations due to errors, excessive resource use, or failed verification, feedback loops (represented by red dashed lines in Fig. 1) are used to restart the process from the configuration phase with a new kernel setup.

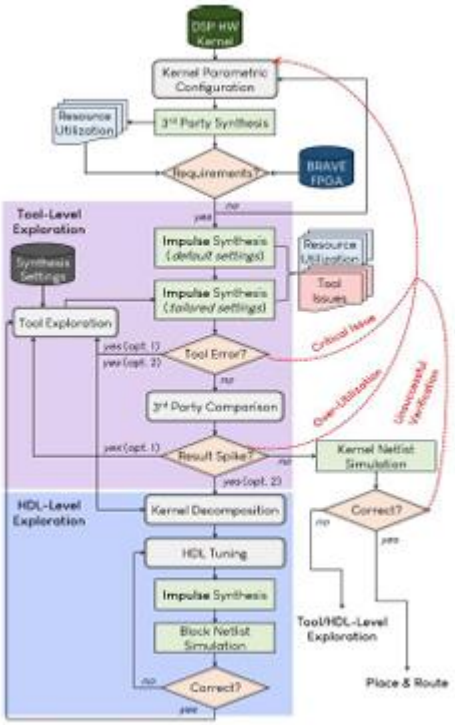


Figure 1: DSP kernel synthesis methodology for BRAVE FPGAs [13]

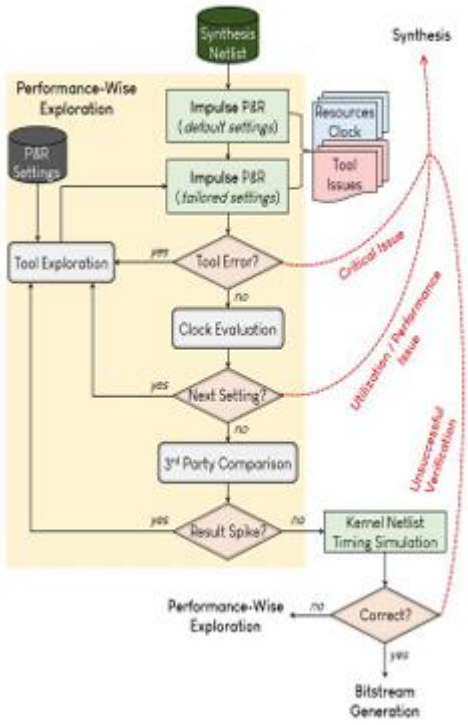


Figure 2: DSP kernel Place & Route methodology for BRAVE FPGAs [13]

The place & route methodology, shown in Fig. 2, starts with the error-free synthesis netlist and focuses on optimizing tool settings to achieve the best possible clock frequency. This involves exploring placement, routing, and physical constraints, while also evaluating specific regions on the FPGA floorplan for mapping. We also fine-tune timing settings, including clock constraints, false paths, maximum delays, and timing-driven placement and routing. As with synthesis, comparisons with third-party tools, simulations, and feedback loops are used to resolve issues like timing violations or inefficient placement, with adjustments made as necessary.

3 Experimental Evaluation & Performance Results

For our experimental evaluation, we used several DSP kernels that are computationally and memory-intensive, all commonly found in modern space applications. Specifically, the Harris Corner Detector and Canny Edge Detector are used for feature detection, such as in vision-based navigation and docking systems, GAD-Disparity and SpaceSweep handle stereo matching for depth extraction in 3D scene reconstruction. The QAM Modulator & Demodulator is employed for 16-order modulation and demodulation in 5G satellite communications. The computer vision kernels were tested using 1 Megapixel input images, and the telecom kernel (QAM) was configured for 16-order modulation. For comparison, we ran tests on 3rd-party FPGAs with similar specifications (AMD/Xilinx Zynq UltraScale+ ZU09CG, Microchip PolarFire MPFS460T, and Intel Arria 10 SX480) alongside NanoXplore's smaller FPGAs (NG-MEDIUM, NG-LARGE). We used Impulse v23.5.1.2 for the European FPGA platforms. Table 1 summarizes the results, comparing NG-ULTRA's performance with the average results of

these other FPGAs. For NG-ULTRA, the reported LUT values include both 4 LUTs and carry logic units, while for 3rd-party FPGAs, we converted their 6-LUT and 8-LUT architectures into 4-LUT equivalents for fair comparison. This was done based on approximate conversion rates: one 6-LUT equals about 1.5x 4-LUTs, and one 8-LUT equals around 2x 4-LUTs. NG ULTRA's results reflect its optimized performance after our systematic exploration.

For the Harris Corner Detector, NG-ULTRA slightly exceeds the average 3rd-party LUT usage (+6.6%), with comparable DFF and DSP utilization. NG-ULTRA performs better in DFF efficiency and uses less block memory (108 RAMBs vs. 182), thanks to its larger RAMB size. We also applied a mapping directive to prioritize DSPs for multipliers and carry logic for adders, which improved NG ULTRA's clock frequency by about 8.7%.

For the Canny Edge Detector, none of the FPGAs utilized DSPs. NG-ULTRA, however, used significantly fewer LUTs (about three times less) and DFFs (2.8 times less) compared to the 3rd-party average, mainly because of additional logic in PolarFire, which inflates the average. NG-ULTRA also showed more efficient RAMB usage, consuming around 60% of the average 3rd-party value. Additionally, manual pad placement was used to further optimize NG-ULTRA's clock frequency, resulting in a 2.6% improvement.

Table 1: DSP Kernel Resource Utilization: NG-ULTRA FPGA [13] vs 3rd-Party FPGAs (Average Values)

Kernel	I/O	LUT (NG-U)	LUT (3rd-P)	DFF (NG-U)	DFF (3rd-P)	DSP (NG-U)	DSP (3rd-P)	RAMB (NG-U)	RAMB (3rd-P)	MHz (NG-U)	MHz (3rd-P)
Harris	1024×32, 8/32b	16882	15831	19470	21016	92	81	108	182	80	183
Canny	1024×1024, 8/4b	2090	5951	2029	5631	0	0	177	293	67	170
Space Sweep	1024×32, 8/32b	11882	17333	10788	15905	50	36	278	401	58	188
GAD-Disparity	1024×32, 8/10b	1306	12817	3838	14369	494	165	105	169	70	217
QAM16	N×8, 8/8b	19818	25119	29193	22463	256	162	0	0	92	303

For SpaceSweep, a similar trend emerges: NG-ULTRA shows lower usage in LUTs, DFFs, and RAMBs compared to 3rd-party FPGAs, with the exception of DSP utilization, which is about 38% higher. In the case of GAD-Disparity, NG-ULTRA demonstrates a significant reduction in both LUTs (1306 vs. 12817) and DFFs (3838 vs. 14369), largely due to our design choice of mapping all multipliers and adders to DSPs (494 vs. 165), aimed at improving clock frequency.

In the QAM16 test, NG-ULTRA uses fewer LUTs but higher DFF and DSP resources. By implementing manual pad placement, we boosted clock frequency by 6%. Although QAM16 delivers NG-ULTRA's highest clock speed at 92 MHz, this remains about one-third of the frequency achieved by 3rd-party FPGAs, a trend that persists across most kernels.

Table 2: DSP kernel throughput on the three generations of NG-ULTRA FPGAs [13]

Kernel	NG-MEDIUM	NG-LARGE	NG-ULTRA
Harris	--**	4 FPS	11 FPS
Canny	15 FPS	9 FPS	21 FPS
SpaceSweep	92 MPDS	67 MPDS	129 MPDS
GAD-Disparity	--**	7 MPDS	20 MPDS
QAM16	454 MBPS	1415 MBPS	2959 MBPS
Avg. Clock	49 MHz	30 MHz	65 MHz

* FPS: Frames Per Second, MPDS: Megapixel Disparities Per Second, MBPS: Megabits Per Second.
** Resource overutilization for the targeted accuracy (even for smaller input image).

Table 2 presents the throughput results for the DSP kernels on NanoXplore's radiation-hardened FPGA models. The kernel configurations mirror those in Table 1, with the exception of NG MEDIUM, where the 1-Megapixel image is split into four 512×512 segments for Canny and sixteen 256×256 segments for SpaceSweep, both processed sequentially. In NG-MEDIUM, the QAM16 parallelization factor is also reduced from 8 to 2. For NG-LARGE, SpaceSweep processes four 512×512 images consecutively. It's important to note that certain kernels could not be accommodated by NG-MEDIUM and NG-LARGE due to excessive resource utilization (LUTs, DSPs, or cycles) even with reduced input sizes.

In the cases of Canny and SpaceSweep, NG-MEDIUM outperforms NG-LARGE in throughput purely due to its higher clock frequency. NG-MEDIUM completes smaller image runs more quickly than NG-LARGE processes a full-sized image. Meanwhile, NG-ULTRA achieves 2.5 to 3 times the throughput of NG-LARGE. For Canny and SpaceSweep specifically, NG-ULTRA offers a 40% throughput increase compared to the serial runs of NG-MEDIUM. These throughput gains are closely tied to clock frequency, with NG-ULTRA operating at 1.4 times the clock speed of NG-MEDIUM and nearly 2.2 times that of NG-LARGE. Overall, the throughput of these kernels meets the expected performance requirements for space applications. However, further improvements could be made with custom kernel redesigns, such as increased parallelization and manual placement tailored for NG ULTRA, to achieve even greater throughput.

4 Conclusion

In this paper, we introduced a methodical approach for designing, testing, and evaluating high performance DSP algorithms on the latest European radiation-hardened FPGAs. Using our methodology, we achieved efficient and well-balanced implementations on NG-ULTRA, showing substantial performance improvements over previous models. Our future research will target the hardware/software co-design for space applications using the NG-ULTRA SoC FPGA.

References

[1] G. Lentaris and others, “High-Performance Embedded Computing in Space: Evaluation of Platforms for Vision-Based Navigation,” AIAA Journal of Aerospace Information Systems, vol. 15, no. 4, pp. 178–192, 2018, doi: 10.2514/1.I010555.

[2] O. Kodheli et al., “Satellite Communications in the New Space Era: A Survey and Future Challenges,” IEEE Communications Surveys & Tutorials, vol. 23, no. 1, pp. 70–109, 2020.

[3] R. D. Torrijos, “GMV and BRAVE FPGAs: From Studies to Flight Hardware Use,” in NanoXplore’s BRAVE Days, 2023, pp. 1–32.

[4] NanoXplore, NG-MEDIUM and FUSIO-RT: Mission completed on the dark side of the Moon. 2024. [Online]. Available: https://nanoxplore.com/index.php/2024/06/07/mediumfusioonmoon/ [

[5] E. Lepape and M. Le Penven, “New Generation of Rad-Hard SoC FPGA,” in Space FPGA Users Workshop (SEFUW), 2023, pp. 1–28.

[6] K. Maragos and others, “Evaluation Methodology and Reconfiguration Tests on the New European NG-MEDIUM FPGA,” in NASA/ESA Conf. on Adaptive Hardware and Systems (AHS), 2018, pp. 127–134. doi: 10.1109/AHS.2018.8541492. [

[7] V. Leon and others, “Development and Testing on the European Space-Grade BRAVE FPGAs: Evaluation of NG-Large Using High-Performance DSP Benchmarks,” IEEE Access, vol. 9, pp. 131877–131892, 2021, doi: 10.1109/ACCESS.2021.3114502.

[8] Y. Barrios and others, “SHyLoC 2.0: A Versatile Hardware Solution for On-Board Data and Hyperspectral Image Compression on Future Space Missions,” IEEE Access, vol. 8, pp. 54269 54287, 2020, doi: 10.1109/ACCESS.2020.2980767.

[9] V. Leon and others, “Systematic Evaluation of the European NG-LARGE FPGA & EDA Tools for On-Board Processing,” in European Workshop on On-Board Data Processing (OBDP), 2021, pp. 1–8.

[10] K. Bravhar and others, “SerDes Integrated Into the SpaceWire Interface Helps in Achieving Higher Data Rates,” IEEE Aerospace and Electronic Systems Magazine, vol. 38, no. 11, pp. 16–27, 2023.

[11] A. Portaluri and others, “Design Techniques for Multi-Core Neural Network Accelerators on Radiation-Hardened FPGAs,” in Int’l. Symposium on Parallel and Distributed Computing (ISPDC), 2023, pp. 16–22.

[12] E. Danard and A. Comolet-Tirman, “OBC-Ultra, the Rad-Hard NG-Ultra-based On Board Computer for Future Applications,” in European Data Handling & Data Processing Conference (EDHPC), 2023, pp. 1–4.

[13] V. Leon et al., “Development of High-Performance DSP Algorithms on the European Rad-Hard NG-ULTRA SoC FPGA,” in 2024 31th IEEE International Conference on Electronics, Circuits and Systems (ICECS), IEEE, Nov. 2024, pp. 1–4.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 91-97

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Autonomous Multi Source Energy Harvesting Multi-Sensor

Ioannis Masklavanos[1]*, Theodoros Georgiadis[2], Vasiliki Naskari[2], Fotios Vartziotis[2] and Gregory Doumenis[2]
[1] *Metis Cyberspace Technology, Athens, Greece*
[2]*Department of Informatics and Telecommunications, University of Ioannina, Arta, Greece*
giannismaskl@gmail.com

Abstract

The paper presents the design and implementation of an autonomous multi-sensor system powered by multi-source energy harvesting for structural health monitoring applications. This system monitors the strain of steel structures on ships while collecting environmental data to assess the correlation with steel degradation. It operates autonomously, using both photovoltaic and electromagnetic energy harvesters, and includes an advanced energy management algorithm to optimize power consumption and ensure continuous operation. The paper details the system's architecture, including energy capture, storage, and distribution strategies, and highlights its efficiency and adaptability to variable environmental conditions. It also explores the power consumption of individual components and discusses the results of simulations and bench experiments conducted to validate the system's energy harvesting capabilities. The findings offer valuable insights into energy autonomy for long-term operation in harsh environments, with implications for future advancements in self-powered monitoring systems.

1 Introduction

In today's rapidly evolving technological landscape, the demand for sustainable and self-powered and energy autonomous sensor systems has been intensified. These devices, capable of capturing and utilizing ambient energy from their surroundings, offer a promising solution to the challenges posed by traditional battery-powered systems. By integrating multi-sensor systems, these devices can enhance their efficiency, adaptability, and reliability, enabling a wide range of applications in various sectors. However, the development and deployment of self-powered and energy autonomous sensor systems are not without their challenges. Factors such as environmental variability, energy conversion efficiency, storage and management, cost and scalability, reliability and durability must be carefully considered. Despite these limitations, the potential benefits of these systems are significant.

In this work, we present the design and implementation of a self-powered and energy autonomous multi-sensor system for structural health monitoring applications. Specifically, the system is designed to monitor the strain of steel structures on ships while simultaneously collecting environmental data to investigate their correlation with steel degradation. The system operates autonomously, acquiring and processing data locally before transmitting it to the proprietary platform owned by METIS

Cyberspace for storage and further analysis. To ensure perpetual operation, the system incorporates advanced energy management algorithms that optimize power consumption and maximize energy harvesting. The primary objective of this paper is to introduce the architecture of the proposed system and highlight its unique features in comparison to existing solutions. Additionally, we will delve into the strategies employed to achieve energy autonomy.

The paper is organized as follows: Chapter 2 presents the system architecture, Chapter 3 provides an overview of each subsystem, and Chapter 4 discusses the results and findings. Finally, Chapter 5 sums up and sets implications for future research.

2 System architecture

The system architecture of an energy harvesting system is a crucial aspect that determines its overall efficiency, reliability, and adaptability to different environments. This chapter delves into the design and implementation of the proposed Wireless Autonomous Device System Architecture (WADSA), focusing on its unique power supply architecture. Section 1.1 presents a typical power supply architecture for reference, while Section 1.2 provides a detailed description of WADSA's innovative approach to energy harvesting, storage, and distribution.

2.1. Typical architecture

Figure 1 presents a typical power supply architecture for an energy harvesting system. It illustrates the sequential flow of energy from the source to the load. Ambient energy is captured by harvesters, such as photovoltaic panels or electromagnetic harvesters. A power path controller manages the energy flow, optimizing its distribution and utilization. A DC/DC converter adjusts the voltage to match the requirements of the load. Excess energy can be stored in an accumulator, such as a battery or supercapacitor, for later use during periods of low ambient energy availability. This architecture is commonly employed in autonomous devices and systems that rely on renewable energy sources, offering a sustainable and environmentally friendly power solution.

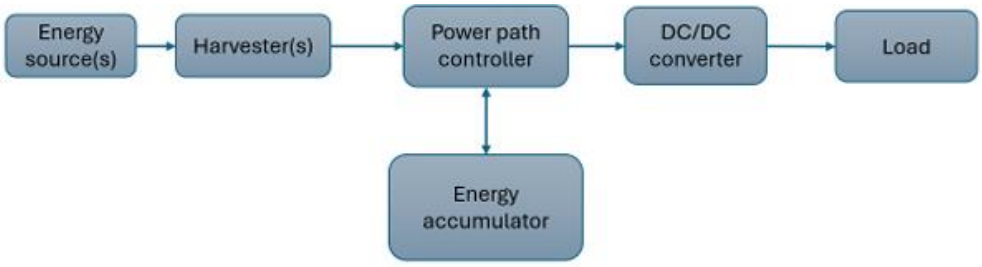


Figure 1: Typical power supply architecture

2.2. WADSA architecture

The system's energy harvesting process is based on integrating multiple energy sources to ensure continuous and reliable power supply. In this design, both photovoltaic (PV) cells and electromagnetic (EM) harvesters are used to capture ambient energy. These energy inputs are managed by a harvester unit (LTC3331), which intelligently controls the flow of energy, directing it to a secondary battery for

storage or directly to a supercapacitor for immediate use. The supercapacitor plays a key role in providing quick bursts of energy when needed, while the power gate controller regulates the flow of energy to the load, ensuring efficient consumption. By balancing energy capture, storage, and distribution, this system ensures that the connected devices can operate autonomously, even in variable energy environments, making it ideal for applications requiring long-term, energy-independent operation.

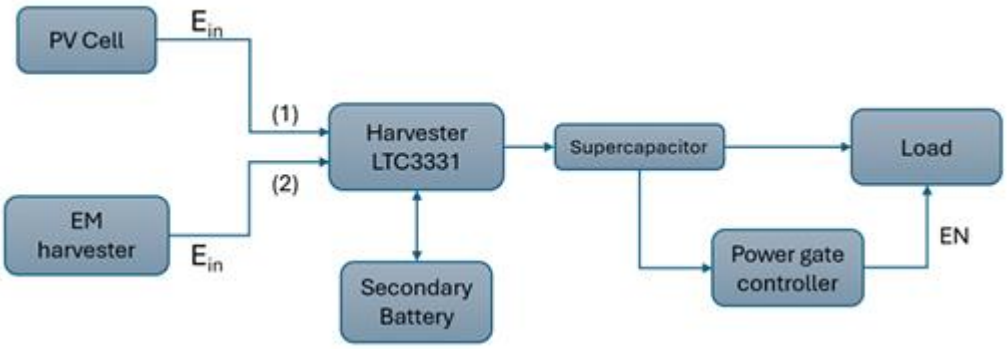


Figure 2: Our power supply architecture

2.3. Application

The main application for the developed system is steel health assessment in naval environments. The used method [1] is a cutting-edge approach that offers a non-invasive way to monitor the strain of steel structures on ships. Unlike traditional methods that require physical access to the steel, this innovative system utilizes advanced sensors to measure strain. This not only minimizes the risk of damage to the steel but also significantly reduces the power consumption required for operation. Additionally, the system is designed to collect environmental data concurrently, allowing for a comprehensive investigation into the relationship between external factors and steel degradation. This dual functionality provides valuable insights for optimizing maintenance schedules and ensuring the structural integrity of ships.

3 Power system design and implementation

Section 3 comprehensively explores WADSA's system design and implementation. This section delves into the intricacies of energy source selection, power management strategies, and multi-source integration.

3.1. Power source and management

The selection of energy sources for WADSA was guided by an evaluation of environmental availability, power density, conversion efficiency, and cost-effectiveness. Indoor light, a ubiquitous and consistent energy source in various indoor environments, was chosen for its potential to provide a reliable and sustainable power supply. The ability of indoor light to penetrate through various materials and its presence in diverse indoor settings made it an attractive option for WADSA's application.

To assess the suitability of different indoor PV cells for WADSA, a comparative analysis was conducted. The power output of several commercially available cells was measured under varying light conditions, as depicted in Figure 3.

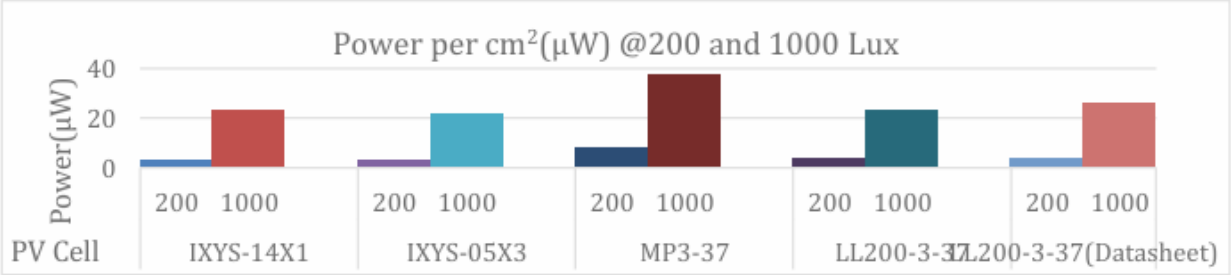


Figure 3: PV cell output at different light intensity

The results indicate that the MP3-37 cell consistently demonstrated superior performance, exhibiting higher power output at both 200 and 1000 lux compared to the other options.

Given that vibrations are abundant in operational ships, the second selected power source for the system were mechanical vibrations. To convert these vibrations into useful electrical power an electromagnetic harvester (EMH) was developed specifically for the needs of the project[2].

3.2. Multi source energy harvesting

In Figure 4, the PCB schematic of the LTC3331 PV + EMH harvesting circuit is presented. A PV cell and an EMH are connected simultaneously to the LTC3331 through connectors J2 and J3. This is a unique configuration, which is not predicted in the literature. As such, the capability of the LTC3331 to receive power simultaneously by PV and EMH is investigated through simulation and bench experiments in section 4.

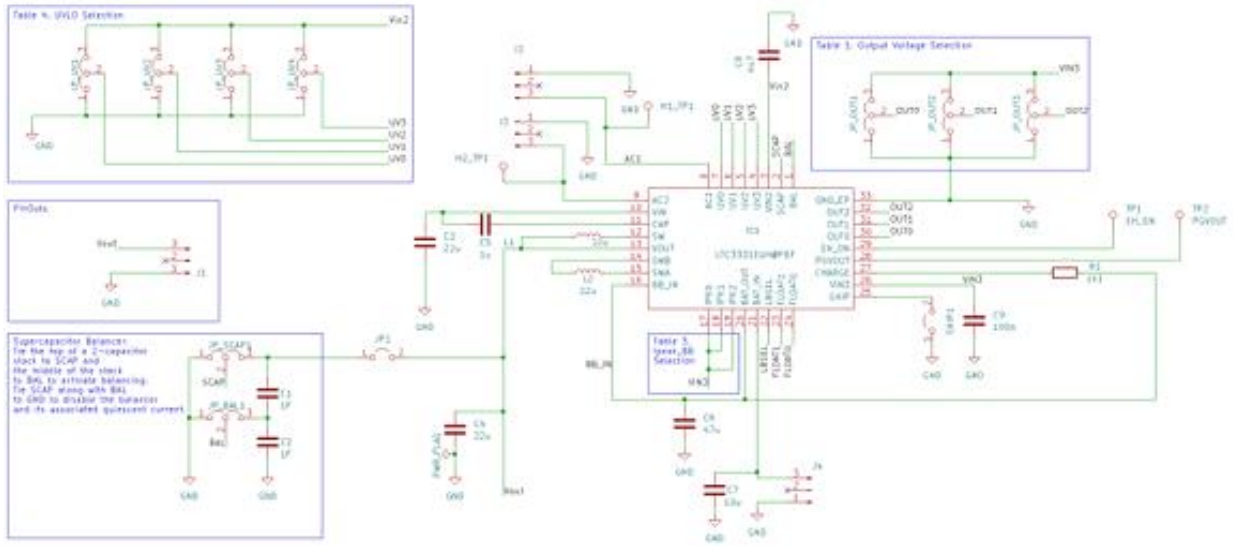


Figure 4: Power supply schematic

3.3. Power consumption analysis

To better understand the power need of the system, a power consumption analysis was conducted. The major components contributing to power consumption are the two microcontrollers, the main microcontroller that reads the data and the RF module. In Table 1, the power consumption of each system component is depicted.

Component	**Operating voltage (V)**	**Current consumption (mA)**	**Power consumption (mW)**
ESP32-C6_DevKit	3.3	33.3	109.89
ESP32-C6_DevKit (LPM)	3.3	0.208	0.6864
Digi Xbee (Tx)	3.3	35.7	117.81
Digi Xbee (LPM)	3.3	0.141	0.4653
Environmental Sensor	5	15	75
IMU	3.3	13	42.9
Steel stress sensor	3.3	3.6	11.88

Table 1: Component current consumption

The energy required for a single active period of the system is 72.45mJ. To optimize the power consumption, a strategy that adjust the duty cycle was developed. However, the energy management algorithm will not be described in this work.

4 Discussion

4.1. System modeling and simulation

PV and EMH are connected to the LTC3331 at the endpoints of the internal rectifier bridge (connection points AC1 and AC2 in figure 5) and charge capacitor C3 attached to VIN. Figure 5 left is a cutout of the LTC3331 block diagram and Figure 5 right is the equivalent Ltspice circuit of the harvester input. Charge from C1 is transferred to supercapacitors (C1 and C2 in Fig. 4) at the output of LTC3331 (pin VOUT in fig. 4), through inductor L1 of the internal buck converter. For an output of 3.3V, A voltage of 4.4V at C3 triggers the buck converter, which discharges C1 down to approx. 3.6V before stopping. The C3 discharge process re-starts when C3 reaches 4.4V again. The Buck operation/C3 discharge process is simulated in LTspice by a voltage-controlled switch S1 (with hysteresis) and the Rs3 “load” resistor.

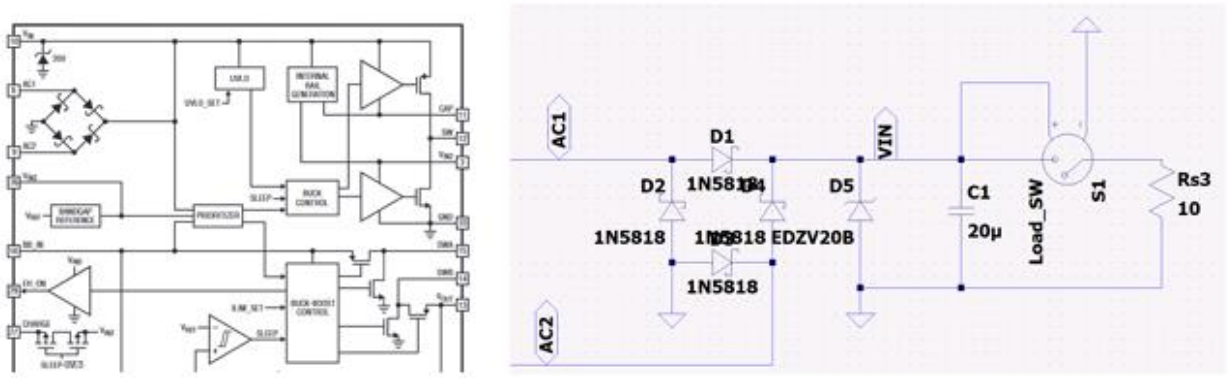


Figure 5: PV cell output at different light intensity

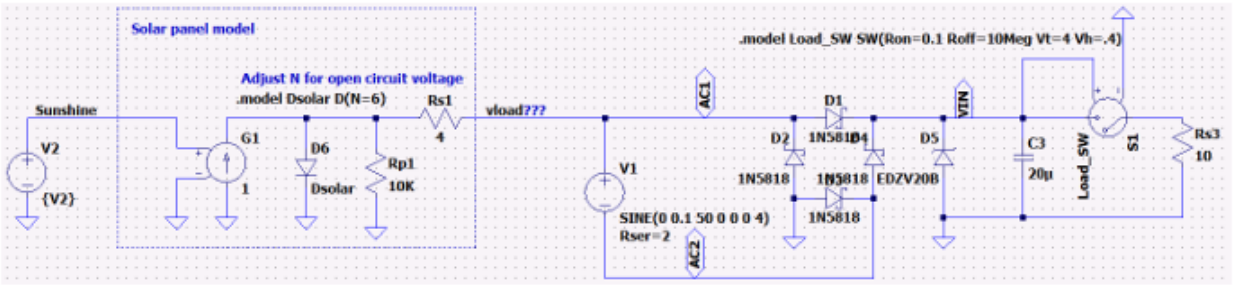


Figure 6. PV Joint PV and EMH operation simulation model

In figure 7, the results of a simulation using two sources at the AC1 and AC2 pins are displayed.

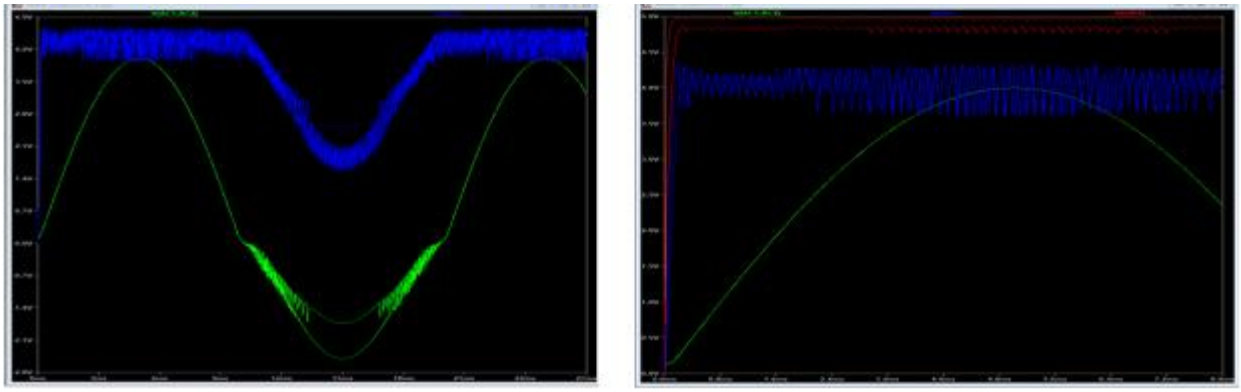


Figure 7. Multi-source simulation

4.2. Bench test and measurements

The experiment was repeated on the bench. The maximum efficiency displayed by the system was 82%. In Figure 8 the bench test is displayed.

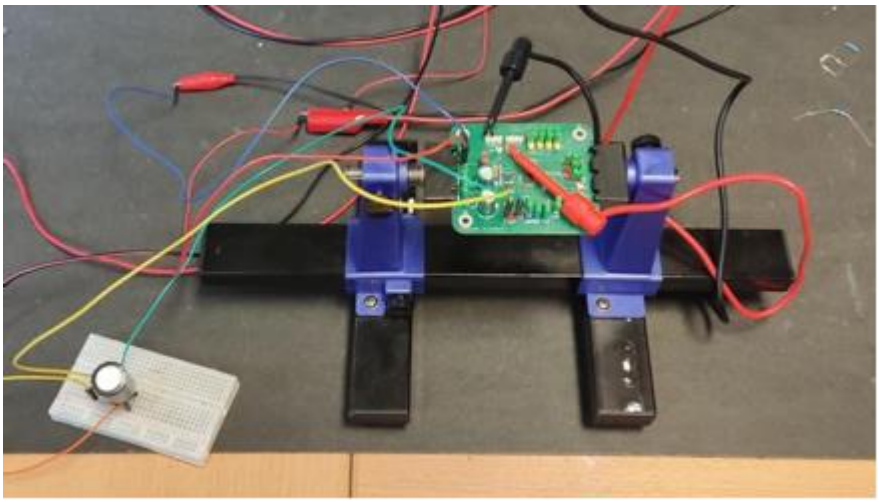


Figure 8. Experimental setup

5 Conclusion

5.1. Sum up

The autonomous multi-source energy harvesting system demonstrated effective performance in energy capture, management, and operational autonomy during bench tests. By integrating photovoltaic and electromagnetic harvesters, the system successfully powered itself while monitoring steel structures in naval environments. The results confirm the system's potential for long-term, self-sustaining operation in harsh conditions.

5.2. Implications for Future Research

Future research could focus on optimizing energy efficiency further and expanding the system's applicability to other structural health monitoring scenarios. Enhancing energy storage capacities and improving adaptability to diverse environmental conditions will enable broader deployment across industries requiring autonomous monitoring solutions.

References

[1] P. Pattakos, A. Katsoulas, S. Angelopoulos, A. Ktena, P. Tsarabaris, and E. Hristoforou, “Development of an Autonomous Magnetic Permeability Sensor,” IEEE Transactions on Magnetics, vol. 59, no. 2, pp. 1–4, Oct. 2023, doi: 10.1109/TMAG.2022.3205106.

[2] P. Pattakos, S. Angelopoulos, A. Katsoulas, A. Ktena, and E. Hristoforou, “Magnetic Harvester for an Autonomous Steel Health Monitoring System Based on Hall Effect Measurements,” Micromachines, vol. 14, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/mi14010028.

EMERGING TECH CONFERENCE 2024
Edge Intelligence

Papers

Session 1.4 | Innovative Approaches

Session Chairs: **Gregory DOUMENIS**

Session Presentation: **Evangelos HRISTOFOROU**

Charting A New Course for Food Logistics: Enhancing River and Ocean Transport Efficiency with CFD-Based Digital Twins, with examples from soybean logistics in the Amazon River

Sotiris Bantas, Miltiadis Xynopoulos, Alexandros Dimitriou, Antonis Sioutas, and Efstathios Kaloudis

Review of Novel Battery Balancing Topologies

Evi Keramida, Sotirios Athanasiou and Fotis Plessas

BoM Reduction for the 5G Radio Unit

Kostas Vryssas and Kostas Papathanasiou

Development and validation of the AcubeSAT nanosatellite communications module

Georgios Kikas, Christina Athanasiadou, Ilias Kamoisis, Ioannis Dimoulios and Alkis Hatzopoulos

Computer vision-based identification of motorcycle helmets using a TPU-based platform

Sotiris Michalakeas and Christoforos Kachris

HETiA
EXPANDING THE DIGITAL FRONTIERS

UNIVERSITY OF THESSALY FOUNDED 1984

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 98 – 103

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Charting A New Course for Food Logistics: Enhancing River and Ocean Transport Efficiency with CFD-Based Digital Twins, with examples from soybean logistics in the Amazon river

Sotiris Bantas[1], Miltiadis Xynopoulos[1], Alexandros Dimitriou[1], Antonis Sioutas[1], and Efstathios Kaloudis[2]

[1]*Centaur Analytics, Inc., 1923 Eastman Ave, Ventura, 93003, CA, USA*

[2]*Computer Simulation, Genomics and Data Analysis Laboratory, Department of Food Science and Nutrition, School of the Environment, University of the Aegean, Myrina, Lemnos, Greece*

sotiris@centaur.ag, stathiskaloudis@aegean.gr

Abstract

Efficient grain transportation via barge along major waterways like the Mississippi and Brazilian Amazon is vital for global food security and economic stability. This paper presents a Computational Fluid Dynamics (CFD)-based digital twin methodology tailored for crop logistics in barge transportation. Integrating sensor data with predictive analytics, the digital twin continuously monitors temperature, humidity, and airflow dynamics to optimize grain storage conditions during transit. Case studies from commercial grain logistics show how this approach enhances efficiency, reduces costs, and ensures grain quality, demonstrating the transformative potential of IoT and digital twin technologies in global food supply chains.

1 Introduction

The efficient transport of grain and oilseeds such as soybeans by barge along major waterways such as the Mississippi River and the Brazilian Amazon is critical to global food security and economic stability. However, ensuring optimal storage conditions during transport presents significant logistical challenges. Traditional methods of monitoring grain quality during transport often rely on manual inspections or static sensors that only provide periodic snapshots of environmental conditions (Caixeta Filho and Péra, 2018). While these approaches are functional, they lack the adaptability and real-time insights required to effectively manage dynamic, long-distance grain logistics.

In recent years, emerging technologies have sought to address these challenges. Previously reported Computational Fluid Dynamics (CFD) models (Lawrence & Maier, 2011; Quemada-Villagómez et al., 2020) have been used to simulate airflow and temperature conditions in static storage environments. However, these models typically utilize predefined data and lack the ability to adapt to real-time changes. Another notable development is the integration of Internet of Things (IoT) technology in logistics management, allowing for real-time data collection and remote monitoring of grain conditions (Dyck et al., 2023). While IoT has enabled better traceability and monitoring, existing solutions often fall short in their ability to predict future conditions and optimize storage environments

proactively.

This study presents a novel CFD-based digital twin methodology tailored to the unique challenges of grain barge logistics. Digital twin technology, which creates a virtual replica of a physical system, has been successfully applied in industries such as manufacturing and precision agriculture (Dyck et al., 2023), but its application in river and marine transport logistics remains relatively unexplored. By integrating real-time sensor data from IoT devices with advanced CFD models, our digital twin continuously monitors and predicts key environmental parameters such as temperature, humidity and airflow dynamics. This enables dynamic adjustment of grain storage conditions throughout the transport process, ensuring grain quality while reducing the risk of spoilage. As demonstrated by (Kollias et al. 2022), digital twin models can be instrumental in predicting and optimizing crop conditions in postharvest logistics, and our application of this methodology in barge transport marks a significant extension of its potential.

Several technologies are available in the same area, including static CFD models for airflow simulation (Lawrence & Maier, 2011; Quemada-Villagómez et al., 2020) and cable sensors for condition monitoring and weather stations that monitor the ambient conditions. However, these technologies typically operate independently and lack the predictive capabilities and real-time optimisation capabilities offered by the proposed digital twin approach.

In terms of market positioning, this digital twin solution occupies a unique space within the grain logistics industry. While traditional solutions focus on monitoring or simulation, the present approach provides a comprehensive system that not only monitors but also optimises grain transport conditions in real time (Kaloudis et al., 2024). This positions the proposed technology as an industry leader in predictive grain logistics management, offering significant cost savings, improved efficiency and enhanced sustainability.

2 Materials and Methods

2.1. Crop monitoring sensors and IoT system

To accurately monitor the storage conditions of grain, a network of IoT-enabled wireless sensors can be installed within the grain and move with it during transport (Figure 2). These sensors are designed to measure critical environmental parameters such as temperature, relative humidity, CO2 and O2 levels (Bantas et al., 2019). The high sampling frequency allows real-time data to be transmitted to a central database, enabling continuous monitoring (Centaur). The collected sensor data is integrated into advanced CFD models to simulate the airflow and heat transfer dynamics within the storage environments. These models solve the Navier-Stokes equations governing fluid flow and include additional equations for heat and mass transfer (Kaloudis et al., 2022). The simulation accurately represents the conditions within the storage systems, including potential hotspots for spoilage. Appropriate boundary conditions are imposed, including ambient temperature, relative humidity, solar radiation, wind speed and water temperature along the barge route. The CFD models are combined with other differential equations used to predict several critical quality metrics of the stored grain, including dry matter loss, the presence of visible mould and germination capacity. Validation of the CFD model is presented in Kollias et al. (2022).

For this study, a representative barge shipment of 3,000 tonnes of soybeans was analyzed, traveling from Miriituba to Bacarena, along the Amazon river. The barge dimensions were 8 meters in height,

30 meters in length, 12 meters in width, and a 5-meter draft. The voyage began on June 20, 2023, and concluded on July 4, 2023. Two wireless sensors were installed on the barge: one near the top layer of the soybeans at a height of 7 meters and the other near the bottom layer at a height of 2 meters.

2.2. Web-based IoT platform and predictive crop quality analytics

Additionally, a web platform (Centaur) enables users and grain managers to access real-time data and predictive insights from their smartphones. This platform displays the status of storage and logistics assets, including predictions of various quality metrics such as dry matter loss, the risk of visible mold, and the germination capacity of stored grain and oilseeds. The predictive power of the incorporated digital twin model, combined with the accessibility of the web platform, demonstrated significant potential to reduce spoilage and ensure the quality and safety of grain transport by barge through informed and timely decision-making.

The platform features a user interface (Figure 2) designed to provide comprehensive insights into the conditions within a grain barge, without requiring users to have advanced technical expertise. A visual representation of key metrics, such as a temperature map (showing product temperature) and a moisture map (indicating water content), allows for intuitive monitoring. Furthermore, the platform provides a detailed breakdown of maximum, average, and minimum values of product condition, for each day of the forecast period. Users can also simulate the entire barge journey, day by day, including potential demurrage periods. This predictive simulation uses the same model described below for real time analysis, projecting future conditions based on current data and trends.

A time-domain CFD simulation is being executed for each day of the journey, utilizing OpenFoam® CFD simulation software (OpenFoam) to model airflow and heat transfer dynamics. Boundary conditions for the simulation are defined by river water temperature and ambient weather conditions sourced from a weather API service, while internal conditions are derived from the real-time sensor readings. This integration of external environmental data and internal sensor metrics allows for highly accurate predictions of quality and storage conditions throughout the voyage.

Figure 1: An example of a wireless sensor installation on a barge. The sensor is connected to a gateway to transmit information in real time to the cloud platform and feed the digital twin model.

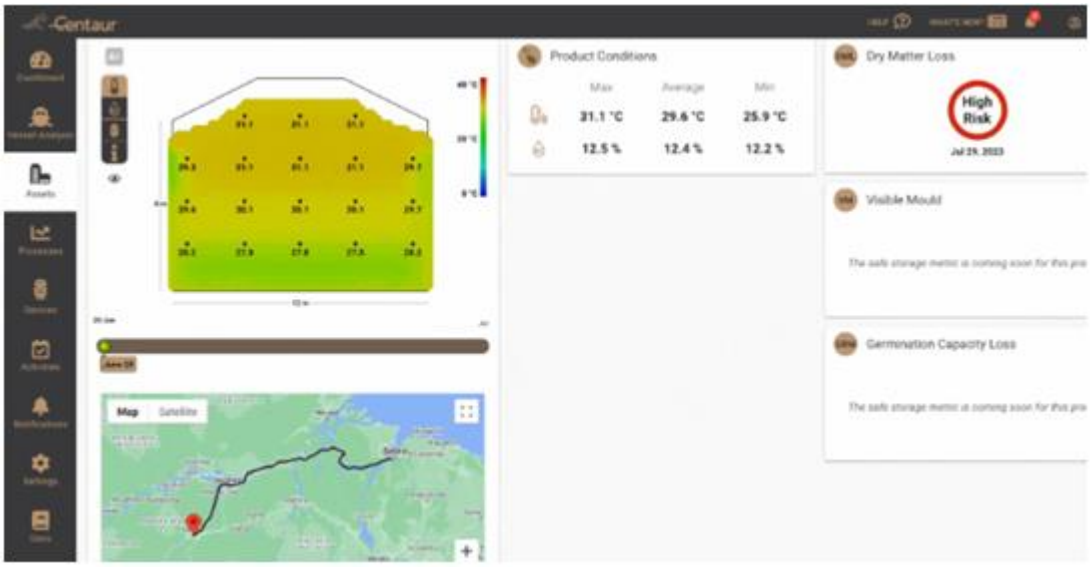


Figure 2: Barge logistics modeling UI

3 Results and Conclusions

Figure 3 shows the evolution of temperature and relative humidity as recorded by the two sensors installed on the barge. The monitoring period covers the entire duration of the barge's journey from Miritituba to Bacarena, along the Amazon river in Brazil. The temperature profile for both sensors shows relatively stable readings throughout the voyage, fluctuating between 28°C and 33°C. The data indicate that the temperatures of the top and bottom layers of the soybeans remained closely aligned, with minimal divergence. In particular, around 27 June 2023, there are fluctuations which correspond to the unloading process. The relative humidity readings from both sensors show consistent values between 65% and 75% throughout the voyage, indicating stable environmental conditions within the barge. Similar to the temperature data, a significant fluctuation is observed around 27 June 2023.

Regarding the number of sensors required for such applications, while increasing the number of sensors would undoubtedly improve the accuracy of the predictions by providing more granular data over the entire shipment, the validation of our model with experimental data has shown that for the volume of grain under consideration, two sensors provide a cost-effective yet reliable solution. These sensors are capable of measuring not only temperature and humidity, but also CO2 and O2 levels. Any abnormal conditions such as spoilage or localised hotspots occurring in areas remote from the sensor locations would likely be detected by changes in gas concentrations, as gases diffuse more efficiently throughout the grain mass than temperature or moisture content. Therefore, the combination of environmental and gas measurements ensures that even with only two sensors, the monitoring system remains robust and reliable for this specific application.

The integration of wireless sensors and CFD modeling provided a comprehensive understanding of the internal conditions of the soybeans in the barge (Figure 4). The CFD model identified specific zones within the barge where temperature and humidity levels were consistently higher. These conditions indicated a higher risk of dry matter losses, an important metric which is related to food loss during transportation and storage.

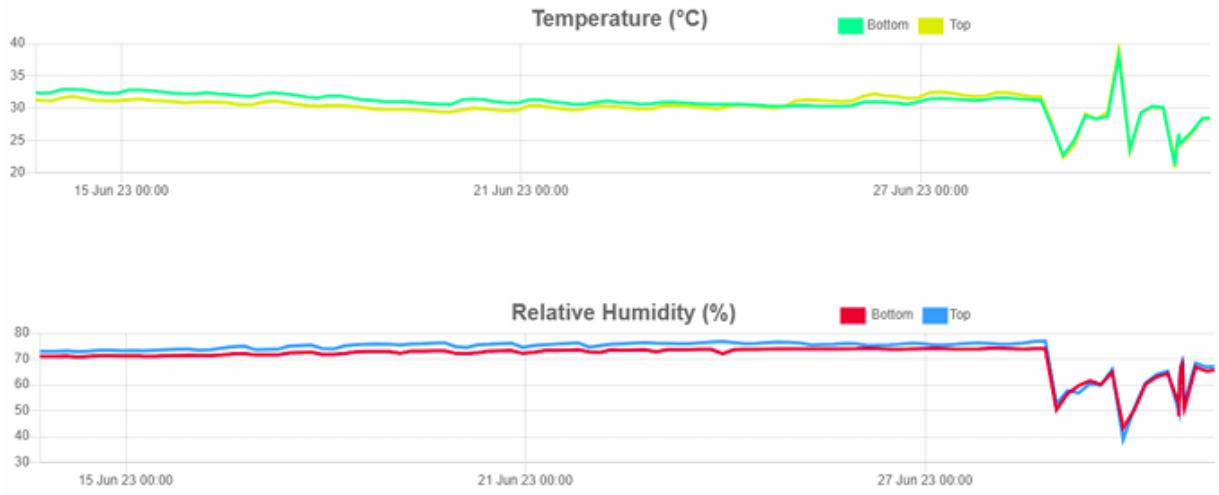


Figure 3: Temperature and relative humidity profiles of transported soybeans, recorded by two wireless sensors

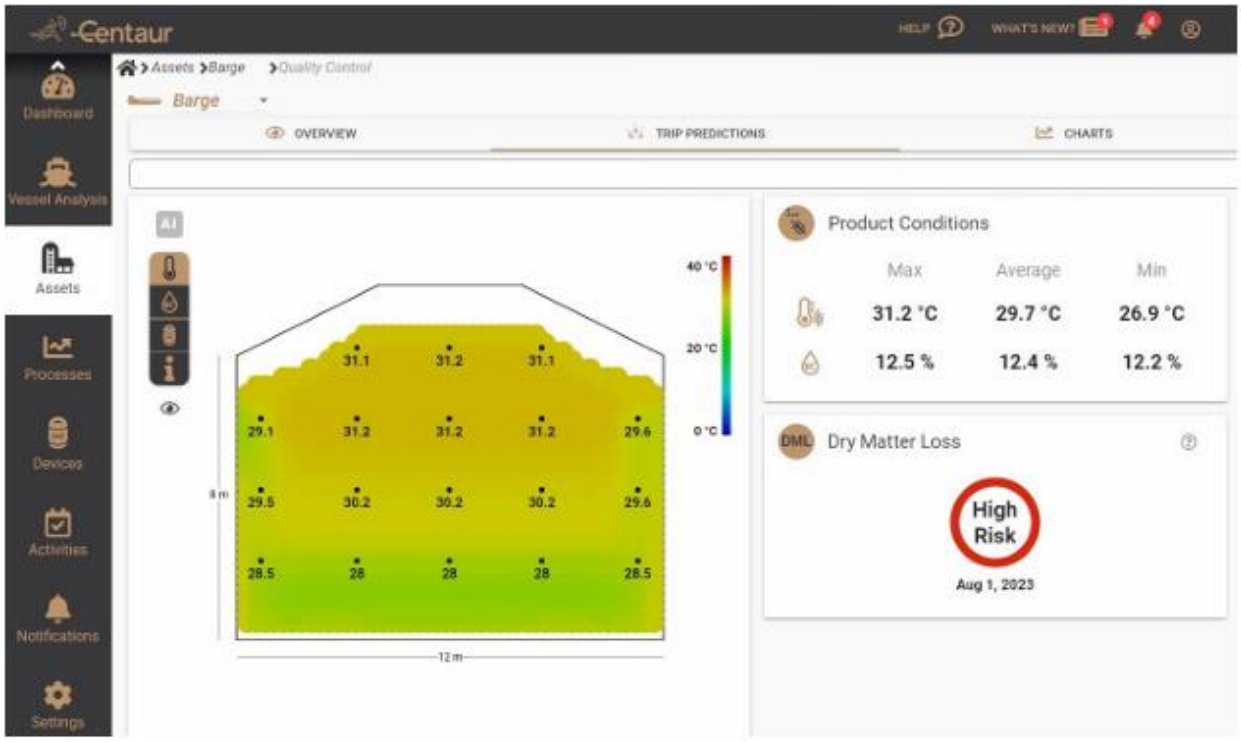


Figure 4: The web platform user interface showing soybean conditions at various points in the barge hold, including forecast values (left).

References

[1] Bantas S., Sotiroudas V., Ham E.H. (2019) Wireless sensor devices for post-harvest crop quality and pest management (US Patent No. 10,296,863-B2). US Patent and Trademark Office.

[2] Caixeta-Filho J.V and Péra T.G, Post-harvest losses during the transportation of grains from farms to aggregation points, International Journal of Logistics Economics and Globalisation 2018 7:3, 209 247.

[3] Centaur (n.d.), Internet-of-Crops® product description. https://centaur.ag. Dyck G., Hawley E., Hildebrand K., Paliwal J., Digital Twins: A novel traceability concept for post- harvest handling, Smart Agricultural Technology, Volume 3, 0223, 100079, https://doi.org/10.1016/j.atech.2022.100079.

[4] Kaloudis E., Bantas S., Lerios A. (2023). Predictive post-harvest stored commodity management methods (US Patent No. 11,605,047). US Patent and Trademark Office.

[5] Kaloudis E., Bantas S., Sotiroudas V., Ham R.E., (2024) Systems and methods for post-harvest crop quality management. US Patent 12,026,655.

[6] Kollias G., Patsianotakis C., Sioutas A. and Bantas S., Digital twin approach for post-harvest crop protection, Emerging Technology Conference – Edge Intelligence (ETCEI) 2022.

[7] Lawrence, J., & Maier, D. E. (2011). Three-dimensional airflow distribution in a maize silo with peaked, levelled and cored grain mass configurations. Biosystems engineering, 110(3), 321-329.

[8] OpenFoam® (n.d.), available at https://www.openfoam.com.

[9] Quemada-Villagómez, L.I., Molina-Herrera, F.I., Carrera-Rodríguez, M. et al. (2020) Numerical Study to Predict Temperature and Moisture Profiles in Unventilated Grain Silos at Prolonged Time Periods. Int J Thermophys **41**, 52

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 104 – 111

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Review of Novel Battery Balancing Topologies

Evi Keramida[1§§§] Sotirios Athanasiou[2****] and Fotis Plessas[1††††]

[1] *Department of Electrical and Computer Engineering, University of Thessaly, Volos, GR.*

[2] *Sunlight Group Energy Storage Systems, 14564 Athens, GR.*

pakerami@e-ce.uth.gr, s.athanasiou@sunlight.gr, fplessas@e-ce.uth.gr

Abstract

Imbalances in battery cells attributed either to the cell production process variability or the operating conditions of the battery, (Conway, 2020) result in the degradation of the overall performance of battery systems and can decrease their life span. Several factors such as the over-voltage, under voltage, overcharge and under charge of battery cells affect the capacity of battery cells. So, several balancing methods have been developed to preserve batteries from damaging and to expand the battery life. The present paper is a brief review about some of the state-of-the-art topologies for battery management systems (BMS). Also, a comparison table is presented. Comparison factors are the circuit design, the balancing speed, the complexity, the efficiency, the cost of production and the size of the presented topologies.

1 Introduction

Sustainable transportation and electric vehicles (EVs) have witnessed an exceptional growth in recent years. The environmental crisis along with the technological development has led to the necessity for more sustainable solutions in transportation. EVs use batteries as fuel and these batteries have to align with strict specifications in order to be environmentally friendly and energy efficient at the same time. Based on the application several types of batteries can be used in EVs such as Lead-acid, Nickel based, and Li-ion (Innovation, 2024). EVs batteries can be connected in series to increase voltage or/and in parallel to increase capacity.

Batteries' performance degrades over the years due to several factors such as temperature, over charging and over discharging, aging, mechanical stresses and etc. (Dinh, 2024). Battery management systems (BMSs) have been developed to monitor and to equalize the charge of battery cells in a battery pack protecting the users and the battery system from damages and increasing the battery life by maintaining the operational conditions in a reliable level (Daowd, 2011).

Two main balancing methods have been developed to overcome battery cells imbalances. The passive balancing method and the active balancing method. The present paper is focused on the

§§§ Writing – original draft preparation, writing – review, editing and visualization
**** Supervision and review
†††† Supervision and review

active cell balancing methods. In active cell balancing, active elements such as capacitors and/or inductors as well as converters are used to transfer energy from the higher energy cell(s) to the lower energy cell(s). The active cell balancing method includes five main subcategories which are: cell bypass, cell-to-cell, cell-to-pack, pack-to-cell and cell-to-pack-to-cell depending on how the energy is transferred among the battery cells of the pack. Many topologies can be found into the literature for the battery balancing methods (Conway, 2020), (Dinh, 2024), (Moghaddam B., 2019), (Pham, 2020), (Shang Z. C., 2019), (Vivert, 2023). Figure 1 depicts how the active balancing methods and their topologies are classified.

This paper discusses the most state-of-the-art topologies and how they provide progress on the overall battery performance, speed of charging, the cost and other factors that are important in a BMS. Also, a comparison table presents a summary between the characteristics of the most recent topologies based on factors like balancing speed, complexity and size of the design and production cost.

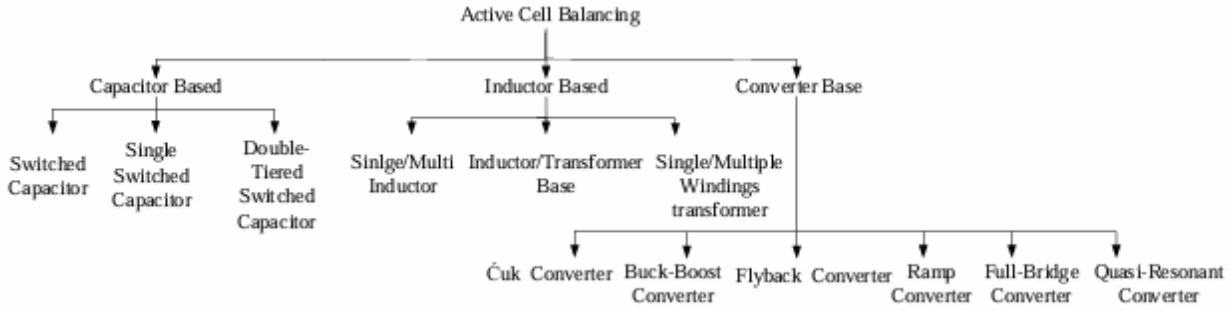


Figure 1: Classification of active cell balancing methods

2 Active Balancing Methods

The active balancing method uses elements such as capacitors, inductors or converters to transfer charge from cells with higher energy to the cells with lower energy. So, the cells in the battery pack become balanced to the same voltage level by delivering the excess charge from the cells with higher voltage levels to the lower voltage cells. The advantages of this method over the passive is that it balances the cells in shorter time and with higher efficiency levels but the topologies of the active method are more complex. Based on their transfer energy elements, active cell equalization topologies can be categorized in capacitive based, inductor or transformer based and converter based.

2.1. Capacitor Based Active Balancing Method

In this active equalization technique capacitors are used for charge transferring. The charge is stored in the capacitor from the cell with the higher voltage and transferred to the cell with the lower voltage. The capacitive based balancing topologies can be classified in the single switched capacitor, multiple switched capacitors and the double-tiered capacitor. An optimized version of the multiple switched capacitor topology is presented in (Shang Z. C., 2019). The topology is based on the Delta-Structured Switched Capacitor (DSSC) structure offering high balancing speed and high efficiency levels. The structure for every pair of two neighboring cells requires one capacitor and two series

connected transistors for each cell. Figure 2 presents the DSSC structure. Accordingly, another optimized topology based on the DSSC topology has been proposed in (Singirikonda, 2021). It is a Closed-Loop Switched Capacitor Structure aims to maintain the constant voltage across each cell. The voltage of each cell is being sensed and compared with a reference voltage and if the voltage of the cell is lower than the reference voltage then the switch is turned on. The procedure is repeated until the cell be balanced. It appears a more compact design, lower cost and very good balancing speed compared to other switched capacitor topologies (Pascual, 1997), (Kim, 2014), (Ye C. F., 2017).

The topology in (Baughman, 2008) which is presented in Figure 3 is based on the double tiered topology and proves how this structure is better in terms of equalization time, efficiency and balancing currents. Another topology that uses the double tiered switched capacitor method is realized in (Ye L. L., 2019) with improved balancing speed and reduced number of required switches making the structure suitable for industrial applications. The proposed topology is a combination of two tiers, as shown in Figure 4. Buck-boost and switched capacitor-based cell balancing techniques are combined.

2.2. Inductor Based Active Balancing Method

Inductor based topologies use inductors to transfer energy among the cells of the battery pack. There are two categories in this method, the single and the multi inductor. An inductor based topology is presented in (Moghaddam B. , 2018) tries to solve the problem of conventional inductor based balancing method that occurs when energy is transferred from the first cell to the last cell of the pack which causes delay in the balancing procedure. Also, in (N, 2020) the proposed structure operates well under charging, discharging and idle regime of battery and can easily be configured for even or odd cell battery packs. This structure is similar to (Moghaddam B., 2018) but it has faster equalization time and less number of components. Figure 5 shows the topology for four cells but it can also be extended to an odd number of cells.

Correspondingly, to single/multi inductor based topologies there are also single and multi transformer topologies. Many topologies have been proposed with single/multi winding transformers. (Moghaddam B., 2019) proposed a multi-winding transformer which compared to the common methods reduces the number of required secondaries and so the size and the cost of the design. In (Conway, 2020) the proposed topology is based on the single transformer per cell balancing method. The architecture provides effective voltage monitoring and cell balancing. The proposed equalizer utilizes four interface units, as show in Figure 6.

2.3. Energy Converters for Active Cell Balancing

Many different converter topologies are used for cell balancing. The main topologies are cuk converter, buck/boost, flyback, ramp, full-bridge and quasi-resonant converter. A cuk converter topology is proposed in (Moghaddam V. d., 2019). The proposed converter requires n switches and n+1 inductors for n cells which leads to less losses and to higher efficiency levels. A simplified circuit of the cuk converter with two cells is depicted in Figure 7.

In cell balancing systems the buck, the boost and the buck-boost converters are used widely. The main drawback of these converters is their high cost of production and their complex design. A buck-

boost converter topology for cell balancing has been proposed in (Liu, 2020) and appears high balancing speed, high efficiency levels, compact size and low cost. The buck-boost converter is presented in Figure 8. Also, in (Raeber, 2021) a high efficiency buck-boost converter is presented which is a combination of a non-isolated DC-DC converter and a low-speed switching matrix with balancing efficiencies over 90% and it is shown in Figure 9.

A flyback converter topology is proposed in (Moghaddam V. d., 2019) and is depicted in Figure 10. The structure provides better efficiency results over the conventional flyback converter as it consists of N secondaries for N battery cells and one switch reducing losses. The full-bridge converter is mainly used for the plug-in hybrid EVs and requires intelligent and complex control method to control the switches but it depicts high power rating. A cascaded full-bridge multilevel converter is presented in (Vivert, 2023) and shown in Figure 11. The proposed work gives the ability to the converter to reconfigure by inserting and removing cells during operation. The topology is a good choice for high power applications. Finally, quasi-resonant converter is a DC- DC converter which not require any complex algorithm to operate. It has high balancing efficiency levels but it also, has high cost of production. An optimized structure of the quasi-resonant converter is proposed in (Shang Z. C., 2015) where the balancing time is reduced. Only one balancing converter is shared by all cells reducing the size of the circuit and the production cost. The core of the proposed equalizer which is a quasi-resonant LC converter is depicted in Figure 12. Another high efficiency LLC resonant converter topology is proposed by (Pham, 2020). The charge is transferred from the highest voltage cell to the lowest voltage cell using a LLC resonant converter which controls the power delivery among the selected cells and a switch network that selects the cells that are need to be balanced. It has simple design, fast balancing time and achieves a maximum efficiency of 94.5%. Figure 13 shows the structure of the proposed LLC resonant converter without the switch network.

3 Comparison

Table 1 is a comparison table among all of the state-of-the-art topologies that has been mentioned and analyzed in the present paper. The comparison factors are the balancing speed, the complexity of the design, the efficiency, the cost and the size of each topology. The significance of each of these metrics varies according to the final application. For example, in automotive systems complexity and the cost of production is an important factor. On the contrary, in military systems or Energy Storage systems the cost is not an important factor as the efficiency of the final product. So, based on the application requirements and specifications there is a tradeoff between these metrics. Therefore, BMSs are a key factor in battery cell balancing as they monitor the safe operation of the battery pack to increase the performance and the cycle life of the battery cells.

Topology	Type of Balancer	Balancing speed	Complexity	Efficiency	Cost	Size
Shang Z. C., 2019 (*Fig.2*)	Capacitor Based	good	simple	excellent (94%)	low	small
Singirikonda, 2021	Capacitor Based	very good	simple	excellent (95%)	low	small
Baughman, 2008 (*Fig.3*)	Capacitor Based	-	-	-	medium	-
Ye L. L., 2019 (*Fig.4*)	Capacitor Based	excellent	simple	-	low	small
Moghaddam, 2018	Inductor Based	very good	medium	-	high	-
N, 2020 (*Fig.5*)	Inductor Based	very good	simple	-	-	-
Moghaddam, 2019	Inductor Based	good	medium	high	low	small
Conway, 2020 (*Fig.6*)	Inductor Based	good	simple	medium	-	-
Moghaddam, 2019 (*Fig.7*)	Converter	very good	-	medium	low	medium
Liu, 2020 (*Fig.8*)	Converter	excellent	-	excellent (96%)	medium	small
Raeber, 2021 (*Fig.9*)	Converter	good	simple	excellent (90%)	-	small
Moghaddam V. d., 2019 (*Fig.10*)	Converter	very good	simple	high	low	small
Vivert, 2023 (*Fig.11*)	Converter	medium	medium	medium	-	-
Shang Z. C., 2015 (*Fig.12*)	Converter	excellent	-	excellent (98%)	low	small
Pham, 2020 (*Fig.13*)	Converter	very good	simple	excellent (94.5%)	medium	-

Table 1: Comparison Table of Active Cell Balancing Topologies

4 Conclusions

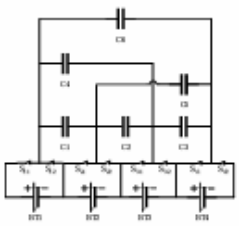


Figure 2: DSSC

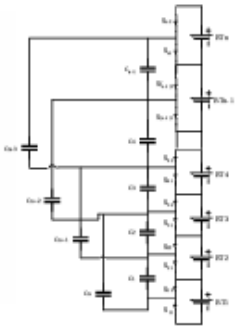


Figure 3: Double Tiered Switched Capacitor

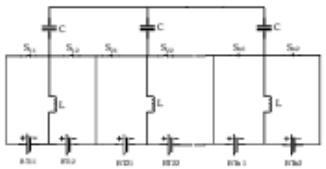


Figure 4: Double-Tiered Topology

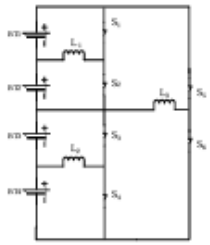


Figure 5: Inductor based Topology

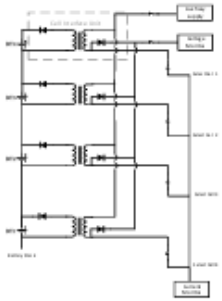


Figure 6: Single Transformer per cell Topology

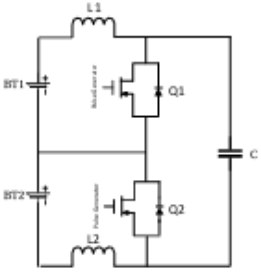


Figure 7: Simplified Cuk Converter Circuit with Two

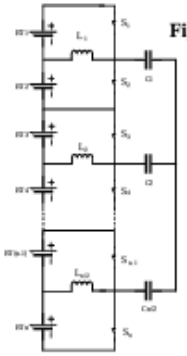


Figure 8: Buck-Boost Converter

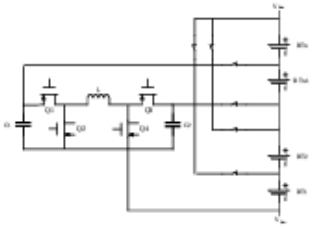


Figure 9: Buck-Boost Converter and Switch Matrix

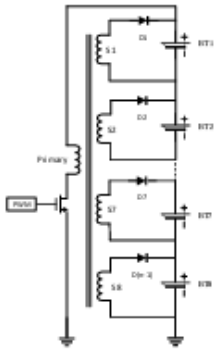


Figure 10: Flyback Converter

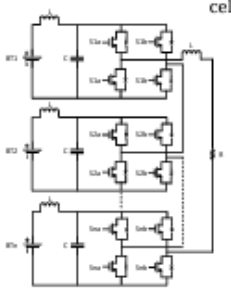


Figure 11: Cascaded Full-Bridge Converter

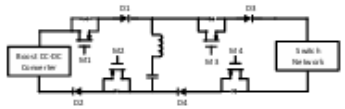


Figure 12: Quasi-Resonant LC Converter

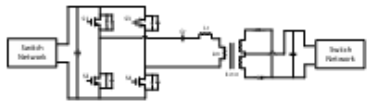


Figure 13: LLC Resonant Converter

The global demand for more environmentally friendly solutions and circular economy has increased the necessity for more efficient designs in BMSs. Future research in BMSs has to be focused on the optimization of the active cell balancing techniques to increase batteries' overall performance and to extend their life cycle. Various smart and efficient methods must be introduced to enhance batteries' efficiency levels according to different circumstances as BMSs can be found in many applications.

References

Baughman, F. (2008). Double-Tiered Switched-Capacitor Battery Charge Equalization Technique. IEEE Transactions on Industrial Electronics, 2277-2285.

Conway. (2020). An Isolated Active Balancing and Monitoring System for Lithium Ion Battery Stacks Utilizing a Single Transformer Per Cell. IEEE Transactions on Power Electronics, 3727-3734.

Daowd, O. B. (2011). Passive and active battery balancing comparison based on MATLAB simulation. IEEE Vehicle Power and Propulsion Conference (pp. 1-7). Chicago: IEEE.

Dinh, L. P. (2024). A Low-Cost and High-Efficiency Active Cell-Balancing Circuit for the Reuse of EV Batteries. Baterries. Retrieved from http://dl.acm.org/citation.cfm?id=2643085&dl=ACM&coll=DL Innovation, E. T. (2024, August 1).https://www.eurobat.org/campaigns-and-initiatives/battery innovation-roadmap-2030/. Retrieved from eurobat.org.

Kim, K. K. (2014). A Chain Structure of Switched Capacitor for Improved Cell Balancing Speed of Lithium-Ion Batteries. IEEE Transactions on Industrial Electronics, 3989 - 3999.

Liu, Y. T. (2020). Automotive Battery Equalizers Based on Joint Switched-Capacitor and Buck Boost Converters. IEEE Transactions on Vehicular Technology, 12716-12724.

Moghaddam, B. (2018). An Active Cell Equalization Technique for Lithium Ion Batteries Based on Inductor Balancing. 9th International Conference on Mechanical and Aerospace Engineering (ICMAE) (pp. 274-278). Budapest: IEEE.

Moghaddam, B. (2019). A Single Transformer for Active Cell Equalization Method of Lithium-Ion Batteries with Two Times Fewer Secondaries than Cells. Electronics.

Moghaddam, V. d. (2019). A Battery Equalization Technique Based on Cuk Converter Balancing for Lithium Ion Batteries. 8th International Conference on Modern Circuits and Systems Technologies (MOCAST).

Moghaddam, V. d. (2019). Flyback Converter Balancing Technique for Lithium Based Batteries. 8th International Conference on Modern Circuits and Systems Technologies (MOCAST) (pp. 1-4). Thessaloniki: IEEE

N, Y. C. (2020). Analysis and Implementation of Inductor Based Active Battery Cell Balancing Topology. IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES) (pp. 1-6). Jaipur: IEEE.

Pascual, K. (1997). Switched capacitor system for automatic series battery equalization. Applied Power Electronics Conference (pp. 848-854). Atlanta: IEEE.

Pham, D. C. (2020). High-efficiency active cell-to-cell balancing circuit for Lithium-Ion battery modules using LLC resonant converter. J. Power Electron, 1037–1046.

Raeber, H. A. (2021). Analysis of an Active Charge Balancing Method Based on a Single Nonisolated DC/DC Converter. IEEE Transactions on Industrial Electronics, 2257-2265.

Shang, Z. C. (2015). A Cell-to-Cell Battery Equalizer With Zero-Current Switching and Zero Voltage

Gap Based on Quasi-Resonant LC Converter and Boost Converter. IEEE Transactions on Power Electronics, 3731-3747.

Shang, Z. C. (2019). A Delta-Structured Switched-Capacitor Equalizer for Series-Connected Battery Strings. IEEE Transactions on Power Electronics, 452-461.

Singirikonda, O. (2021). Active cell voltage balancing of Electric vehicle batteries by using an optimized switched capacitor strategy. Journal of Energy Storage.

Vivert, D.,. (2023). Decentralized Control for the Cell Power Balancing of a Cascaded Full-Bridge Multilevel Converter. Energies.

Ye, C. F. (2017). Topology, Modeling, and Design of Switched-Capacitor-Based Cell Balancing Systems and Their Balancing Exploration. IEEE Transactions on Power Electronics, 4444-4454.

Ye, L. L. (2019). Double-Tiered Cell Balancing System With Switched-Capacitor and Switched Inductor. IEEE Access, 183356 - 183364.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 112 – 120

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

BoM Reduction for the 5G Radio Unit

Konstantinos Vryssas
Argo Semiconductors, Athens, Greece
kostantis@argosemi.com

Abstract

This paper explores strategies for reducing the bill of materials (BoM) in 5G massive MIMO radio units, with a focus on optimizing RF component integration while maintaining performance standards. We examine the architecture of Argo Semiconductors product series "AS0102" RF-sampling companion IC and analyze challenges related to electromagnetic interference (EMI) and crosstalk that arise from passive component coupling. Through comprehensive EM simulations, we identify key factors influencing TX-TX and RX-RX isolation, such as transformer placement and decoupling capacitor positioning. Our findings suggest that carefully managing floor planning and signal integrity can significantly improve isolation, enabling efficient BoM reduction without compromising on system performance. The insights from this study provide valuable guidelines for the design of future 5G radio units.

1 Introduction

Massive Multiple Input Multiple Output (m-MIMO) is a key enabler for the advancement of 5G cellular communications, providing higher capacity, improved coverage, and enhanced spectral efficiency. As the demand for 5G continues to expand, optimizing the infrastructure of m-MIMO systems, particularly in the sub-6GHz band, is crucial for reducing costs and facilitating widespread adoption.

One critical area of optimization is the Bill of Materials (BOM) for the radio unit, where reducing the number of RF components plays a pivotal role in packing a large number of radios into a single radio unit. Achieving this requires sophisticated integration techniques that condense multiple RF functions into fewer integrated circuit (IC) packages or even a smaller number of dies. However, this high level of integration introduces several challenges, such as noise coupling and increased heat dissipation, which can degrade system performance if not properly addressed.

In this paper, we present a novel circuit design that integrates multiple TX/RX paths on a single die. The design is based on advanced CMOS-SOI technology and incorporates various strategies to mitigate the challenges arising from dense integration, including the use of guard bands to reduce crosstalk and interference. The paper will also explore approaches to optimize component layout and improve isolation between transmission paths, enabling a reduction in the total number of key components required in m-MIMO radio units by more than 15%.

The proposed solution demonstrates significant BOM reduction while maintaining performance metrics such as power output and efficiency. The system's modular design, with both two-core and

four core configurations, highlights its scalability and adaptability to different deployment scenarios. By optimizing RF integration and system design, this approach contributes to more efficient and cost effective m-MIMO infrastructure for 5G networks.

2 Typical m-MIMO Architecture and Integration Challenges

In a typical massive MIMO (m-MIMO) radio unit, such as the 32T32R configuration, the radio architecture includes both digital and analog components, with most of the BOM originating from the Analog Front End (AFE). In fact, the AFE accounts for approximately 97% of the key components, which significantly impacts the overall complexity and cost of the radio unit.

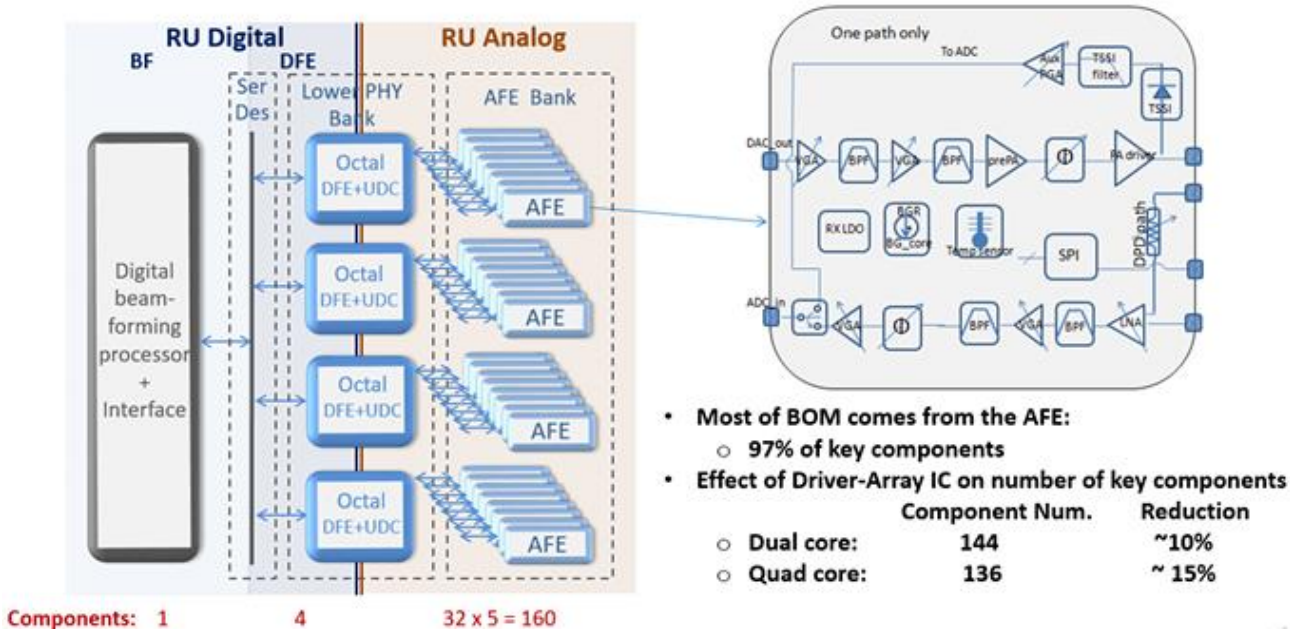


Figure 1: Typical Architecture of 32T32R massive MIMO Radio

One of the critical factors influencing the BOM is the Driver-Array IC architecture. As shown in our design analysis, using a dual-core or quad-core configuration for the driver array directly impacts the number of key components required for the m-MIMO system:

Component Configuration	Component Number	Reduction
Dual Core	144	~10%
Quad Core	136	~15%

Table 1: Effect of Driver-Array IC on number of key components

By integrating more cores into the design, significant reductions in the BOM are achieved, which is crucial for achieving a more compact and cost-effective radio unit.

3 Integration Challenges: EMI and Crosstalk

As RF components become more densely integrated into smaller IC packages, issues such as Electromagnetic Interference (EMI) and crosstalk become increasingly problematic. These phenomena are particularly prevalent in RF ICs due to the high density of interconnects and close proximity of sensitive components.

EMI can originate from several sources within the IC, including direct radiation from the surface of the IC and conducted noise through signaling ports. The impact of EMI can be severe, leading to synchronization issues, clock misalignment, and signal degradation. One of the most challenging aspects of EMI is the high-frequency noise signals conducted through the power lines, which can interfere with the critical timing and signaling of the system.

Crosstalk occurs when unwanted electromagnetic coupling between interconnects, bonding wires, power pins, or signal pins degrades the performance of the IC. This interference can manifest in several ways, such as signal propagation delays, failures, pulse distortions, and emissions. As IC geometries shrink and the distance between components reduces, the risk of crosstalk increases, posing a significant threat to signal integrity.

The effects of crosstalk and EMI are directly responsible for poor signal integrity and an increase in noise margins, ultimately degrading overall system performance. In the context of m-MIMO systems, where precision and timing are crucial, these issues can severely impact the reliability and efficiency of the radio unit. To address these challenges, several techniques can be applied to the layout and design of the IC: Noise and Signal Separation: Ensuring that noise-generating nets and sensitive signal nets are physically separated on the layout can help to reduce EMI and crosstalk. Careful layout optimization can minimize the parasitic coupling between various nets, reducing the likelihood of signal interference.

Additionally, the use of advanced electromagnetic (EM) modeling is essential to accurately predict and mitigate the risks associated with EMI and crosstalk. As IC designs move towards more advanced process nodes, the complexity of design rules, interconnect parasitics, and wire resistance continues to grow, requiring even more precise modeling and simulation techniques to ensure optimal performance.

4 Case Study: AS0102 Sub-6GHz AFE RF-Sampling Companion IC

The AS0102 Sub-6GHz AFE RF-Sampling Companion IC from Argosemi is a highly integrated solution designed for 5G radio systems, offering both transmit (TX) and receive (RX) paths. The architecture is optimized for sub-6GHz 5G radio units, and it provides high levels of performance and integration for a variety of m-MIMO radio configurations. It supports 4x4 or 2x2 radio configurations, making it ideal for high-throughput applications. Each core in the AS0102 integrates both the 5G-Tx and 5G-Rx paths. This compact design makes it well-suited for high-density integration into massive MIMO radio units, but it also introduces specific challenges in terms of signal integrity and noise management, as explored in the following sections. A critical requirement for this IC is achieving greater than 35 dB of Tx-Tx and Rx-Rx isolation, which is essential for maintaining optimal Error Vector Magnitude (EVM) and Signal-to-Noise Ratio (SNR) performance.

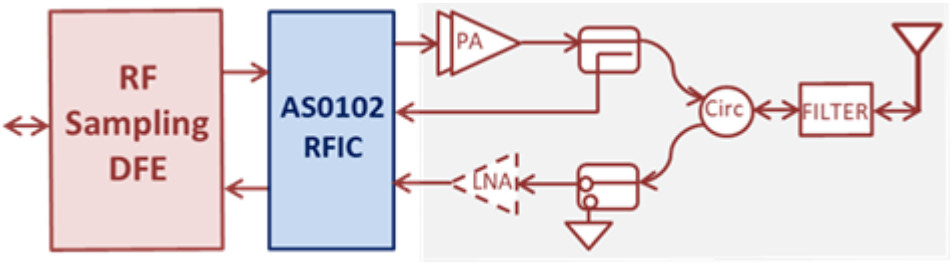


Figure 2: AS0102-Sub-6 GHz AFE RF-Sampling Companion IC

Below, we outline the key features of the AS0102's TX and RX paths.

AS0102 Features	TX	RX
Freq. Range	3.3-4.9 GHz	
Bandwidth	450 MHz	
Gain	25dB	
O1dBCP	28dBm	
Noise Figure	1.8dB	

Table 2: AS0102 Selected Key Metrics

5 Isolation Study

Isolation between signal paths is a critical factor in the design of 5G radio units, especially for m MIMO systems where multiple transmission and reception paths are in close proximity. Electromagnetic (EM) interference between these paths can degrade performance, and the isolation level must be accurately characterized and optimized. We conducted isolation simulations for both TX TX and RX-RX paths in the AS0102 architecture. The goal was to ensure high isolation between paths to minimize interference and cross-talk, thus preserving signal integrity.

5.1. TX Isolation: Identifying Worst Coupling between TX cores

To simulate the isolation between two TX paths, the following workflow was implemented: All passive components, such as PGA, prePA, and PA transformers, were moved to the top hierarchy level. This setup allowed easy switching between different passive models. A secondary TX path (TX2) was added to the top hierarchy, powered on, and sharing the same control logic as the primary TX path (TX1). However, no input signal was applied to the secondary path (TX2). Isolation is defined as the difference in output power (in dB) between TX1 and TX2. This provides a measure of how well TX1 and TX2 are isolated from each other. S-Parameter files generated through EMX were used for the isolation simulation, considering different separations and dimensions. The simulation yielded isolation values of approximately:

- 33.3 dB for 1000 μm separation
- 43 dB for 2500 μm separation
- 52.6 dB for 3000 μm separation

Note: AS0102 package models were included in the simulation but without TX-TX coupling. These models were only used for evaluating TX performance and output matching.

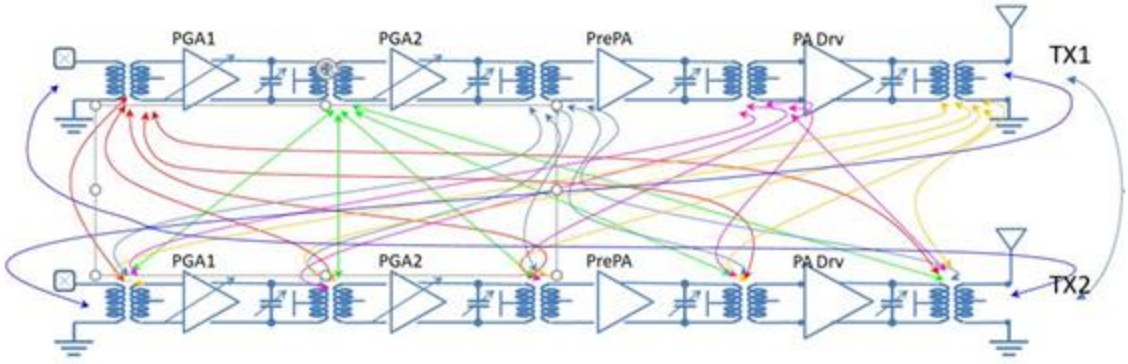


Figure 3: All Cross-talks combinations in 2x2 TX paths

In our design for the AS0102 architecture, understanding and mitigating electromagnetic interference (EMI) and crosstalk between TX paths is essential for maintaining signal integrity. The major contributors to coupling and interference are the passive components in the transmission chain, particularly transformers and inductors. Power Amplifier (PA) transformer act as major contributor. The isolation between TX1 and TX2 slightly improved when moving the "PAC=1" source from the PGA1 output to the PA output. However, it was clear that the PA transformer was the main source of coupling from TX1, primarily due to the significantly larger power it handles compared to other passives. This result was expected and confirmed by several test scenarios, demonstrating that the TX1 PA transformer contributes most to the power observed at the TX2 PA output.

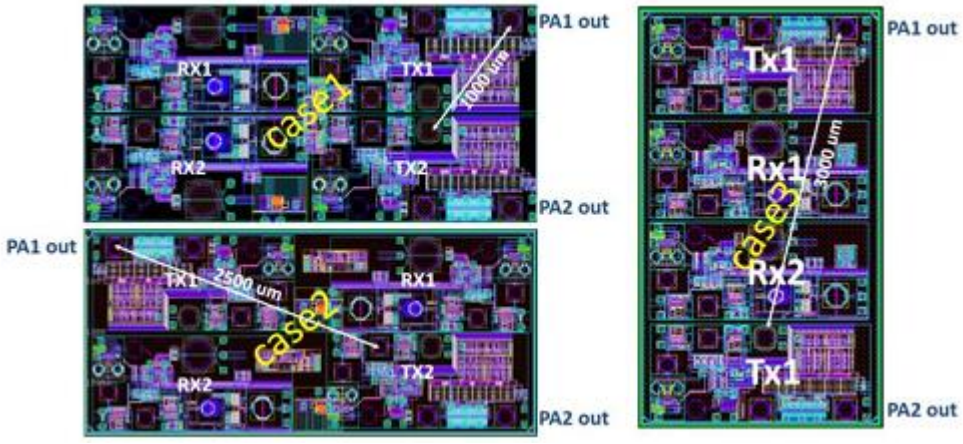


Figure 4: Three different floor plan cases of the 2x2 Transceiver identifying worst coupling combinations between passives

To isolate the worst coupling combinations, we employed a systematic approach: In each test, one passive component was treated as non-coupled by using its standalone S-Parameter model, which had no coupling with the other five passives. The remaining five passives were simulated with coupled S Parameter models. The major improvement in isolation occurred only when the TX1 PA transformer was treated as non-coupled, further emphasizing its significant role in the observed interference.

From the data of our study, it became evident that the major coupling occurs between the TX1 PA transformer and the TX2 PGA1 inductor. When non-coupled models for TX2 prePA and PA transformers were used, only a slight improvement in isolation was observed. However, significant improvement occurred when using a non-coupled TX2 PGA balun, confirming that this inductor-balun interaction is a primary source of interference. Improvements in the prePA transformer coupling were minor, with a maximum possible improvement of only about 3 dB. Thus, further efforts to reduce coupling in the prePA transformer would not yield substantial benefits.

We conducted the study for three different floor plan configurations of the 4x4 transceiver, with results aligning with the previous observations.

Case 1 was selected as the optimal configuration due to its superior isolation performance compared to the other two cases. A general rule of thumb emerged from the simulations: for every 300 μm increase in separation, we observed a 2-3 dB increase in isolation. This trend is consistent for typical inductors with diameters up to 200-250 μm. A 1000 μm pitch distance between the two transmitters was found to be insufficient, with the analysis suggesting that a separation of at least 1500 μm would provide safer isolation levels. The PA1 transformer and its associated bumps were identified as the main sources of interference, particularly affecting the prePA2 and PA2 transformers in the adjacent TX path. As a potential solution, a rearrangement of the AS0102 Tx5G output transformer bumps could be considered to increase separation and reduce coupling. Another solution to consider would be the use of an 8 shaped prePA or PGA transformer, which, while having a lower Q-factor, would help in mitigating crosstalk due to its reduced coupling footprint.

The study of TX-TX isolation revealed that the PA transformers are the main contributors to coupling and EMI, and that significant improvements in isolation can be achieved through floor plan optimization and passive component design. By carefully analyzing different coupling scenarios and floor plan arrangements, we identified key strategies for improving isolation in the AS0102 architecture. Going forward, rearranging the layout of the PA transformers and their bumps as well as exploring alternative transformer designs, such as the 8-shaped transformer, will be critical in minimizing EMI and ensuring optimal performance.

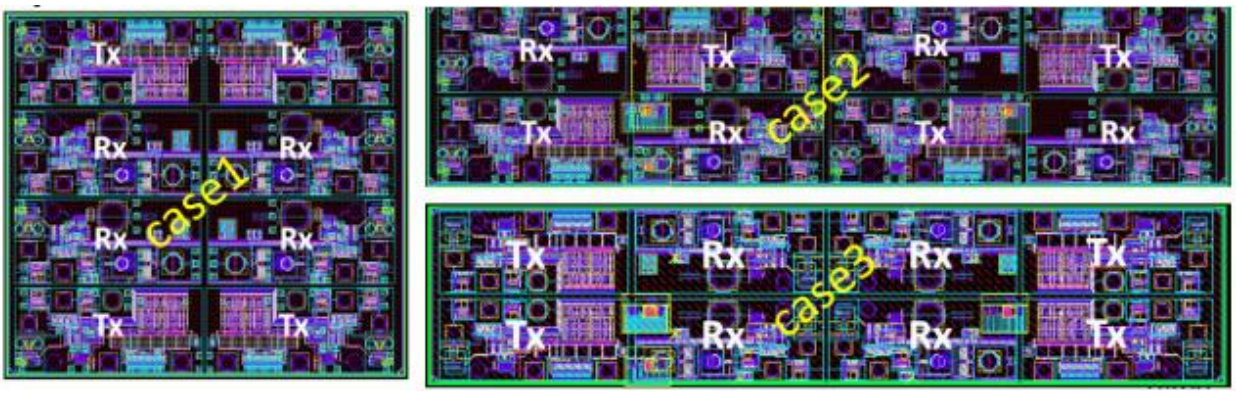


Figure 5: Three different floor plan cases of the 4x4 Transceiver identifying worst coupling combinations between passives

5.2. RX Isolation: Identifying Worst Coupling Combinations

The RX-RX isolation simulation was conducted in a similar manner, with modifications specific to the RX paths. The simulation testbench consisted of two complete RX paths, the aggressor and the victim:

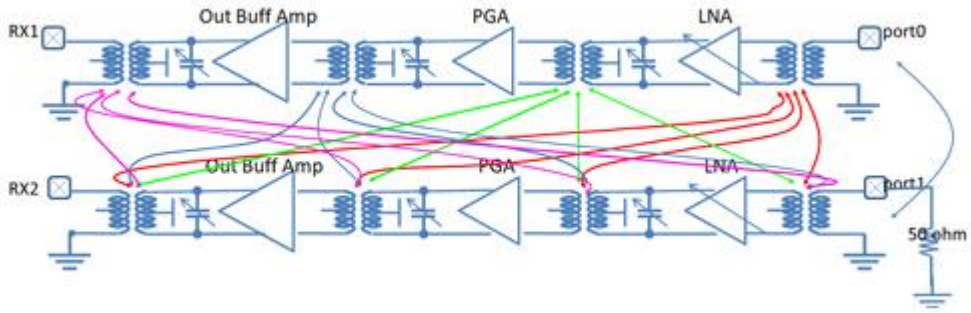


Figure 6: RX coupling combinations between passives on a 2x2

The total RX gain was set to approximately 52 dB, measured at the Output Buffer Amplifier. The input signal to RX core 0 was set at -50 dBm. All passive components (input/output transformers and routings) were moved to the top hierarchy, allowing the sharing of S-parameter models between RX core 0 (RX0) and RX core 1 (RX1). The RX path hierarchy was also modified to allow the connection of the S-parameter files to both RX paths, ensuring a complete and realistic simulation environment. Harmonic Balance simulations were used to assess the performance and isolation levels between the RX paths. Similar to the TX simulation, a secondary 5G-RX path (RX1) was added to the top hierarchy. RX1 shared the same control logic as RX0 but had no input signal. Instead, RX0 received an input signal, while RX1 was terminated with a 50 Ohm load at port 1. The RX-RX isolation was defined as the difference in output power (in dB) between RX0 and RX1, measured at the Output Buffer Amplifier. The simulation results demonstrated the importance of accurate S-parameter modeling and passive component placement in achieving high levels of isolation, particularly as IC designs become more compact and densely integrated.

Through detailed electromagnetic simulations we were able to achieve strong isolation results, demonstrating the effectiveness of the design and the importance of precise EM modeling ensuring that the AS0102 architecture provides the necessary isolation between signal paths and meet performance requirements. As with TX isolation, identifying the worst coupling combinations in the RX path is crucial for maintaining signal integrity, particularly for sensitive reception stages. Our methodology aims to evaluate the effectiveness of the current floor planning, analyze how isolation is affected by different passive components, and provide insights on optimizing the decoupling capacitor setup for better RX-RX isolation. To assess RX-RX isolation, we inject the RF signal from Port 0 and measure the signal level difference at the Buffer Amp outputs of RX Core 0 (aggressor) and RX Core 1 (victim). The isolation is determined by the output power difference between these two cores, with Port 1 terminated using a 50 Ohm to ground configuration. This setup provides a reliable measurement for isolation and enables us to observe how the two cores interact under different coupling scenarios. We utilize harmonic balance simulations to model the non-linear behavior of the circuit and to accurately capture the coupling effects between RX paths. The current floor planning with a 1000 μm spacing between RX cores is used as the baseline for these simulations. We aim to evaluate whether this spacing can achieve the target isolation of 35 dB. To predict the worst-case coupling scenarios, we applied a method similar to the one used for TX isolation. By sweeping the

capacitances on the VDD nets, we can observe potential resonances that might degrade isolation. This approach helps us identify problematic points where the impedance of the power nets (surrounding transformers or inductors) affects the isolation between RX cores.

Results from the capacitance sweep indicated that, as seen in the TX case, the impedance characteristics of the VDD nets surrounding the passive components play a significant role in inter-core isolation. Higher capacitance values generally improved RX-RX isolation, as they help to suppress the resonance effects that may arise at specific frequencies.

The value of the decoupling capacitors has a direct impact on RX-RX isolation. In general, larger capacitance values lead to better isolation, as they improve filtering and reduce noise on the power and ground planes. However, in addition to the capacitance value, the position of the decoupling capacitors also plays a critical role. To study this effect, we conducted simulations where the decoupling cap bank was connected between various VDD and VSS ports, positioned at different spots along the power and ground paths. Results showed that the RX-RX isolation is highly dependent on the exact points of connection between the VDD-VSS rails and the decoupling capacitors.

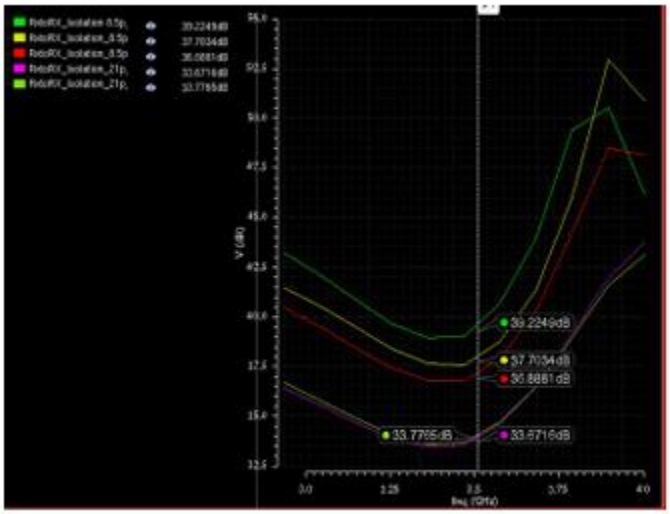


Figure 7: RX resonance vs decoupling capacitance

The optimal placement of these capacitors significantly reduces coupling between RX cores, improving isolation by filtering out the unwanted noise and resonance effects that can degrade signal performance. The results from the RX-RX isolation study showed reasonably consistent behavior when compared to the TX-TX isolation findings. Similar trends were observed, particularly regarding the influence of transformer placement and power net impedance on isolation performance.

As seen in the TX isolation study, spacing between components plays a crucial role, and increasing the spacing beyond 1000 μm is likely to improve isolation significantly.

The study highlights the importance of both decoupling capacitor value and positioning in achieving optimal RX-RX isolation. Increasing the value of the decoupling capacitors and carefully selecting their placement on the power and ground planes can lead to significant improvements in isolation between RX cores. Based on the harmonic balance simulations and capacitance sweep results, we recommend increasing the spacing between RX cores to above 1500 μm where feasible, and optimizing the layout of decoupling capacitors to minimize noise and resonance effects.

6 Conclusion

The development of 5G massive MIMO radio units, particularly in the context of reducing the bill of materials (BoM) while maintaining performance integrity, presents several key challenges related to component integration, electromagnetic interference (EMI), and signal isolation. Through a detailed exploration of both TX and RX paths, this paper has highlighted the critical impact of passive component coupling, floor planning, and decoupling strategies on overall system performance.

The findings show that as RF component densities increase, especially with advanced RF-sampling companion ICs like the AS0102, effective management of EMI and crosstalk is essential for ensuring system reliability and signal quality. Our EMX simulations have demonstrated the importance of passive component separation and optimized decoupling capacitor placement in improving TX-TX and RX-RX isolation, with notable improvements achieved by increasing separation distances and strategically positioning decoupling capacitors.

Moreover, the study reinforces that while component miniaturization and integration are key to BoM reduction, careful attention must be paid to the physical layout of passive elements and power distribution networks to mitigate the risks of EMI, crosstalk, and other forms of interference that can degrade signal integrity.

In conclusion, achieving optimal performance in 5G m-MIMO radio units requires a delicate balance between BoM reduction and the mitigation of signal integrity risks. By applying advanced simulation techniques and focusing on the careful integration of RF components, future designs can continue to push the boundaries of performance while meeting the stringent demands of next-generation wireless communication systems. The ongoing evolution of process technologies, such as the AS0102, offers promising paths for further innovation, but they must be accompanied by equally advanced strategies for addressing the growing complexity of RF system integration.

References

[1] EMX + Calibre design flow document for GF 28nm, Integrated software incorporation

[2] EMX Virtuoso interface tutorial, March 2015, Integrated software incorporation

[3] Introduction to EMX Planar 3D Solver, Rapid Adoption Kit (RAK), Product Version: Virtuoso ICADVM20.1 / EMX 6.0 / Spectre 19.1 May 2021

[4] Modeling of Integrated RF Passive Devices, Sharad Kapur and David E. Long Integrand Software, Inc. Berkeley Heights, NJ 07922 USA, http://www.integrandsoftware.com

[5] EMX: Overcoming Silicon Chip EM Simulation Challenges for Passive Circuit Analysis and Model Development, John M. Dunn; Sharad Kapur; David Long, 2021 International Applied Computational Electromagnetics Society Symposium (ACES) August 2021

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 121 – 126

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Development and validation of the AcubeSAT nanosatellite communications module

Georgios Kikas[‡‡‡‡]*, Christina Athanasiadou, Ilias Kamoisis, Ioannis Dimoulios, and Prof. Alkis Hatzopoulos
Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

AcubeSAT is an open-source nanosatellite project which will study the effects of micro gravity and radiation on microorganism cells. Its communication needs are served by the SatNOGS COMMS Board, a board initially designed by Libre Space Foundation (LSF) and modified accordingly by SpaceDot for its mission requirements. In this paper, we provide an overview of the functionality of the board, the in-house software and FPGA development as well as a detailed presentation of the board's testing and validation procedures.

1 Introduction

The AcubeSAT nanosatellite is designed by SpaceDot, a non profit, interdisciplinary student research team in the Aristotle University of Thessaloniki (AUTh), to serve an open-source biological mission, under the auspices of the "Fly Your Satellite!" 3 programme of the ESA Education Office. The scientific output of the mission consists of images of microorganisms, captured by the on-board microscope and camera setup, in order to study their response to radiation and microgravity. The downlink of the scientific data, even though they will be compressed, sets strict requirements for the Communications Subsystem, due to the limited communication windows and power restrictions. In this paper, the design and testing of the Communications board will be presented.

2 Subsystem Description

The novel biological experiment and the environmental conditions in Low Earth Orbit (LEO), have posed several challenging requirements to the definition of the telecommunications sub system, which all reflect on the design and testing approaches to be discussed. The AcubeSAT communications subsystem consists of two in-house designed antennas, a UHF deployable turn stile antenna and an S-Band patch antenna both designed in-house [7] and the COMMS Board, which this paper is going to expand upon.

The AcubeSAT COMMS Board is an in-house modified version of the SatNOGS COMMS hardware, designed by Libre Space Foundation [13], featuring in-house software and hardware IP cores. It is designed around an STM32H743 microcontroller unit that coordinates the sensors, RF circuitry, power supply unit and inter-board communication, a AC7Z020 SoM featuring an XC7Z020 FPGA SoC that is responsible for the generation of the modulated baseband signals along with their encoding

‡‡‡‡ Corresponding author: kikasgeorg@ece.auth.gr

and an AT86RF215 RF transceiver module that up-converts the FPGA modulated samples. The hardware differences with the original SatNOGS COMMS are mainly related to a better compatibility with the AcubeSAT mission in terms of power consumption, RF circuitry and mechanical properties (design changes in connector placement and the EMI shield design) [18].

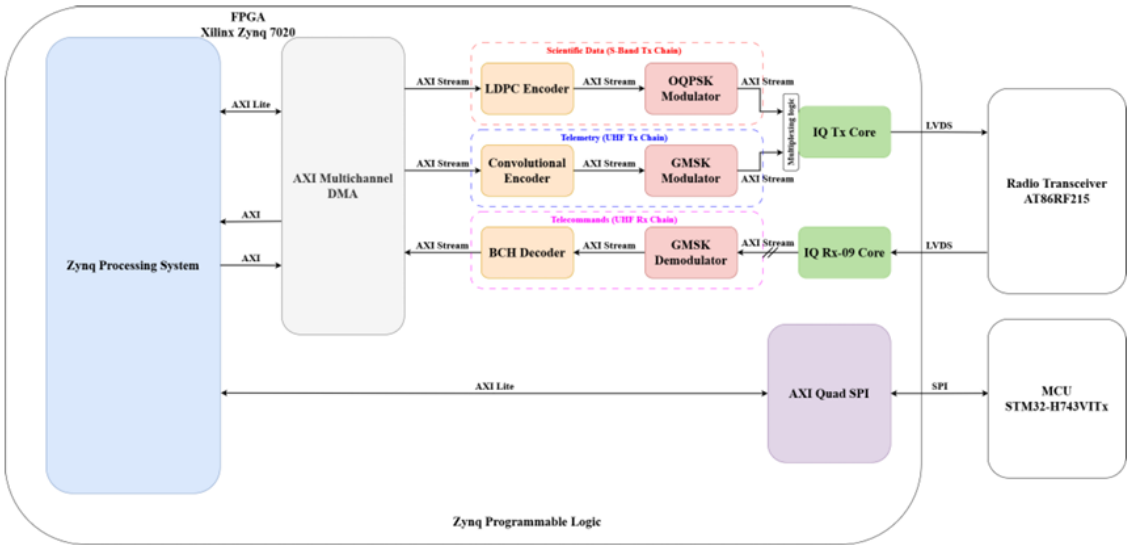


Figure 1: FPGA Physical Layer System Architecture

The in-house software design consists of the following main components: the MCU firmware [17], built around the STM32 HAL and the FreeRTOS kernel [1] for real-time task scheduling, the CCSDS data link layer module [14], implementing the CCSDS TM and TC Space Data Link standards [9, 8] and the physical layer module [16], hosting the error correction, modulation and synchronization algorithms. The specifications of the codes, modulations and data packet formats used are thoroughly described in [6]. The integration and testing philosophy of all on-board software is described in [11].

Furthermore, regarding the physical layer, the main task the team has taken is to trans fer the majority of the algorithms to the on-board FPGA to increase performance and data throughput. A significant hindrance in implementing the entirety of the physical layer on FPGA is its significant power consumption, which comes in contrast to the limited power generation capabilities of a 3U LEO CubeSat. For this reason, the power footprint of each IP core is the primary metric used for optimization during design. In addition, apart from power simulations and testing in ambient conditions, the power footprint shall be closely monitored in flight rep resentative conditions as described in the following section. All in all, the block diagram of the in-house FPGA IP Cores is shown in Figure 1.

At the time of writing, three of the IP Cores have been fully developed and tested [15], including an OQPSK modulator, an LDPC encoder [12] and a BCH Decoder. It shall be noted that the configuration that was environmentally validated in the testing campaign included solely the OQPSK Tx chain. This choice is deemed acceptable as during an orbital pass, one Tx channel will be used at a time (UHF with GMSK or S-band with OQPSK), and S-band transmission has the highest power footprint on the board, thus simulating the worst scenario possible.

3 Environmental Testing and Validation

An environmental testing campaign is the means of testing the readiness of the subsystem to be sent to flight as a part of the whole satellite system and be space-qualified. These tests follow the philosophy “Test as you fly, fly as you test” [19], consisting of vibration sequences that mimic the launch mechanical stresses, followed by multiple Thermal Vacuum Chamber (TVAC) cycles, to test against the thermal cycles the CubeSat is going to endure in orbit. Being part of the FYS!3 programme enabled the team to utilize the CubeSat Support Facility at the ESA Education Office, ESEC-Galaxia, Transinne, Belgium, to conduct the aforementioned tests. The environmental testing campaign of AcubeSAT’s Communications Board was conducted in July 2024.

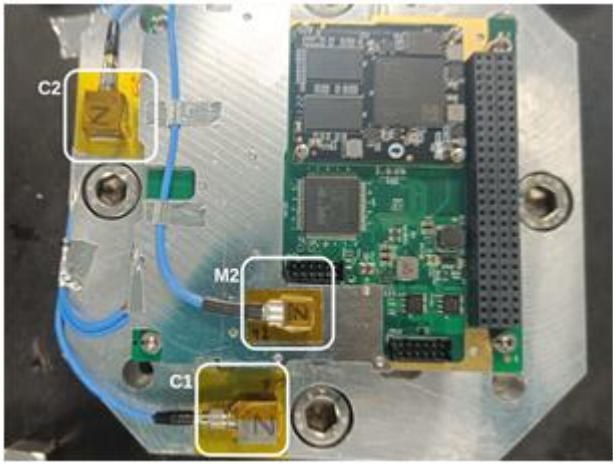


Figure 2: Accelerometers placed on the COMMS Board (VIBE Test)

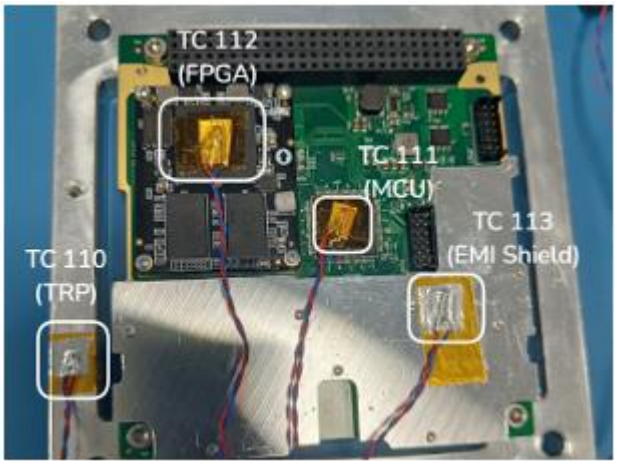


Figure 3: Thermocouples placed on the COMMS Board (TVAC Test)

The vibration testing sequence consisted of the following per axis: modal searches, to identify the resonant modes of the board, a random vibration sequence to simulate the launch mechanical stresses (adhering to the GEVS profile [2]) and a second post-stress modal search to detect any system mode deviation from the previous search. Any significant changes to the modes may propose permanent mechanical alteration of the system, and are included in the pass/fail criteria of the test. This sequence is conducted for all axis X, Y and Z and the test data are correlated with FEM simulations conducted beforehand [5]. The three axes sequences were successfully completed, apart from a minor non-conformance that was issued after the posttest visual inspection. Notably, 2 Surface-mount Technology (SMT) standoffs, on which the FPGA daughter board was mounted, were slightly detached from the main board due to the mechanical stress induced upon them. Eventually, this had no impact on the functionality of the board, as no electrical interface was damaged, verified by the post-vibration full functional test and the TVAC tests that followed. Finally, the vibration test was considered successful, with the minor non-conformance formally issued to ESA with a proposed rework for the flight model that ensures the stability of the mechanical interface.

For the second phase, after installing the thermocouple sensors, as seen in Figure 3, to monitor the temperatures of the board, a total of four thermal cycles were run, including 1 non-operational and 3 operational cycles as per [10], after defining the minimum and maximum operational target temperatures of the chamber at-25 ◦C and 42 ◦C respectively [4]. These were chosen in order to mitigate the overheating of the FPGA, which was noticed during an initial thermal balance test to have

an increased temperature gradient on power-on even in ambient conditions. To further accommodate for this issue, a more time-restricted testing approach, representative of the operational time of the downlink chain in orbit, was employed. Specifically, 8 minutes were chosen to account for a best-case pass and 4 minutes for a nominal pass, as also shown in [3]. A collective summary of all the thermal cycles can be seen in Figure 4, where the thermocouple measurements over time are plotted.

The functional testing approach during TVAC consists of two functional modes. These are the reduced functionality mode tests throughout the cycling, simulating the ”Idle” state on which the communications subsystem will be when there is no line-of-sight contact with the ground station, and the full functionality mode tests that are conducted at the cold and hot temperature plateaus, to simulate the ”Tx/Rx” state during a ground station communication window. The plateaus refer to the constant temperature dwell sequences in the maximum and minimum operational temperatures after each thermal half-cycle. The full functional tests are conducted there in order to test the performance of the system under the most unfavorable conditions. In Figure 4 the plateaus for the first non-operational cycle are evident. For the next three operational cycles the dwells also showcase temperature spikes in certain thermocouples (e.g. TC-112 mounted on the FPGA), due to heating during the full functional test.

In terms of functionality, no issues were identified during the hot plateaus and the board performance was well within operational requirements. During the cold plateaus, some issues emerged initially, which the team promptly analyzed. The primary issue was that high inrush current caused the over-voltage protection circuits to be triggered in colder temperatures, thus disabling the powering on of the board. The issue was overcome by altering the power-on sequence of the board components via firmware and the rework was tested rigorously to verify the reliability of the power supply circuitry during colder temperatures. Therefore, after the rework, the following thermal cycles progressed nominally and the Communications Board performance was validated throughout them.

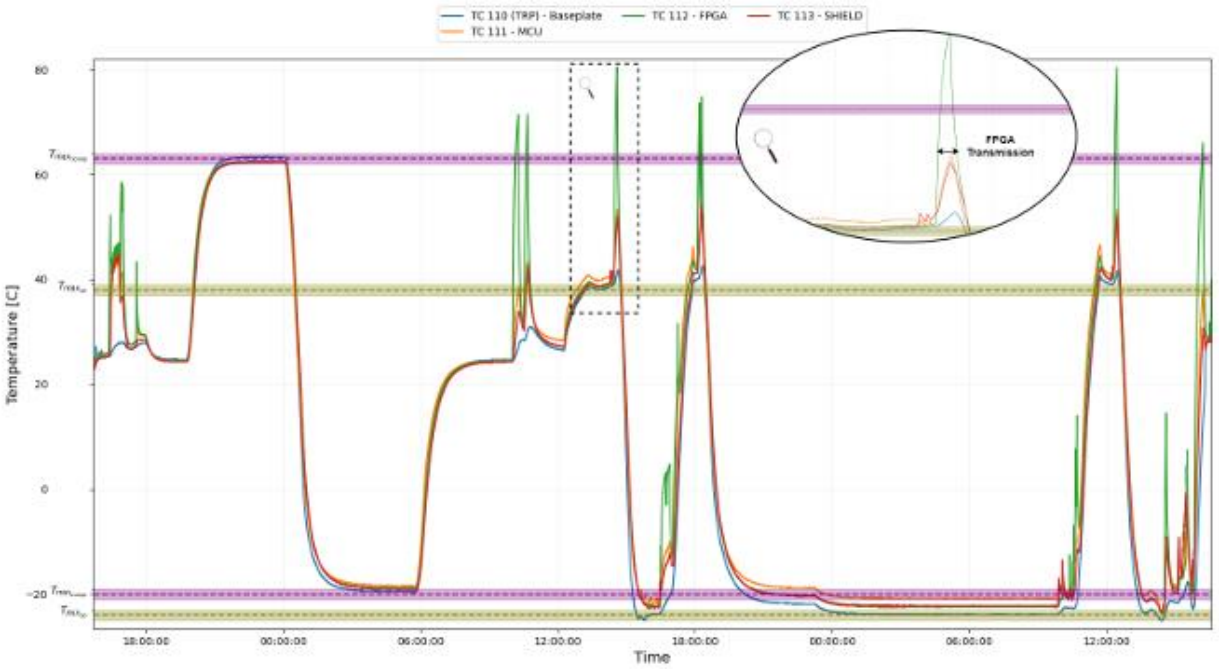


Figure 4: TVAC thermocouple data plotted over time

4 Conclusions

In conclusion, the environmental test of the AcubeSAT communications module was successful, and the subsystem is now considered qualified in thermal cycle and vibration testing, achieving a TRL-7 (Technology Readiness Level) status. Some future directions currently include continuing the FPGA IP Core development to include more modulation and encoding schemes and integrating the communication module with the rest of the system on a FlatSat level, before the CubeSat assembly, verifying the implementation of the data layer and operational procedures architecture on a system level. Finally, the environmental qualification of the hardware, described in this paper, and the ongoing software integration verification, establish the reliability of this mission-critical component in space conditions.

References

[1] FreeRTOS. https://www.freertos.org/.

[2] General Environmental Verification Standard (GEVS) for GSFC Flight Programs and Projects.

[3] AcubeSAT Team. Baseline Mission Analysis Appendix.

[4] AcubeSAT Team. COMMS EQM Thermal Vacuum Test Specification & Test Procedures.

[5] AcubeSAT Team. COMMS EQM Vibration Test Specification & Test Procedures.

[6] AcubeSAT Team. TTC Design Definition & Justification File. https://gitlab.com/acubesat/documentation/cdr-public/-/blob/master/DDJF/DDJF_TTC.pdf?ref_type=heads.

[7] P. Bountzioukas, G. Kikas, C. Tsiolakis, D. Stoupis, et al. The Evolution from Design to Verification of the Antenna System and Mechanisms in the AcubeSAT mission. http://dx.doi.org/10.48550/arXiv.2310.16134, 2023.

[8] The Consultative Committee for Space Data Systems. TC Space Data Link Protocol.

[9] The Consultative Committee for Space Data Systems. TM Space Data Link Protocol.

[10] European Cooperation for Space Standardization. Testing. https://ecss.nl/standard/ecss-e-st-1003c-testing/.

[11] A. Keremidis, S. Tzelepis, and A. Hatzopoulos. The Integration and Testing Procedures for the AcubeSAT Nanosatellite's Software. In International Conference on Synthesis, Modeling, Analysis and Simulation Methods, and Applications to Circuit Design (SMACD) 2024 Proceedings, 2024.

[12] Christos Sidiras and AcubeSAT Team. tory. AcubeSAT LDPC Encoder GitLab Reposi https://gitlab.com/acubesat/comms/fpga-development/ldpc-encoder-core/-/tree/development?ref_type=heads.

[13] Manolis Surligas, Agis Zisimatos, Ilias Daradimos, Aris Nikas, Dimitris Moustroufis, Pierros Papadeas, Manthos Papamathaiou, Alfredos-Panagiotis Damkalis, Vasillis Tsiligiannis, and Victoria Malyshkina. SatNOGS-COMMS: Turnkey Nanosatellite Communications.

[14] AcubeSAT Team. AcubeSAT CCSDS Space Data Link Protocols GitLab Repository. https://gitlab.com/acubesat/comms/software/ccsds-telemetry-packets.

[15] AcubeSAT Team. AcubeSAT FPGA GitLab Repository. https://gitlab.com/acubesat/comms/fpga-development.

[16] AcubeSAT Team. AcubeSAT Physical Layer GitLab Repository. https://gitlab.com/acubesat/comms/software/physical_layer.

[17] AcubeSAT Team. AcubeSAT COMMS EQM Software GitLab Repository. https://gitlab.com/acubesat/comms/software/comms-eqm-software/-/tree/campaign?ref_type=heads, 2024.

[18] AcubeSAT Team. AcubeSAT COMMS Hardware GitLab Repository. https://gitlab.com/acubesat/comms/hardware/satnogs-comms-hardware, 2024.

[19] Julia White. Test Like You Fly: Assessment and Implementation Process. AEROSPACE RE PORT, (8591), 2010.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, Pages 127 – 132

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Computer vision-based identification of motorcycle helmets using a TPU-based platform

Sotiris Michalakeas and Christoforos Kachris
University of West Attica Department of Electrical and Electronics Engineering Athens, Greece

Abstract

The use of helmets can significantly reduce fatal injuries in motorcycle accidents. How ever, there are still many instances where helmets are not used. In this paper, we present a computer-vision system for the automatic identification of helmet use by motorbike riders, that can be used to encourage the use of helmets. The system is developed on a Google Coral Dev Board with an embedded EdgeTPU, connected to a display and a buzzer that are used to provide feedback to the people it detects. Firstly, we present the training of the specific model with several optimizations and then we evaluate two versions of the SSD-MobileNetV2 model in terms of performance, accuracy and energy consumption. The proposed system can achieve high accuracy, low-latency and it consumes less than 4.2 Watts on a TPU board.

1 Introduction

According to the latest report from the European Transport Safety Council, respectively 15.5% and 2.9% of all road fatalities in the EU in 2018 were motorcyclists and moped riders. In Greece, from the 621 fatal accidents that were recorded in 2023, 211 of them were motorbike riders, while the overall reduction of the fatal accidents compared to 2022 was just 5% [1].

According to the data, helmets can significantly reduce the amount of fatalities among motorcycle riders in road accidents. In this paper we develop a low-cost, low-latency embedded system that automatically detects helmet use by motorbike riders using computer vision and provides feedback to promote helmet usage.

The system consists of a camera that feeds images in real time to the Google Coral Dev Board, which runs an object detection model and displays its results in a monitor. At the same time it sends logic signals for each class to an external microcontroller. The microcontroller processes these signals and controls an LCD and a passive piezo buzzer to provide both visual and auditory feedback to the motorbike riders.

The main contributions of this paper are the followings:

- Acomputer-vision system based on a low-cost Google Coral TPU that utilizes the advan tages of the TPU accelerators.
- Some optimization options for the training data and the training process to increase the accuracy of the helmet identifications.

- Integration of the computer vision system with a display and an audio device for the encouragement of helmet use when it is not used and positive feedback for when it is used.
- Performance evaluation of the proposed system using the SSD-MobileNetV2 and the SSD MobileNetV2-FPNlite architectures in terms of accuracy, inference throughput, and power consumption.

2 Related work

In recent years, the application of computer vision for helmet recognition has gained significant traction, particularly in enhancing workplace safety across various industries. Several studies have explored different methodologies to effectively identify helmet usage among workers, contributing to safer working environments.

In [3], it is presented a survey of several efforts for the automatic detection of helmets using computer vision algorithms.

One notable approach is presented by Wu and Zhao [4], who developed an intelligent vision based system that integrates pedestrian detection with helmet identification. Their method employs image recognition technologies to ensure compliance with safety protocols in construction sites, demonstrating the potential of computer vision to reduce accident rates significantly.

Another significant contribution comes from a study that proposed a dual-function system for helmet and identity recognition [5]. This research highlights the integration of helmet detection with worker identification, which is crucial for monitoring compliance in real-time. The authors conducted extensive testing under varying visual conditions, indicating that such systems can effectively enhance safety management in construction environments.

Moreover, advancements in deep learning have led to the development of robust helmet detection models that can operate in real-time. For instance, a study utilizing convolutional neural networks achieved high accuracy in detecting safety helmets across different scenarios, showcasing the effectiveness of these technologies in practical applications [2].

These advancements illustrate the critical role of computer vision in promoting safety through effective helmet recognition, paving the way for further research and development in this vital area. However, in these cases the identification of the helmet is restricted to in dustrial environments with low processing requirements. In the case of helmet detection for motorcycle riders, the system needs to use hardware accelerators to achieve low latency and high throughput as the objects are moving in high speed.

3 Training of the Helmet Detector

The training of the helmet detector was initially performed using a dataset from Kaggle that consists of 774 images of people with and without helmets. After a careful analysis of the dataset, it was found that many images (249) were duplicates and were removed, making the next trained model very inaccurate. To improve it, we added 500 more images and relabeled the entire dataset with two classes, ”With Helmet” and ”Without Helmet” using labelImg. These images were carefully sourced from the internet and contained objects in different scenarios, such as varying lighting levels, distances and backgrounds. The new dataset contained a total of 1015 annotated images in Pascal VOC format, but some of them were removed, because of the lower data quality that they provided. For example,

cases where people weren't facing the camera, or were wearing a headscarf or a burqa. The final dataset would contain 958 images. During preprocessing, a python script would randomly split the full dataset into a training subset (80%), a validation subset (10%) and a testing subset (10%). The total amount of images is quite small to fully train an object detection model from scratch. However, we would be using a method called transfer learning, in which a pre-trained model is re-trained with a new custom dataset. If done right, this method can achieve great results with much fewer data and training time.

For the training of the models we used the TensorFlow framework and the Tensorflow Object Detection API, running on Google Colab, so that we can utilize a CUDA enabled GPU to massively accelerate the training process. We split the dataset into training, validation and testing data. The xml files, that contain the annotation data for each image, were processed and converted into csv files for each subset. These files would later be converted to TFRecords to make them compatible with TensorFlow. The pre-trained models (SSD-MobileNetV2 and SSD-MobileNetV2-FPNlite) were imported from the Tensorflow 2 Detection Model Zoo.

The training of the models required around 38,000-40,000 steps, as after that, the total loss metric was not decreasing any more. The batch size was set to 16 because of memory limitations, the warmup learning rate was set to 0.0266 and the base learning rate to 0.08. These hyperparameters were applied to both of the architectures. After the completion of the training process, we converted the TensorFlow models to TFlite and we quantized the models to make them compatible with the Edge TPU board.

4 Integration and Inference on Coral TPU board

The Coral TPU Development Board uses an iMX SoC from NXP (with a Quad-core Arm Cortex-A53, plus a Cortex-M4F) integrated with a TPU engine specialized for edge applications. The board we used has 4GB LPDDR4. Two folders containing each model were compressed (zip files) and transferred to the Google Coral board via a microSD card. With the camera and the monitor connected to the board, the next step would be to set up the resolution and the minimum confidence threshold. The inputs of the models were 510x510 for the SSD MobileNetV2 and 320x320 for the SSD-MobileNetV2-fpnlite, so the resolution of the camera would be set close to these resolutions (640x480) to avoid any further distortion of the images. The framerate would be set to 30 frames per second, as it was the maximum the camera could achieve. After some testing, the optimal minimum confidence threshold was found at 30%.

Figure 1: The detection results from a camera test, in the monitor

After some changes to the inference python script, two logic signals that contain the de tection status of each of the two classes would be transferred via two GPIO (General-Purpose Input/Output) pins to the external microcontroller. If any object is detected as class 2 (With out Helmet) then the negative visual and auditory feedback would be played. Positive feedback will only be played if every detected object is classified as class 1 (With Helmet).

Figure 2: Back and Front panel of the Integrated device using the Coral TPU board, LCD screen and the display

5 Performance evaluation

To evaluate the models' performance we used two different sets of images. The first one was the testing subset from the custom dataset which was used to get a first estimate of its accuracy. Due to the fact that this set was randomly created each time a new model was trained, we couldn't use these results to directly compare them with each other. So we gathered 11 more images, containing 23 instances of each class (46 total), with varying distances, lighting levels and environments (backgrounds) to compare the two final models.

SSD-MobileNetV2 (45 FPS)			
TARGET / OUTPUT	With Helmet	Without Helmet	SUM
With Helmet	12 35.29%	2 5.88%	14 85.71% 14.29%
Without Helmet	4 11.76%	16 47.06%	20 80.00% 20.00%
SUM	16 75.00% 25.00%	18 88.89% 11.11%	28 / 34 82.35% 17.65%

Figure 3: Confusion matrix for the SSD-MobileNetV2

SSD-MobileNetV2-FPNlite (10 FPS)			
TARGET / OUTPUT	With Helmet	Without Helmet	SUM
With Helmet	10 31.25%	0 0.00%	10 100.00% 0.00%
Without Helmet	7 21.88%	15 46.88%	22 68.18% 31.82%
SUM	17 58.82% 41.18%	15 100.00% 0.00%	25 / 32 78.13% 21.88%

Figure 4: Confusion matrix for the SSD-MobileNetV2-FPNlite

Fig. 1 and Fig. 2 show the confusion matrices of the two models using the SSD-MobileNetV2 and the SSD-MobileNetV2-FPNlite architectures. As it is shown in the figures, the SSD MobileNetV2 offers slightly better overall accuracy and a faster inference time, but the SSD MobileNetV2-FPNlite is better in terms of identifying the riders that do not wear helmets.

The following equations show the Mean Average Precision of the SSD-MobileNetv2 and SSD-MobileNetV2-FPNlite.

SSD-MobileNetV2: $mAP = \frac{1}{n}\sum_1^n \frac{TP}{TP+FP} = \frac{1}{2}\cdot\left(\frac{12}{12+2} + \frac{16}{16+4}\right) = 0{,}8286$

SSD-MobileNetV2-FPNlite: $mAP = \frac{1}{n}\sum_1^n \frac{TP}{TP+FP} = \frac{1}{2}\cdot\left(\frac{10}{10+0} + \frac{15}{15+7}\right) = 0{,}8409$

Based on the testing of the models, the SSD-MobileNetV2-FPNlite seems to work very well on good lighting and on relatively small distances (up to 7m). This model is quite conservative in the identification of the helmets and it tends to classify more objects as ”Without Helmet”. Due to its complex architecture, that embeds an FPN in the SSD-MobileNetV2 model, many operations could not be transferred to the TPU, resulting in slower inference times. Specifically, the SSD-MobileNetV2-FPNlite achieves up to 11 fps, meaning that it could be ideal for cases when the objects (motorcycles) are moving no more than 30km/h.

Due to its more simple architecture (nearly all of its operations are handled by the TPU), the SSD-MobileNetV2 can achieve up to 110 frames per second with its default input size of 300x300x3. Since the camera’s framerate cannot surpass 30 fps, it would be a good idea to increase the input size to get better detection accuracy. The reduction in performance doesn’t affect the system, as long as the inference time doesn’t drop below 33 ms. After increasing the input size to 510x510x3, the model achieved 40-45 fps and was better at identifying objects in short distances (up to 5 meters). Its low inference time makes it a great choice, when the system is installed in roads where the motorcycles can be moving in higher speeds.

Generally, both models had great accuracy in scenarios where the lighting was good and the distance of the object didn't exceed 5 meters. They can work well in small roads or parking lot entrances/exits, but will struggle in bigger roads and cannot perform at all during nighttime.

The device is portable and needs to be self powered (i.e. using batteries powered by solar power panels). Therefore the power consumption of the system needs to be low. To this end, we have measured the power consumption of the TPU board running the 2 models for computer vision.

Table 1 shows the power consumption of the TPU board (using a 5V power supply) for the 2 different object detection models. As it is shown, the SSD-MobileNetV2 consumes approxi mately 3,74% less power compared to the FPNlite model.

Model	**Power (W)**
SSD-MobileNetV2	3.876
SSD-MobileNetV2-FPNlite	4.021

Table 1: Comparison of power consumption for the 2 computer vision models

6 Conclusions

In this paper, we show an embedded system for the identification of helmet usage by motorbike riders, using computer vision, in a low-cost TPU board that can achieve the strict requirements in terms of high accuracy and low-latency. The performance evaluation shows that the TPU board consumes low energy, making it ideal for self-powered systems (e.g. using a solar panel). The models can perform very well in well lit areas with objects that are not further than 5-7 meters and can achieve low enough inference times to detect in real time. We have high hopes that a system like this can increase helmet usage across the country, by providing feedback to the motorbike riders to increase awareness and decrease the number of deaths or severe injuries.

Acknowledgments: The registration fees were totally covered by the University of West Attica.

References

[1] EU. European road safety observatory facts and figures– motorcyclists and moped riders- 2023, 2023.

[2] Yange Li, Wei Han, Han Zheng, Huang Jianling, and Wang Weidong. Deep learning-based safety helmet detection in engineering management based on convolutional neural networks. Advances in Civil Engineering, 2020:1–10, 2020.

[3] Arya K M and Ajith K K. A review on deep learning based helmet detection. Proceedings of the International Conference on Systems, Energy Environment (ICSEE) 2021, 2020.

[4] Hao Wu and Jinsong Zhao. An intelligent vision-based approach for helmet identification for work safety. Safety Science, 109:123–130, 2018.

[5] Wei Zhang et al. Worker's helmet recognition and identity recognition based on computer vision. Scientific Research Publishing, 12:1–10, 2024.

EMERGING TECH CONFERENCE 2024
Edge Intelligence

Papers

Session 1.5 | Circuits/Systems & CAD

Session Chairs: **Nestor EVMORFOPOULOS & Christos SOTIRIOU**

Design and Electromagnetic Analysis of Interconnects in Silicon Interposer

Nefeli Metallidou, Thorsten Baumheinrich and Vasilis Pavlidis

Continuous and Symmetric Drain Current Compact Model for Nanoscale Triple-Gate FinFETs

Militadis K. Nakos, Andreas Tsormpatzoglou, Theodoros A Oproglidis, Dimitrios H. Tassis and Charalabos A. Dimitriadis

A low-rank balanced truncation approach for large-scale RLCk model order reduction based on extended Krylov subspace and a frequency-aware convergence criterion

Christos Giamouzis, Dimitrios Garyfallou, Nestor Evmorfopoulos and Georgios Stamoulis

An Efficient Design Methodology for RF and mmWave Voltage – Controlled Oscillators

Alexandros Chatzis, Anastasios Michailidis, Vasiliki Gogolou and Thomas Noulis

Automated Low-Noise Amplifier Design Methodology for Wireless Sensor Nodes

Panagiota Tsimpou, Anastasios Michailidis, Vasiliki Gogolou and Thomas Noulis

HETiA
EXPANDING THE DIGITAL FRONTIERS

UNIVERSITY OF THESSALY
FOUNDED 1984

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 133 – 138

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Design and Electromagnetic Analysis of Interconnects in Silicon Interposer

Nefeli Metallidou [1,2], Thorsten Baumheinrich[2] and Vasilis F. Pavlidis[1]
[1] *Electrical and Computer Engineering Department, Aristotle University of Thessaloniki*
[2] *Ansys*
nefeli.metallidou@ansys.com, thorsten.baumheinrich@ansys.com, vpavlid@ece.auth.gr

Abstract

In the post-Moore era, novel technologies are developed to support higher speeds, with significant emphasis in the field of advanced packaging. One of the technologies in this domain is 2.5D integration, where one or more chiplets are interconnected in the same substrate, called interposer. This integration technology allows for the chiplets to be placed in the same package, and thus the interconnects are much shorter than if the chiplets are placed in separate packages, resulting in lower interconnect latency. In this paper, signal integrity issues for a typical interposer that connects two dies are investigated. Three different types of interconnects are investigated. The metrics used to compare signal integrity among the three configurations are the reflection coefficient, insertion loss, coupling coefficient, and near- and far-end crosstalk between two differential interconnect pairs.

1 Introduction

Recently, the scaling of transistors has slowed down, due to thermal, technological, and economical limits. However, modern applications require continuously larger computing power and speed, emphasizing the need for novel integration techniques. One of these technologies lies in the area of advanced packaging and is called 2.5D integration, briefly described in the following section.

2 2.5D Integration

In traditional packaging, a semiconductor die is placed in a single package and the packages of different integrated circuits are connected on the Printed Circuit Board (PCB). The delay and power increase because the length of the traces that connect two different ICs on a PCB are in the millimeter scale. 2.5D integration is a more advanced technique, in which two or more semiconductor dies can be juxtaposed in the same package [6] - [10]. The dies are placed on top of a special substrate called an interposer, on which the connections between the dies are fabricated. An interposer is called passive if no active devices are contained in the substrate.

The connections on the interposer have a much shorter length, which introduces a smaller delay than in traditional 2D integration and, also, a lower power consumption. 2.5D integration has the advantage of reduced congestion when routing signals on the PCB. Furthermore, it enables heterogeneous integration, which means that dies of different technologies can be integrated in the same package, thus integrating e.g., high speed digital logic with analog circuits and power or

telecommunication electronics.

3 Interposer topologies

A passive silicon interposer has been designed using the GPDK45 library provided by Cadence, as well as the Virtuoso software. The PDK provides 11 available metals, of which the top four are used for signal routing. On the top metal (M11), rectangular 40 μm X 40 μm pads are placed for each chiplet, so that the chiplets can be attached to the pads using a flip-chip technology. For each chiplet, 16 pads are used in a 4x4 configuration. The pitch between the pads is 40 μm. The pitch between the corresponding pads of the two different chiplets is 3.2 mm. The inner column of each pad array is used for signal routing, using two differential pairs of signals. The other 12 pads are used as ground pads. The pads and the connectivity for die1 can be seen in Fig. 1. The ground plane configuration and the traces connecting the chiplets differ in the three configurations that are investigated in the following sections.

The operating data rate of the circuit is assumed to be at 2 Gbps, which corresponds to a 1 GHz baseband frequency, plus the first harmonic taken into account, thus reaching a 2 GHz bandwidth. The nominal operating voltage is at 1 V with each single-ended signal fluctuating from -0.5 V to 0.5 V. The maximum acceptable voltage drop for the signals is 10% of the nominal Vdd, such that the high level of the signals must be at least 0.4 V [4].

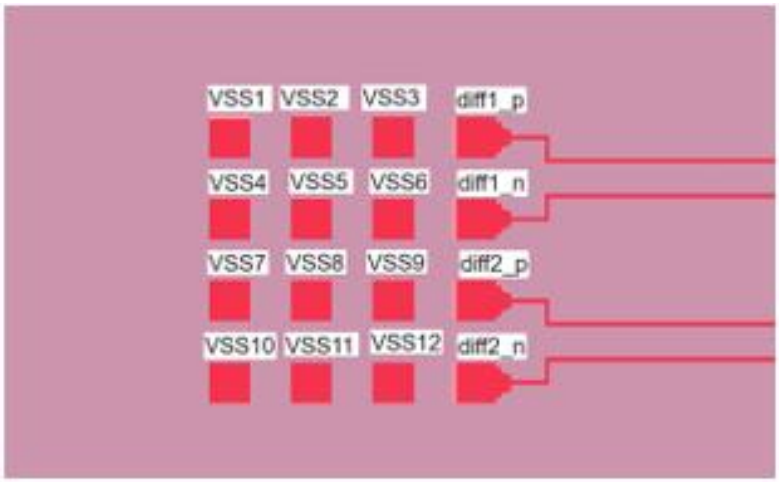


Figure 1: Interposer die1 pads

3.1. Interposer topologies

The simplest configuration for the interconnects is the microstrip line [11], [12]. The traces are placed on M11 in differential pairs and a continuous ground plane is placed on M8. The microstrip is easy to manufacture, and uses minimal space on the chip, since the signal traces do not require any guard traces. For this reason, no congestion to other signals is caused, but the drawback is that it does not offer any shielding between the two pairs, resulting in non-negligible crosstalk between the signal lines. The configuration is shown in Fig. 2.

Figure 2: Microstrip line topology

The geometric properties of the microstrip are chosen to match the differential impedance of both pairs to 100 Ω, using available tools [1]. The trace thickness and dielectric height are parameters given by the technology used. The trace width t and differential spacing s can be varied to achieve the desired impedance of 100 Ω.

Figure 2: Microstrip line cross section

Parameter	Value
h_1 (μm)	3.50
h_2 (μm)	6.30
ε_1	3.99
ε_2	5.63
t (μm)	1.40
w $(\mu\mu)$	54.17
s (μm)	30.16
Z_d (Ω)	100.03

Table 1: Microchip parameters

The design is analyzed in HFSS 3D Layout, using a setup with a mesh frequency at 2 GHz, and an interpolating sweep from 0 to 15 GHz with a 20 MHz step. The simulation also models the skin effect inside the conductors. The results for the microstrip are listed in Table 4.

3.2. Coplanar waveguide

The second configuration investigated is the coplanar waveguide. The waveguide resembles the microstrip, but the differential pairs are placed in between ground traces also placed on M11. These ground traces are wider than the signal traces and provide sufficient capacitive shielding between the differential pairs by blocking the electric field lines from one pair to the other [5], [11], [12].

The geometric characteristics are again chosen for appropriate impedance matching, using typical formulae [2]. These characteristics are reported in Table 2.

Parameter	Value
$h\ (\mu m)$	3.50
ε	4.32
$t\ (\mu m)$	1.40
$w\ (\mu\mu)$	5.59
$s\ (\mu m)$	20.00
$d\ (\mu m)$	5.00
$Z_d\ (\Omega)$	100.05

Table 2: Coplanar waveguide parameters

The results for the coplanar waveguide are reported in Table 4. It can be seen that the reflection coefficient is lower than for the microstrip and the insertion loss is higher. Thus, signal transmission is better through the waveguide. The most important result is the crosstalk, which has been reduced by 55% for the near-end and 75% for the far-end, respectively.

3.3. Coplanar stripline

The last configuration explored is the coplanar stripline. The stripline resembles the coplanar waveguide, but the signal traces are now placed on M9 instead of M11. On M11, an additional ground plane is utilized for enhanced shielding. For impedance matching, there are no analytical equations, such that the Q2D tool has been used. The cross section of the transmission line is simulated, with the geometric characteristics varying. The chosen characteristics are the features that provide 50 Ω impedance (for the single ended trace) at 2 GHz. The geometric properties are listed in Table 3.

Parameter	Value
$h\ (\mu m)$	3.50
ε	4.32
$t\ (\mu m)$	1.40
$w\ (\mu\mu)$	2.60
$s\ (\mu m)$	20.00
$d\ (\mu m)$	5.00
$w_{gs}\ (\mu m)$	20.00
$w_{gb}\ (\mu m)$	109.80
$Z_d\ (\Omega)$	100.05

Table 3: Coplanar stripline parameters

The results are reported in Table 4. The reflection coefficient has risen significantly in comparison to both of the previous configurations, while the insertion loss is reduced. This behavior can also be seen from the eye diagram, where the positive voltage has dropped below the acceptable limit of 0.4 V. Thus, the signal propagation is not acceptable. The crosstalk has dropped by 63% for the near end and by 89% for the far-end. These results are better than the coplanar waveguide but come at the cost of poor signal propagation.

Parameter	Value		
	Microstrip	Coplanar waveguide	Coplanar stripline
Reflection coefficient (dB)	-16.29	-17.46	-10.23
Insertion loss (dB)	-1.46	-1.10	-3.19
NEXT coupling coefficient (dB)	-86.54	-93.1	-96.65
FEXT coupling coefficient (dB)	-94.03	-96.6	-100.69
Eye amplitude (mV)	865.66	898.26	783.77
Eye height (mV)	842.97	890.35	770.34
NEXT crosstalk (%)	0.043	0.019	0.016
FEXT crosstalk (%)	0.032	0.0078	0.0035

Table 4: Comparison of different interconnect structures

4 Comparison & Conclusions

Comparing the three configurations investigated, the microstrip is the simplest configuration but suffers the most from crosstalk between the differential pairs. In this case study, the crosstalk generated is still very low, but in larger applications with many signals placed adjacent to each other, this crosstalk can become significant and degrade normal operation. The coplanar waveguide offers very high shielding between the pairs and exhibits the best signal transmission characteristics out of the three configurations. The coplanar stripline can offer the best shielding but at the cost of poor transmission characteristics. Thus, the best choice between the three configurations is chosen to be the coplanar waveguide. The waveguide is also simpler to manufacture than the coplanar stripline, which makes this interconnect the most suitable choice.

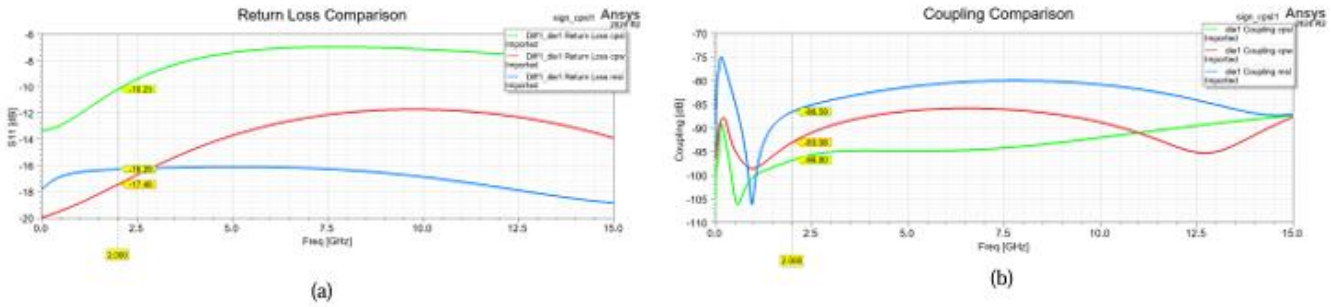


Figure 4: Return loss and coupling comparison

References

[1] Embedded Differential Microstrip Impedance Calculator. Retrieved from https://impedance.app.protoexpress.com/?appid=EMDPIMPCAL

[2] R. N. Simons (2001). Coplanar Waveguide Circuits, Components, and Systems. John Wiley and Sons, Inc.

[3] Effective dielectric calculation of two dielectric blocks in series. Retrieved from https://www.toppr.com/ask/question/a-parallel-plate-capacitor-is-made-of-two-dielectric-blocks-in series-one-of-the/

[4] Power management for processor core voltage requirements (Texas Instruments). Retrieved from https://www.ti.com/lit/an/slyt261/slyt261.pdf?ts=1708101831342&ref_url=https%253A%252F%252Fwww.google.com%252F

[5] Achieve High Routing Density with Grounded Coplanar Waveguide Technology (Cadence). Retrieved from https://resources.system-analysis.cadence.com/blog/msa2022-achieve-high-routing density-with-grounded-coplanar-waveguide-technology

[6] About 2.5D Technology. Retrieved from https://nhanced-semi.com/technology/about-2-5d technology/

[7] 2.5D and 3D IC Packaging. Retrieved from https://ase.aseglobal.com/3d-ic-packaging/

[8] Silicon Interposer: Ultimate Guide. Retrieved from https://anysilicon.com/semipedia/interposer/

[9] Systems on Interposer: “Chips” are So Passé. Retrieved from https://www.eejournal.com/article/20130813-interposer/

[10] Interposers: Fast, low-power inter-die conduits for 2.5D electrical signals. Retrieved from https://semiengineering.com/knowledge_centers/packaging/advanced-packaging/2-5d-ic/interposers/

[11] Ultimate Guide to Microstrip, Stripline and CPW in PCBs. Retrieved from https://jhdpcb.com/blog/microstrip-vs-stripline-vs-coplanar-waveguide/

[12] Comparing Microstrip and Grounded Coplanar Waveguide. Retrieved from https://www.microwavejournal.com/blogs/1-rog-blog/post/24374-comparing-microstrip-andgrounded-coplanar-waveguide

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 139 – 145

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Continuous and Symmetric Drain Current Compact Model for Nanoscale Triple-Gate FinFETs

M. K. Nakos, A. Tsormpatzoglou, T. A. Oproglidis, D. H. Tassis and C. A. Dimitriadis, Senior Member, IEEE[§§§§]

Abstract

In this work, we present a drain current compact model for nanoscale triple gate FinFETs satisfying continuity and source/drain symmetry. This is achieved by upgrading key equations of our original compact model. The formulated model is compact, valid in all regions of operation and its accuracy is validated through comparison with simulation data of nanoscale triple-gate transistor. The symmetry conditions are investigated and validated by performing the Gummel Symmetry Test.

Index Terms—Nanoscale triple-gate FinFETs, continuous and symmetric compact model.

1 Introduction

TRIPLE-GATE (TG) field-effect transistors (FETs) such as fin-shaped FETs (FinFETs) have been proposed for the 7 nm technology node and beyond [1], which exhibit excellent gate control for suppressing the short channel effects (SCEs) compared to the conventional planar MOSFETs [2], [3]. Generally, FinFETs are inherently symmetric devices and the source and drain contacts are interchangeable. Therefore, the compact models describing their behavior have to be electrically symmetric around the drain voltage of Vds = 0, a condition not usually satisfied.

The Gummel Symmetry Test (GST) was introduced as a benchmark test to qualify compact models developed for MOSFET circuit simulation and design [4], [5]. It has been demonstrated that a compact model can succeed to pass the GST when the potentials are referenced to the bulk [4], [6]. In the GST, the source and drain are driven in the opposite direction simultaneously by the applied voltage Vx. When Vx at the drain is swept from a negative to an equal positive value, the transistor under test is driven symmetrically from a reverse to a forward mode. In order to pass the GST, the symmetric drain current versus Vx plot must be an odd function and its nth order derivative should exist. One of the basic criteria used to decide the efficacy of a compact model is that the high-order derivatives of the drain current must be continuous at Vx = 0. Furthermore, the higher-order derivatives continuity implies that the second-order derivative should be equal to zero at Vx = 0.

From the compact models developed for multiple-gate transistors, most of them examine the

[§§§§] M. K. Nakos and A. Tsormpatzoglou are with the Department of Informatics and Telecommunications, University of Ioannina, Arta, Greece.

T. A. Oproglidis, D. H. Tassis and C. A. Dimitriadis are with the Department of Physics, Aristotle University of Thessaloniki, 54124 Thessaloniki Greece. (e-mail: cdimitri@physics.auth.gr).

symmetry condition either for surface potential-based or charge-based compact models in independent-gate operation double-gate (DG) MOSFETs [7] or symmetric DG MOSFETs [8]-[13]. For the surface-potential based BSIM-CMG compact models in common-gate TG FinFETs presented in [14], [15], the symmetry condition has been examined in long-channel devices. For our charge-based compact model presented in [16] for nanoscale TG FinFETs, analogue gain-stage circuit was used for assessing its linear behavior, showing that the circuit's usable range is limited for input signal of power below about -15 dB. However, this compact model failed to pass the GST.

In this work, we update our previous drain current compact model for TG FinFETs [16] to a continuous and symmetric by modifying drastically the channel length modulation effect, the effective carrier mobility, the threshold voltage and the subthreshold swing coefficient, in a way satisfying the GST. The properly formulated initial charge-based compact model first is evaluated with simulation data in order to validate its efficiency. Then, the source/drain symmetry is verified by presenting high order derivatives of the drain current.

2 FORMULATION OF THE INITIAL COMPACT MODEL

The charge-based analytical compact model developed in our previous work [16] for TG FinFETs includes the SCEs, saturation velocity, source/drain series resistance and mobility degradation effects. For gate/source/drain potentials referenced to the bulk potential V_b [4], [6], the normalized inversion charge of the original model is formulated by the principal branch of the Lambert function W0(x) as follows:

$$q = W_0 e \left[e^{\frac{V_g - V_b - V_i - (V - V_b)}{2V_{th}}} \frac{e^{\frac{V_g - V_b - V_i - (V - V_s)}{2\eta V_{th}}}}{4e^{\frac{V_i + V_b - \phi_1}{4e}} + e^{\frac{V_g - V_b - V_i - (V - V_b)}{2\eta V_{th}}}} \right] \tag{1}$$

where V_g is the gate voltage, V_t is the threshold voltage, V_{th} is the thermal voltage, V is the quasi-Fermi potential which varies from 0 at the source to V_{ds} at the drain, V_0 = 1 V represents a normalizing factor, η is the subthreshold swing coefficient, $V_{fb} = \Delta\varphi - \left(\frac{kT}{q}\right) \ln\left(\frac{N_A}{n_i}\right)$ is the flat-band voltage, $\Delta\varphi$ is the gate work function referenced to silicon, N_A is the doping concentration of the silicon channel, ni is the intrinsic carrier concentration and c1 is a fitting parameter for adjusting the transition from the below to the above threshold region. The use of the Lambert function $W_0(x)$ in (1) allow us to exploit the function's inherent continuity in order to eliminate discontinuities and asymmetries by expressing key equations as a function of the normalized charge density. The following analytical approximation for the Lambert function $W_0(x)$ is used, with relative error less than 0.1% for wide range of positive x values:

$$W_0(x) = \frac{W_1(x) W_2(x)}{\sqrt{W_1^2(x) + W_2^2(x)}} \tag{2}$$

where

$$W_1(x) = (1 + 0.233878) \cdot \ln\left[\frac{12(x + 0.24)}{5 \ln[2.4(x + 0.761034)]}\right] - 0.233878 \cdot \ln[2(x + 0.24)],$$
$$W_2(x) = x \tag{3}$$

In order to model the channel length modulation (CLM) effect, we originally included a mathematical smoothing function for the effective drain-source voltage ($V_{ds,eff}$) in our compact model that was valid in the saturation region, shortening the total length by a factor ΔL [16]. Performing the GST, while using this equation for $V_{ds,eff}$, leads to asymmetries because ΔL has a value only in the saturation region; otherwise it was fixed to zero. In general, a compact model has to include expression for $V_{ds,eff}$ that is inherently symmetric. In this work, we propose an alternative continuous function for ΔL in terms of the normalized inversion charge density to render the smoothing function symmetric and thus enable the compact model to pass the GST. In the forward mode operation (i.e. $V_{ds} > 0$ and $q_s > q_d$), the value of ΔL equivalent to that obtained using the $V_{d,eff}$ in [16] is

$$\Delta L_{eff} = \lambda_{av} \ln\left[1 + \frac{(V_d - V_s) - 2.2V_{th}(q_s - q_d)}{V_E}\right] \quad (4)$$

where λ_{eff} is the effective natural length of the TG FinFET, q_s and q_d are derived from (1) for $V = V_s$ and $V = V_d$, respectively and V_E is a fitting parameter. When the source and drain contacts are reversed (i.e. $V_{ds} < 0$ and $q_s < q_d$), ΔL will acquire in reverse mode a different value than that of the forward mode and as a result the model will be asymmetric. In order to bypass this issue, the expression (4) for ΔL is written as

$$\Delta L = \lambda_{eff} \ln\left[1 + \frac{V_{dn} - 2.2\left(\frac{V_d - V_s}{V_{dn}}\right)V_{th}(q_s - q_d)}{V_E}\right] \quad (5)$$

where

$$V_{dn} = \frac{\ln\left[1 + e^{100(V_d - V_s)} + e^{-100(V_d - V_s)}\right]}{100} \quad (6)$$

In relation (6), V_{dn} maintains positive value regardless of the source/drain reversal. In this way, the hyperbolic tangent contained in V_{dseff} of [16] is removed.

Moreover, it has been reported that asymmetries and discontinuities may also be introduced in a compact model because of the expression used for the effective carrier mobility (μ_{eff}) [17]. Following [18], μ_{eff} is analytically expressed in terms of the normalized inversion charge density q_0, allowing the compact model to satisfy the symmetry condition:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta_1 V_{th} q_0 + \theta_2 (V_{th} q_0)^2} \quad (7)$$

where μ_o is the low-field mobility, θ_1 and θ_2 are the mobility attenuation coefficients of first and second order, respectively and q_0 is described by the Lambert function $W_0(x)$

$$q_0 = W_0\left[\left(e^{\frac{V_g - V_b - V_t}{2V_{th}}}\right)\right] \quad (8)$$

The mobility attenuation coefficient θ_1, which includes the linear mobility attenuation coefficient $\theta_{1,0}$, the saturation velocity υsat and the series resistance R_{sd} [16], is expressed as a function of V_{dn} by:

$$\theta_1 = \theta_{1,0}\left[1 + \frac{\mu_0 V_{dn}}{v_{sat}(L - \Delta L)}\right] + \frac{\mu_0 W_{eff} C_{ox}}{L - \Delta L} R_{sd} \quad (9)$$

where $W_{eff} = 2H_{fin} + W_{fin}$ is the effective gate width of the TG FinFET, H_{fin} is the fin height and W_{fin} is the fin width. The mobility attenuation parameters μ_o, $\theta_{1,0}$, θ_2 and R_{sd} can be extracted using our new Y-

function based methodology presented in [18].

Finally, performing the GST the impact of Vds on the threshold voltage Vt and the ideality coefficient η is taken into account. Following [16], the device parameters V_t and η are redefined in terms of V_{dn} by the analytical expressions

$$V_t = V_{fb} - \frac{A_1(V_{bi}+V_{dn})+A_2V_{bi}}{1-(A_1+A_2)} - \frac{V_{th}}{1-(A_1+A_2)} \ln\left[\frac{Q_n N_A}{n_i^2 W_{fin}}\right] \quad (10)$$

$$\eta = \frac{1}{1-(A_1+A_2)} \quad (11)$$

where the built-in potential across the source/drain junctions is $V_{bi} = V_{th} \ln\left(\frac{N_D N_A}{n_i^2}\right)$ and N_D is the doping concentration of the source/drain contacts. The parameters Q_{th}, A_1 and A_2 presented in [16] are redefined through the modified drain source bias voltage V_{dn} in the minimum potential at the conductive path.

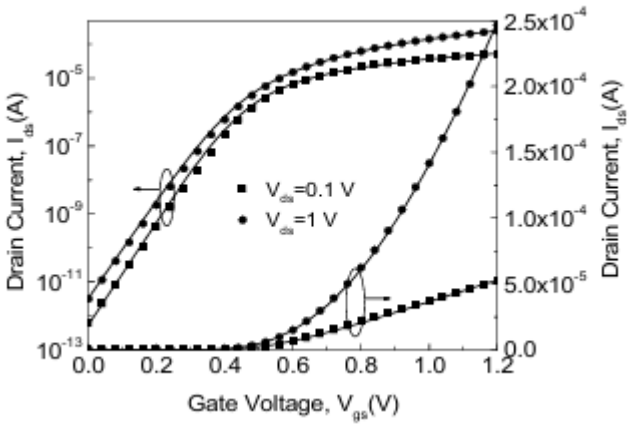


Fig. 1. Transfer characteristics in linear and semi-logarithmic representations of TG FinFET with dimensions L = 40 nm, Hfin = 20 nm, Wfin = 20 nm and tox = 1 nm. Symbols represent simulation data and solid lines the model results using the extracted model parameters: c1 =3, VE = 0.35 V, Vgc = 0.4 V, θ2 = 0, θ1,0 = 0.1 V-1, μ0 = 200 cm2/Vs, and Rsd = 125 Ω.

Our previous analytical drain current compact model [16], upgraded to a new compact model being continuous and symmetric by including the new modifications for ΔL, μ_{eff}, V_t and η presented in this work, is formulated as:

$$I_{ds} = W_{eff}\mu_{eff}C_{ox}(2V_{th})^2\left[\frac{q_s-q_d}{L} + \frac{1}{2}\frac{q_s^2-q_d^2}{L-\Delta L}\right] \quad (12)$$

The modified compact model (12) has been validated with simulation data of TG FinFET with channel length L = 40 nm, fin height H_{fin} = 20 nm, fin width W_{fin} = 20 nm, equivalent gate oxide thickness t_{ox} = 1 nm, doping concentration of silicon channel N_A = 1.45 × 10^{10} cm^{-3}, doping concentration of the source/drain contact regions N_D = 10^{20} cm^{-3} and mid-gap gate metal work-function 4.71 eV corresponding to Δφ = 0 (i.e.V_{fb} = 0) [18]. The results of the validation are presented in Fig. 1. Using the extracted model parameters presented in the caption of Fig. 1, the agreement between simulation data and model results is very good.

3 THE GUMMEL SYMMETRY TEST

The model GST results for the TG FinFET of Fig. 1 were investigated with the gate/source/drain potentials referenced to zero bulk potential (Vb = 0). The inset of Fig. 2(a) shows the schematic of the GST circuit, where Vx is swept from -30 mV to +30 mV, while the gate voltage maintains a constant value. The model GST results for gate voltages Vg = 0, 0.25, 0.5 and 1 V, normalized to their absolute maximum values, are shown in Figs. 2(a)-2(d). It is evident that the derivatives of the modeled drain current up to the third order maintain continuity and symmetry around Vds = 0 for all gate voltages and the second derivative is zero at Vx = 0, validating the symmetric nature of the upgraded compact model.

4 CONCLUSION

In this work, we discuss certain formulations in our previous charge-based analytical drain current compact model for nanoscale TG FinFETs [16]. The new model is continuous and symmetric around V_{ds} = 0, verified by performing the Gummel Symmetry Test.

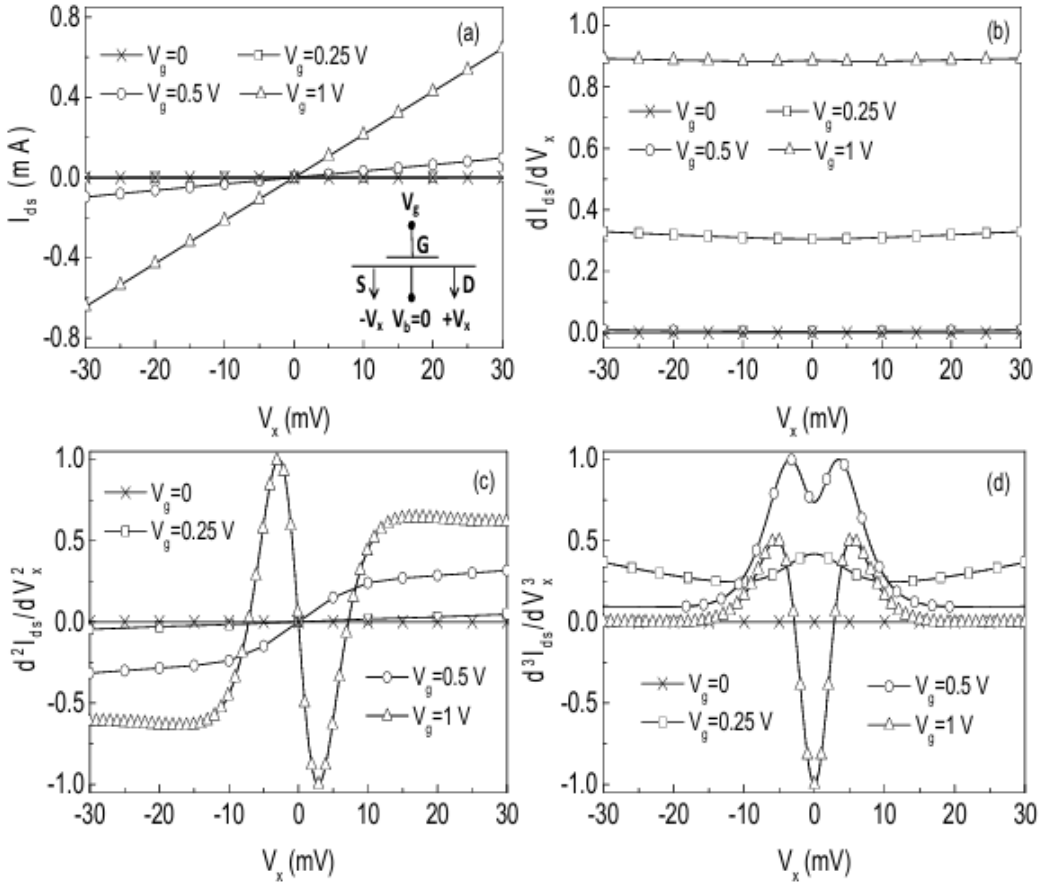


Fig. 2. Model GST results for derivatives of the drain current up to the third order at Vg = 0, 0.25, 0.5 and 1 V for TG FinFET with dimensions L = 40 nm, Hfin = 20 nm and Wfin = 20 nm. The current Ids(Vx) and its derivatives have been normalized to their absolute maximum values. The inset in Fig. 2(a) shows the schematic of the GST circuit.

The validation with simulated results, as well as the continuity and the symmetric nature of the formulated model, certify that it is suitable for implementation in circuit simulation tools.

References

[1] E. Sicard and L. Trojman, "Introducing 5-nm FinFET technology in microwind," 2021 (https://hal.science/hal-03254444).

[2] J.-T. Park and J.-P. Colinge, "Multiple-gate SOI MOSFETs: device design guidelines," IEEE Trans. Electron Dev., vol. 49, no. 12, pp. 2222-2229, 2002.

[3] W. Zhang, J. G. Fossum, L. Mathew, and Y. Du, "Physical insights regarding design and performance of independent-gate FinFETs," IEEE Trans. Electron Dev., vol. 52, no. 10, pp. 2198-2206, 2005.

[4] C. C. McAndrew, "Validation of MOSFET model source/drain symmetry", IEEE Trans. Electron Devices, vol. 53, no. 9, pp. 2202 2206, 2006, doi: 10.1109/TED.2006.881005.

[5] C. C. McAndrew, H. K. Gummel, K. Singhal and K. Singhal, "A comparison and evaluation of state-of-the-art compact MOSFET models", AT&T Bell Laboratories Memorandum, 52864-910119-01TM.

[6] G. H. See, X. Zhou, K. Chandrasekaran, S. B. Chiah, Z. Zhu, G. H. Lim, C. Wei, S. Lin and G. Zhu, "Gummel symmetry with higher-order derivatives in MOSFET compact models", Proc. NSTI-Nanotech, vol.3, pp. 613-616, 2007.

[7] D. D. Lu, M. V. Dunga, C-H. Lin, A. M. Niknejad, and C. Hu, "A multi-gate MOSFET compact model featuring independent-gate operation," 2007 IEEE International Electron Devices Meeting, Washington, DC, USA, 2007, pp. 565-568, doi: 10.1109/IEDM. 2007.4419001.

[8] V. Hariharan, J. Vasi, V. R. Pao, "Drain current model including velocity saturation for symmetric double-gate MOSFETs", IEEE Trans. Electron Devices, vol. 55, no. 8, pp. 2173-2180, 2008, doi: 10.1109/ TED.2008.926745.

[9] G. D. J. Smit, A. J. Scholten, N. Serra, R. M. T. Pijper, R. van Langevelde, A. Mercha, G. Gildenblat, and D. B. M. Klaassen, "PSP based compact FinFET model describing dc and RF measurements," in IEDM Tech. Dig., pp. 175–178, 2006, doi: 10.1109/IEDM. 2006.346990.

[10] G. Zhu, G. H. See, S. Lin, and X. Zhou, "Ground-referenced" model for three-terminal symmetric double-gate MOSFETs with source/drain symmetry", IEEE Trans. Electron Devices, vol. 55, no. 9, pp. 2526 2530, 2008, doi: 10.1109/TED.2008.928022.

[11] G. Zhu, X. Zhou, G. H. See, S. Lin, C. Wei, J. Zhang, "A unified compact model for FinFET and silicon nanowire MOSFETs," Proc. NSTI-Nanotech, vol. 3, pp. 588-591, 2009.

[12] J. Song, B. Yu, Y. Tuan, and Y. Taur, "Review on compact modeling of multiple-gate MOSFETs," IEEE Trans. on Circuits and Systems, vol. 56, no. 8, pp. 1858-1869, 2009, doi: 10.1109/TCSI.2009.2028416.

[13] X. Zhou, G. Zhu, M. K. Srikanth, S. Lin, Z. Chen, J. Zhang, C. Wei, Y. Yan, R. Selvakumar and Z. Wang, "Xsim: Benchmark tests for the unified DG/GAA MOSFET compact model," Proc. NSTI-Nanotech, vol. 2, pp. 785-788, 2010.

[14] N. Paydavosi, S. S. Venugopalan, Y. Singh, J. P. Duarte, S. Jandhyala, A. M. Niknejad, and C. C. Hu, "BSIM-SPICE models enable FinFET and UTB IC designs," in IEEE Access, doi: 10.1109/ACCESS. 2013.2260816.

[15] J. P. Duarte, S. Khandelwal, A. Medury, C. Hu, P. Kushwaha, H. Agarwal, A. Dasgupta, and Y. S. Chauhan, "BSIM-CMG: Standard FinFET compact model for advanced circuit design," ESSCIRC Conference 2015 - 41st European Solid-State Circuits Conference (ESSCIRC), Graz, Austria, 2015, pp. 196-201, doi: 10.1109/ ESSCIRC. 2015.7313862.

[16] N. Fasarakis, A. Tsormpatzoglou, D. H. Tassis, I. Pappas, K. Papathanasiou, M. Bucher, G. Ghibaudo, and C. A. Dimitriadis, "Compact model of drain current in short-channel triple-gate FinFETs," IEEE Trans. Electron Devices, vol. 59, no. 7, pp. 1891-1898, 2012, doi: 10.1109/TED.2012.2195318.

[17] S. B. Chiah, X. Zhou, K.-Y. Lim, L. Chan and S. Chu, "Source-Drain symmetry in unified regional MOSFET model", IEEE Trans. Electron Device Lett., vol. 25, no. 5, 2004, doi: 10.1109/LED.2004.826513.

[18] A. Tsormpatzoglou, K. Papathanasiou, N. Fasarakis, D. H. Tassis, G. Ghibaudo, and C. A. Dimitriadis, "A Lambert-function charge based methodology for extracting electrical parameters of nano-scale FinFETs," IEEE Trans. Electron Devices, vol. 59, no. 12, pp. 3299 3305, 2012, doi: 10.1109/TED.2012.2222647.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 146 – 153

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

A low-rank balanced truncation approach for large-scale RLCk model order reduction based on extended Krylov subspace and a frequency-aware convergence criterion FinFETs

Christos Giamouzis, Dimitrios Garyfallou, Nestor Evmorfopoulos, and George Stamoulis
Dept. of Electrical and Computer Engineering, University of Thessaly, Volos, Greece
{cgiamouzis, digaryfa, nestevmo, georges}@e-ce.uth.gr

Abstract

Model order reduction (MOR) is essential in integrated circuit design, particularly when dealing with large-scale electromagnetic models extracted from complex designs. The nu merous passive elements introduced in these models pose significant challenges in the sim ulation process. MOR methods based on balanced truncation (BT) help address these challenges by producing compact reduced-order models (ROMs) that preserve the original model's input-output port behavior. In this work, we present an extended Krylov subspace based BT approach with a frequency-aware convergence criterion and efficient implemen tation techniques for reducing large-scale models. Experimental results indicate that our method generates accurate and compact ROMs while achieving up to ×22 smaller ROMs with similar accuracy compared to ANSYS RaptorX™ ROMs for large-scale benchmarks.

1 Introduction

Electromagnetic model extraction is crucial for designing and verifying integrated circuits (ICs), enabling precise simulation of the passive elements of the design. However, simulating extracted RLCk models with millions of elements and multiple ports is extremely computationally expen sive. Model order reduction (MOR) can reduce the complexity of such models while maintaining accurate input/output port behavior [1, 2]. By constructing reduced-order models (ROMs) that capture the essential dynamics of the original system, MOR can significantly reduce simulation time, enabling faster design iterations in IC development.

There are two main approaches to MOR. Moment matching (MM) methods are preferred for their efficiency, but they require manual selection of the number of moments [1]. Most importantly, they correlate the final ROM size with the number of moments and ports, limiting scalability. On the contrary, balanced truncation (BT) provides explicit theoretical bounds for the approximation error and is independent of the number of ports [2]. However, BT is restricted to small-scale models due to the high computational complexity of solving Lyapunov equations [2].

In this paper, we introduce an efficient low-rank BT technique to address the main scalability issue of the conventional BT approach. Specifically, we employ the extended Krylov subspace (EKS) method,

which effectively solves the Lyapunov equations, drastically reducing the computational load of BT [3]. Additionally, we incorporate a frequency-aware convergence criterion, ensuring accuracy in the frequency range of interest. Experimental evaluation indicates that the proposed method can be integrated into commercial extraction tools, such as the ANSYS RaptorX™ [4], to generate more compact ROMs of large-scale multi-port RLCk models.

2 Background

Consider the modified nodal analysis (MNA) description [5] of an n-node, m-branch (inductive), p-input, and q-output RLCk circuit in the time domain:

$$\begin{pmatrix} G_n & E \\ -E^T & 0 \end{pmatrix}\begin{pmatrix} v(t) \\ i(t) \end{pmatrix} + \begin{pmatrix} C_n & 0 \\ 0 & M \end{pmatrix}\begin{pmatrix} v(t) \\ i(t) \end{pmatrix} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix} u(t),\ y(t) = \begin{pmatrix} L_1 & 0 \end{pmatrix}\begin{pmatrix} v(t) \\ i(t) \end{pmatrix} \quad (1)$$

where $\boldsymbol{G_n} \in R^{n\times n}$G (node conductance matrix), $\boldsymbol{C_n} \in R^{n\times n}$ (node capacitance matrix), $\mathrm{M} \in R^{\mathrm{m}\times\mathrm{m}}$ (branch inductance matrix), $\mathbf{E} \in R^{\mathrm{n}\times\mathrm{m}}$ (node-to-branch incidence matrix), $\mathbf{v} \in R^{\mathrm{m}}$ (vector of node voltages), i ∈ Rm (vector of inductive branch currents), $\mathbf{u} \in R^{p}$ (vector of v(t) ≡ dv(t) input excitations), $\boldsymbol{B_1} \in R^{n\times\mathrm{p}}$ (input-to-node connectivity matrix), $\mathbf{y} \in R^{q}$ (vector of output measurements), and $\boldsymbol{L_1} \in R^{\mathrm{q}\times\mathrm{n}}$ (node-to-output connectivity matrix). Moreover, we denote $\dot{v}(t) \equiv \frac{dv(t)}{dt}$ and $i(t) \equiv \frac{di(t)}{dt}$. If we now define the model order as N ≡ n+m, the state vector as $x(t) \equiv \begin{pmatrix} v(t) \\ i(t) \end{pmatrix}$, and also

$$G = -\begin{pmatrix} G_n & E \\ -E^T & 0 \end{pmatrix},\quad C \equiv \begin{pmatrix} C_n & 0 \\ 0 & M \end{pmatrix},\quad B \equiv \begin{pmatrix} B_1 \\ 0 \end{pmatrix},\quad \mathrm{L} \equiv \begin{pmatrix} L_1 & 0 \end{pmatrix})$$

then Eq. (1) can be written in the generalized state-space form, or so-called descriptor form:

$$C\frac{dx(t)}{dt} = Gx(t) + Bu(t),\quad y(t) = Lx(t) \quad (2)$$

The objective of MOR is to produce an equivalent ROM:

$$\tilde{C}\frac{d\tilde{x}(t)}{dt} = \tilde{G}\tilde{x}(t) + \tilde{B}u(t),\quad \tilde{y}(t) = \tilde{L}\tilde{x}(t) \quad (3)$$

where $\tilde{\mathbf{G}}, \tilde{\mathbf{C}} \in \mathrm{R}^{\mathrm{r}\times\mathrm{r}}$, $\tilde{\boldsymbol{B}} \in \mathrm{R}^{\mathrm{r}\times\mathrm{p}}$, $\tilde{\boldsymbol{L}} \in \mathrm{R}^{\mathrm{q}\times\mathrm{r}}$ the reduced order r << N, and the output error is bounded as:$||\tilde{\boldsymbol{y}}(t) - \boldsymbol{y}(t)||_2 < \varepsilon||\boldsymbol{u}(t)||_2$ for given **u**(t) and small ε. The output error bound in the frequency domain is:$||\tilde{y}(s) - \boldsymbol{y}(s)||_2 < \varepsilon||\boldsymbol{u}(s)"|_2$ via Plancherel's theorem [6]. If

$$\mathbf{H}(\mathrm{s}) = \mathbf{L}(\mathrm{s}\mathbf{C} - \mathrm{G})^{-1}\mathbf{B},\quad \tilde{\mathbf{H}}(\mathrm{s}) = \tilde{\mathrm{L}}\left(\mathrm{s}\tilde{\mathbf{C}} - \tilde{\mathbf{G}}\right)^{-1}\tilde{\mathbf{B}}$$

are the transfer functions of the original model and the ROM, the corresponding output error is:

$$||\tilde{\boldsymbol{y}}(s) - \boldsymbol{y}(s)||_2 = ||\tilde{\boldsymbol{H}}(s)\boldsymbol{u}(s) - \boldsymbol{H}(s)\boldsymbol{u}(s)||_2 \le ||\tilde{H}(s) - \boldsymbol{H}(s)||_\infty|\boldsymbol{u}(s)||_2$$

where $||\,.\,||_\infty$ is the $\mathcal{L}_2$ matrix norm or H_∞ norm of a rational transfer function. Thus, to bound this error, we need to bound the distance between the transfer functions: $||\tilde{H}(s) - \boldsymbol{H}(s)||_\infty < \varepsilon < \varepsilon$.

3 MOR by Balanced Truncation

BT relies on the computation of the controllability Gramian **P** and observability Gramian **Q**, which are calculated as the solutions of the following **L**yapunov matrix equations [2]:

$$(\boldsymbol{C}^{-1}\mathbf{G})\mathbf{P} + \mathbf{P}(\mathbf{C}^{-1}\mathbf{G})^T = -(\boldsymbol{C}^{-1}\mathbf{B})(\boldsymbol{C}^{-1}\mathbf{B})^T,\ (\mathbf{C}^{-1}\mathbf{G})^T\mathbf{Q} + \mathbf{Q}(\boldsymbol{C}^{-1}\mathbf{G}) = -\boldsymbol{L}^{\mathrm{T}}\mathbf{L} \quad (4)$$

The controllability Gramian **P** describes the degree to which the states are controllable by the inputs, while the observability Gramian **Q** reflects the degree to which the states are observ able at the outputs. A ROM can theoretically be generated by eliminating the states that are difficult to control or observe. However, in the original state-space coordinates, certain states may be easy to control but difficult to observe, and vice versa. The process of "balancing" transforms the state vector to a new coordinate system, where the controllability and observ ability of each state are balanced, meaning each state is equally difficult to control and observe. An appropriate transformation **Tx**(*t*) exists, leading to the balanced state-space model:

$$\boldsymbol{TCT}^{-1}\frac{d\left(\boldsymbol{Tx}(t)\right)}{dt} = \boldsymbol{TGT}^{-1}\left(\boldsymbol{Tx}(t)\right) + \boldsymbol{TBu}(t), \quad y(t) = \boldsymbol{L}T^{-1}\left(\boldsymbol{Tx}(t)\right)$$

This balanced representation preserves the system's transfer function **H**(s) and simplifies to **P** = **Q**=*diag*($\sigma_1, \sigma_2, \ldots, \sigma_N$) [2], where σ_i are the Hankel singular values (HSVs). These HSVs are the square roots of the eigenvalues of the product **PQ**, i.e., $\sigma_i = \sqrt{\lambda_i(PQ)}$. In the above balanced model, the states with the largest HSVs are the easiest to both control and observe. If *r* of them are retained (truncating the *N* −*r* states associated with the smallest HSVs), the error between the original and the reduced-order transfer functions is bounded as:

$$||\boldsymbol{H}(s) - \tilde{\boldsymbol{H}}(s)||_\infty \leq 2(\sigma_{r+1} + \sigma_{r+2} + \cdots + \sigma_N)$$

The above serves as an "a-priori" criterion that offers flexibility by allowing either the specification of a ROM size r to compute the error or a target error (target error) to determine the number r of HSVs to be preserved. This adaptability is a key advantage of BT over MM methods.

Algorithm 1 MOR by balanced truncation

Inputs: $\mathbf{G}, \mathbf{C}, \mathbf{B}, \mathbf{L}$
Outputs: $\tilde{\mathbf{G}}, \tilde{\mathbf{C}}, \tilde{\mathbf{B}}, \tilde{\mathbf{L}}$

1: Solve the Lyapunov equations to obtain the Gramian matrices $\mathbf{P}$ and $\mathbf{Q}$ [7]
2: Compute the SVD of the Gramian matrices: $\mathbf{P} = \mathbf{U}_P\mathbf{\Sigma}_P\mathbf{V}_P^T$ and $\mathbf{Q} = \mathbf{U}_Q\mathbf{\Sigma}_Q\mathbf{V}_Q^T$
3: Find the square root of the Gramian matrices: $\mathbf{Z}_P = \mathbf{U}_P\mathbf{\Sigma}_P^{1/2}$ and $\mathbf{Z}_Q = \mathbf{U}_Q\mathbf{\Sigma}_Q^{1/2}$
4: Compute the SVD of the product of the roots: $\mathbf{Z}_Q^T\mathbf{Z}_P = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$
5: Compute transformation matrices: $\mathbf{T}_{(r\times N)} = \mathbf{\Sigma}_{(r\times r)}^{-1/2}\mathbf{U}_{(r\times N)}\mathbf{Z}_Q^T$, $\mathbf{T}_{(N\times r)}^{-1} = \mathbf{Z}_P\mathbf{V}_{(N\times r)}\mathbf{\Sigma}_{(r\times r)}^{-1/2}$
6: Compute ROM: $\tilde{\mathbf{G}} = \mathbf{T}_{(r\times N)}\mathbf{G}\mathbf{T}_{(N\times r)}^{-1}$, $\tilde{\mathbf{C}} = \mathbf{T}_{(r\times N)}\mathbf{C}\mathbf{T}_{(N\times r)}^{-1}$, $\tilde{\mathbf{B}} = \mathbf{T}_{(r\times N)}\mathbf{B}$, $\tilde{\mathbf{L}} = \mathbf{L}\mathbf{T}_{(N\times r)}^{-1}$

The main steps of the BT procedure are summarized in Algorithm 1. The main limitation of BTisits high computational and memory cost, which makes it impractical for large-scale models (with N over a few thousand states). This is due to the computationally expensive operations required, such as solving Lyapunov equations and performing singular value decomposition (SVD), both of which have a complexity of O(N3). Additionally, they are applied on dense matrices, since the Gramians **P, Q** are dense even if the system matrices **C, G, B, L** are sparse.

However, the products $(\boldsymbol{C}^{-1}\mathbf{B})(\boldsymbol{C}^{-1}\mathbf{B})^T$ and $\boldsymbol{L}^T\boldsymbol{L}$ have a much lower numerical rank com pared to *N*, as *p,q* << *N*. This results in low-rank Gramians that can be approximated using low-rank techniques, significantly reducing the complexity and memory requirements for solving the Lyapunov equations and performing SVD, which are now performed with a complexity of order k rather than N.

3.1. Low-rank BT MOR

The essence of low-rank BT MOR is to iteratively project the Lyapunov equations onto a lower dimensional Krylov subspace and solve the resulting small-scale equations to obtain low-rank approximate solutions of Eq. (4). The k-dimensional standard Krylov subspace is defined as:

$$\mathrm{K_k}(\mathbf{G_C}, \mathbf{B_C}) = \mathrm{span}\{\mathbf{B_C}, \mathrm{G_C}\mathbf{B_C}, \mathbf{G_C^2}\mathbf{B_C}, \dots, \mathbf{G_C^{k-1}}\mathbf{B_C}\}$$

where:$\boldsymbol{G_C} \equiv \boldsymbol{C^{-1}G}, \quad \boldsymbol{B_C} \equiv \boldsymbol{C^{-1}B}$. If $\mathbf{K} \in R^{N\times k}$ (k << N) is a projection matrix whose columns span the *k*-dimensional standard Krylov subspace, then the projected Lyapunov equation (for the controllability Gramian **P**) onto $K_k(\boldsymbol{G_C}, \boldsymbol{B_C})$ is:

$$\left(\mathbf{K^T G_C K}\right)\mathbf{X} + \mathbf{X}\left(\mathbf{K^T G_C K}\right)^{\mathbf{T}} = -\left(\mathbf{K^T B_C B_C^T K}\right). \tag{5}$$

(the same holds true for the observability Gramian **Q** with $\boldsymbol{G_C^T}$, $\boldsymbol{L^T}$ in place of $\boldsymbol{G_C}$, B_C). The solution $\mathbf{X} \in \boldsymbol{R^{k\times k}}$ of Eq. (5) can be back-projected to the *N*-dimensional space to give an approximate solution $\mathbf{P} \approx \left(\boldsymbol{KXK^T}\right)$P ≈ KXKT for the original large-scale Eq. (4), and a low-rank factor $\mathbf{Z} \in \boldsymbol{R^{N\times k}}$ of **P** can be obtained as $\mathbf{Z = KU\Sigma^{1/2}}$, where [**U,Σ,V**] = *SVD*(**X**) and $\mathbf{P \approx ZZ^T}$.

While the projection process is independent of the chosen subspace, its effectiveness heavily relies on it. The convergence to an accurate solution can be accelerated by enhancing the stan dard Krylov subspace $\boldsymbol{K_k}(\boldsymbol{G_C}, \boldsymbol{B_C})$ with information from the subspace $K_k(G_C^{-1}, \boldsymbol{B_C})$, which corresponds to the inverse matrix G_C^{-1}, leading to the EKS [3, 8]:

$$\boldsymbol{K_k^C}(\mathbf{G_C}, \mathbf{B_C}) = \boldsymbol{span}\{\mathbf{B_C}, \mathbf{G_C^{-1}B_C}, \mathbf{G_C B_C}, \mathbf{G_C^{-2}B_C}, \mathbf{G_C^2 B_C}, \dots, \mathbf{G_C^{-(k-1)}B_C}, \mathbf{G_C^{k-1}B_C}\} \tag{6}$$

The EKS method (EKSM) begins with the vectors $\{\boldsymbol{B_C}, \boldsymbol{G_C^{-1}B_C}\}$. and iteratively builds an EKS $\boldsymbol{K_k^C}(\boldsymbol{G_C}, \boldsymbol{B_C})$) of increasing dimension, solving the projected Lyapunov Eq. (5) in each iteration, until a sufficiently accurate approximation of the solution of Eq. (4) is achieved. The complete EKSM is presented in Algorithm 2. Below are some efficient implementation details:

- **Matrix inversion by linear solves**: Algorithm 2 uses the system matrices $\boldsymbol{G}, C, G^T$. $\boldsymbol{G^T}$instead of
  $\boldsymbol{G_C} \equiv \boldsymbol{C^{-1}G}$ or $\boldsymbol{G_C^T} \equiv \boldsymbol{(C^{-1}G)^T}$ since the (generally dense) inverse matrices are only required for products with *p* vectors (in step 2) and *2pj* vectors (in steps 4 and 11 of each iteration), which can be handled as linear solves like CY = R and **GY = R** (or $\mathbf{C^TY=R}$, $\mathbf{G^TY=R}$), using either direct or iterative methods [9].
- **Handling of sparse/dense matrices**: Matrix **M** of Eq. (1) is typically very dense due to the huge number of mutual inductances. To efficiently handle both sparse ($\mathbf{C_n}$) and dense (**M**) matrix blocks of **C**, we use specialized data structures and numerical techniques. This includes parallel CPU-optimized methods for sparse matrices and GPU-accelerated techniques [10] for dense matrices.
- **Solution of the small-scale Lyapunov equations**: To solve the small-scale (*2pj* × *2pj*) Lyapunov equations in step 5 of each iteration, we employ the Bartels-Stewart algorithm [7].
- **Convergence criterion**: The error estimation [11] relies on the ROM transfer function ˜ H(s) and is described by:

$$\max_{i=1,\dots,l} \frac{|\widetilde{\boldsymbol{H}_j}(s_i) - \widetilde{\boldsymbol{H}_{j-1}}(s_i)|_\infty}{|\widetilde{\boldsymbol{H}_j}(s_i)|_\infty}$$

where $\widetilde{\boldsymbol{H}_j}(s_i)$ is the ROM transfer function at the *j-th* iteration (calculated at frequency $s_i = 2\pi f_i$) and l is the number of evaluated frequency points evenly distributed across a frequency range [f_{min}, f_{max}]. The proposed criterion offers insight into the extent to which the transfer function changes between iterations at the frequencies of interest. Moreover, it proves to be practical and effective for circuit simulation problems, where designers are only interested in the circuit's behavior in certain frequency windows. The iterative procedure stops when the error remains below a certain threshold (*tol*) for three consecutive iterations.

4 Experimental Evaluation

We evaluated EKSM using RLCk models extracted from various circuits via ANSYS Rap torX™ [4]. The evaluated designs consist of a phase-locked loop (PLL), an analog mixer, a time-interleaved digital-to-analog converter (TI DAC), an injection-locked frequency multiplier (ILFM), a VGA circuit, hybrid couplers (HCs), Wilkinson power dividers (WPDs), and typical transceiver blocks, such as low-noise-amplifiers (LNAs) and oscillators (VCO). Their detailed characteristics are listed in Tables 1 and 2. Two experiments were conducted: in the first one, we used small-scale benchmarks (< 30K nodes), where the original and ROM transfer functions could be directly compared; in the second one, we used large-scale benchmarks and compared EKSM to golden RaptorX™ROMs through S-parameter plotting. For the reduction process, target error and tol were set to 1e-2 and the number of frequencies l was set to 20. Experiments were performed on a Linux server with a 2.80 GHz 16-thread CPU and 64 GB of memory.

For the first experiment, the results are presented in Table 1, where the error refers to the max relative error between the transfer functions of the original models and ROMs, which is

Algorithm 2 Extended Krylov subspace method for low-rank solution of Lyapunov equations

Input: $\mathbf{G}_C \equiv \mathbf{C}^{-1}\mathbf{G}, \mathbf{B}_C \equiv \mathbf{C}^{-1}\mathbf{B}$ (or $\mathbf{G}_C^T, \mathbf{L}^T$)
Output: $\mathbf{Z}$ such that $\mathbf{P} \approx \mathbf{Z}\mathbf{Z}^T$
$j = 1$; $p = size\_col(\mathbf{B}_C)$
$\mathbf{K}^{(j)} = Orth([\mathbf{B}_C, \mathbf{G}_C^{-1}\mathbf{B}_C])$
while $j < maxiter$ **do**
  $\mathbf{A} = \mathbf{K}^{(j)T}\mathbf{G}_C\mathbf{K}^{(j)}$; $\mathbf{R} = \mathbf{K}^{(j)T}\mathbf{B}_C$
  Solve $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T = -\mathbf{R}\mathbf{R}^T$ for $\mathbf{X} \in \mathbb{R}^{2pj \times 2pj}$
  if converged **then**
    $[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = \mathbf{SVD}(\mathbf{X})$; $\mathbf{Z} = \mathbf{K}^{(j)}\mathbf{U}\mathbf{\Sigma}^{1/2}$
    break
  end if
  $k_1 = 2p(j-1)$; $k_2 = k_1 + p$; $k_3 = 2pj$
  $\mathbf{K}_1 = [\mathbf{G}_C\mathbf{K}^{(j)}(:, k_1+1 : k_2), \mathbf{G}_C^{-1}\mathbf{K}^{(j)}(:, k_2+1 : k_3)]$
  $\mathbf{K}_2 = Orth(\mathbf{K}_1)$ w.r.t. $\mathbf{K}^{(j)}$
  $\mathbf{K}_3 = Orth(\mathbf{K}_2)$
  $\mathbf{K}^{(j+1)} = [\mathbf{K}^{(j)}, \mathbf{K}_3]$
  $j = j + 1$
end while

Table 1: Detailed characteristics of RLCk models and experimental results of EKSM

Model	Order	# ports	# mutual induct.	$\|\|\mathbf{G}_C\|\|_F$	$cond(\mathbf{G}_C)$	ROM order	Error	Reduction time	Memory
PLL @ 28 GHz	1474	4	251680	7.03e+21	4.41e+16	96	1.21e-03	3.8 s	72 MB
Mixer @ 28 GHz	1498	10	79794	4.13e+23	1.81e+18	100	2.04e-04	2.09 s	58 MB
TI_DAC @ 28 GHz	3869	160	365494	4.16e+22	9.62e+15	1280	1.14e-06	3 min	1.18 GB
LNA @ 56 GHz	4274	6	1988882	1.73e+24	1.37e+19	204	7.98e-04	31 s	341 MB
LNA @ 28 GHz	6956	6	5360490	1.55e+23	3.15e+18	144	4.48e-03	30 s	594 MB
ILFM @ 14 GHz	15665	11	18394794	4.68e+27	9.02e+22	176	1.19e-03	1 min	1.54 GB
LNA @ 2.4 GHz	25602	6	72959220	6.23e+23	5.14e+19	144	1.42e-03	3 min	6.62 GB

calculated as $||\boldsymbol{H}(s) - \tilde{\boldsymbol{H}}(s)||_\infty / ||\boldsymbol{H}(s)||_\infty$ at the designated frequencies. For every benchmark, a base frequency of 100 MHz is chosen (f_{min} = 1e+8) and the maximum frequency is set to twice the resonance frequency of each circuit (e.g., f_{max} = 56e+9 for PLL@28GHz). As can be seen, EKSM generates accurate and compact ROMs across every type of benchmark with a maximum error below 0.14%. Additionally, the convergence criterion effectively strikes a balance between error and final ROM size while being computationally efficient. This is also visible through the performance results, where the reduction time remains below 3 min for benchmarks with less than 30K nodes and the memory requirements are not significantly high.

For the second experiment, the results are demonstrated in Table 2. The S-parameters plots of Figure 1 indicate that EKSM achieves accuracy close to that of RaptorX™ while producing on average ×13.2 more compact ROMs. Although EKSM has higher reduction time and memory requirements, they are still reasonable and can be significantly improved in futurk work.

Table 2: ROM order and MOR performance of EKSM vs RaptorX™

Model	Initial order	#ports	#mutual induct.	ROM order		Reduction time		Memory (GB)	
				RaptorX™	EKSM	RaptorX™	EKSM	RaptorX™	EKSM
VGA @ 28 GHz	95189	13	126766838	4744	286	1 min	11 min	32.63	19.14
HC @ 56 GHz	98024	5	165802476	1267	120	2 min	8 min	24.05	25.05
WPD @ 56 GHz	100888	4	193641938	765	40	3 min	4 min	24.79	29.76
VCO @ 13 GHz	104367	4	188436057	407	56	2 min	4 min	26.48	28.96
LNAC @ 56 GHz	128574	9	169339965	2172	378	1 min	19 min	25.82	26.01
WPD @ 28 GHz	129087	4	259462454	885	40	3 min	5 min	25.35	34.57
HC @ 28 GHz	134710	5	264162513	787	90	4 min	8 min	24.31	35.62
LNAC @ 28 GHz	162881	11	323090671	4768	308	6 min	27 min	78.52	49.58

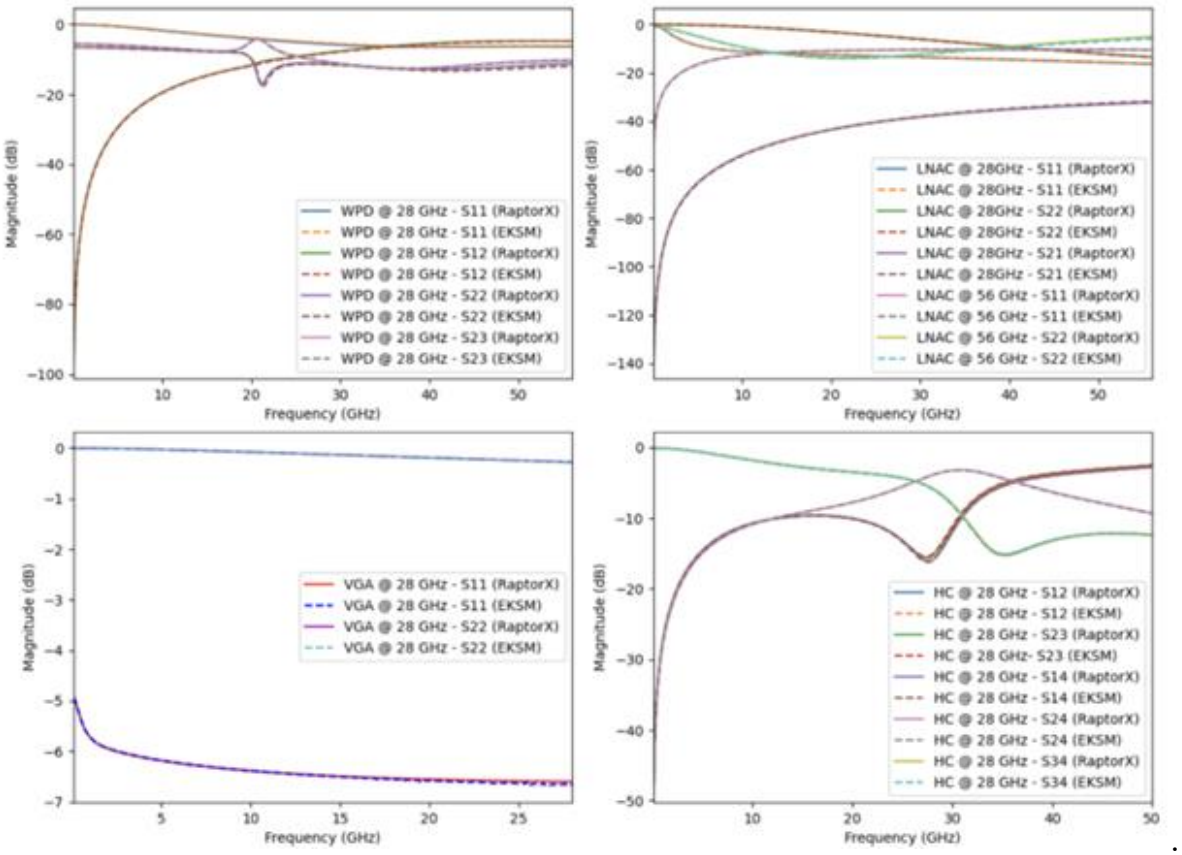


Figure 1: Comparison of accuracy between EKSM and RaptorX™ ROMs.

5 Conclusions

An alternative MOR technique for accurately reducing large-scale RLCk models is introduced. The proposed low-rank BT approach incorporates an iterative EKS projection method with a frequency-aware convergence criterion to produce accurate and compact ROMs. Experimental results demonstrate that our method provides up to ×22 smaller ROMs than ROMs obtained by ANSYS RaptorX™ for large-scale benchmarks with negligible deviations in S-parameters.

References

[1] A. Odabasioglu et al., "Prima: Passive reduced-order interconnect macromodeling algorithm," IEEE Trans. on CAD of Integrated Circuits and Systems, vol. 17, no. 8, pp. 645–654, 1998.

[2] S. Gugercin et al., "A survey of model reduction by balanced truncation and some new results," International Journal of Control, vol. 77, no. 8, pp. 748–766, 2004.

[3] C. Giamouzis et al., "Reduction of large-scale rlck models via low-rank balanced truncation," arXiv:2311.08478 [math.NA], 2023.

[4] "Ansys-RaptorX." [Online]. Available: www.ansys.com/products/semiconductors/ansys-raptorh

[5] C.-W. Ho et al., “The modified nodal approach to network analysis,” IEEE Transactions on Circuits and Systems, vol. 22, no. 6, pp. 504– 509, 1975.

[6] K. Gr¨ochenig, Foundations of Time-Frequency Analysis. Birkh¨auser, 2001.

[7] D. Lathauwer et al., “Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition,” SIAM Journal on Matrix Analysis and Applications, vol. 26, no. 2, pp. 295–327, 2004. [

[8] C. Chatzigeorgiou et al., “Exploiting Extended Krylov Subspace for the Reduction of Regular and Singular Circuit Models,” in Proc. of the 26th Asia South Pacific Design Automation Conference, pp. 773–778, 2021.

[9] E. Bavier et al., “Amesos2 and Belos: Direct and Iterative Solvers for Large Sparse Linear Sys tems,” Sci. Program., vol. 20, no. 3, p. 241–255, 2012.

[10] D. Garyfallou et al., “A Combinatorial Multigrid Preconditioned Iterative Method for Large Scale Circuit Simulation on GPUs,” in Proc. of the 15th International Conference on Synthesis, Model ing, Analysis and Simulation Methods and Applications to Circuit Design, pp. 209–212, 2018.

[11] C. Giamouzis et al., “Low-rank balanced truncation of rlck models via frequency-aware rational krylov-based projection,” in Proc. of the 20th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design, 2024

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 154 – 159

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

An Efficient Design Methodology for RF and mmWave Voltage – Controlled Oscillators

A.Chatzis, A. Michailidis, V. Gogolou and T. Noulis

Electronics Laboratory, Physics Department, Aristotle University of Thessaloniki Thessaloniki Greece

alekchat@physics.auth.gr, anamicha@physics.auth.gr, vgogolou@physics.auth.gr, tnoul@physics.auth.gr

Abstract

This paper presents an analytical design methodology approach for Voltage – Controlled Oscillators (VCOs). The selected topology is a cross coupled complementary LC – VCO, with an oscillation frequency of 5 GHz, designed using a 65nm RF CMOS process node. A detailed design methodology flowchart is presented to ease the work of the designer while, the proposed methodology is applied accordingly as to produce an LC – VCO while achieving a tuning range of 850 MHz, low phase noise performance of -106 dBc/Hz at 1 MHz offset from carrier frequency and low power consumption of 0.9 mW.

Keywords: Voltage Controlled Oscillator, LC – VCO, design methodology

1 Introduction

Voltage Controlled Oscillators are one of the most important building blocks in wireless communication systems, since they provide a steady periodical signal wherever is demanded. To ease the work of the designer, an analytical design methodology approach is presented, which is focused mainly on the challenges of the LC tank characterization and the transistor sizing optimization.

In this paper, the methodology is developed using a complementary LC – VCO, tuned at 5 GHz, since this topology is widely preferred due to its low power consumption, low phase noise and easier start – up [1 - 4]. The combination of the NMOS and PMOS devices gives a higher transconductance [1], but more importantly produces twice the voltage swing for a given bias current and inductor design. Another advantage of this topology is that it naturally provides an output common mode (CM) level approximately equal to $V_{DD}/2$ [2]

The core of the proposed methodology is the characterization of the LC tank which provides the needed transistor sizing according to the desired frequency of oscillation. From the parallel resistance of the tank, the required transistors' *gm* value is extracted [2], [5], [6]. Knowing the exact value of the transconductance needed, the minimum size of the transistors can be considered thus making this method favorable for transistor sizing optimization.

2 The proposed Design Methodology

It is well known that the core of the LC Voltage Controlled Oscillator is the LC tank, which is a lossy resonator. For that reason, an active circuitry must replenish the energy lost in each period as to sustain the oscillation [2]. The lossy LC tank can be modeled as a lossless tank with ideal components (capacitor and inductor) and a parallel resistance Rp which drains some of the capacitor energy in every cycle. If this lossy LC tank is connected with an active circuit which exhibits an input resistance of $-Rp$, the lost energy is replenished and the oscillation can be sustained [5]. The topology that was used for this study is the cross-coupled LC VCO, as depicted in Fig. 1 (a).

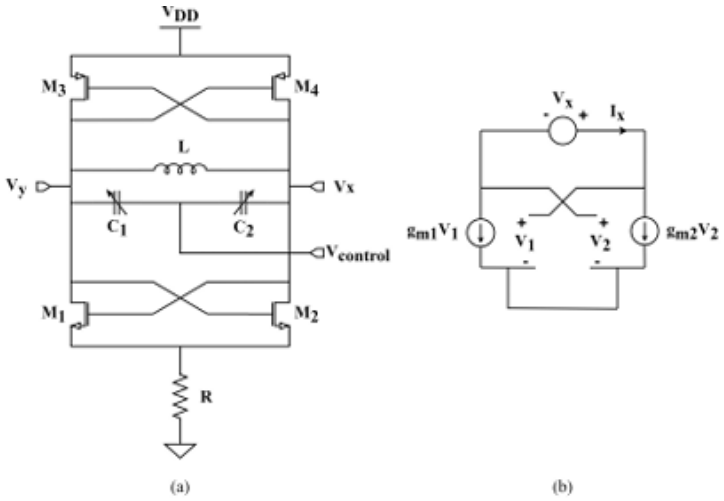


Fig. 1: (a) Cross – coupled LC VCO, (b) negative resistance small-signal equivalent circuit.

From the equivalent small-signal circuit depicted in Fig. 1 (b) it follows that

$$\frac{V_x}{I_x} = -\left(\frac{1}{gm_1} + \frac{1}{gm_2}\right) \tag{1}$$

And for $g_{m1} = g_{m2} = g_m$, it reduces to

$$\frac{V_x}{I_x} = -\left(\frac{2}{g_m}\right) \tag{2}$$

The negative resistance must cancel the loss of the tank, hence

$$\left(\frac{2}{g_m}\right) \leq 2R_p \tag{3}$$

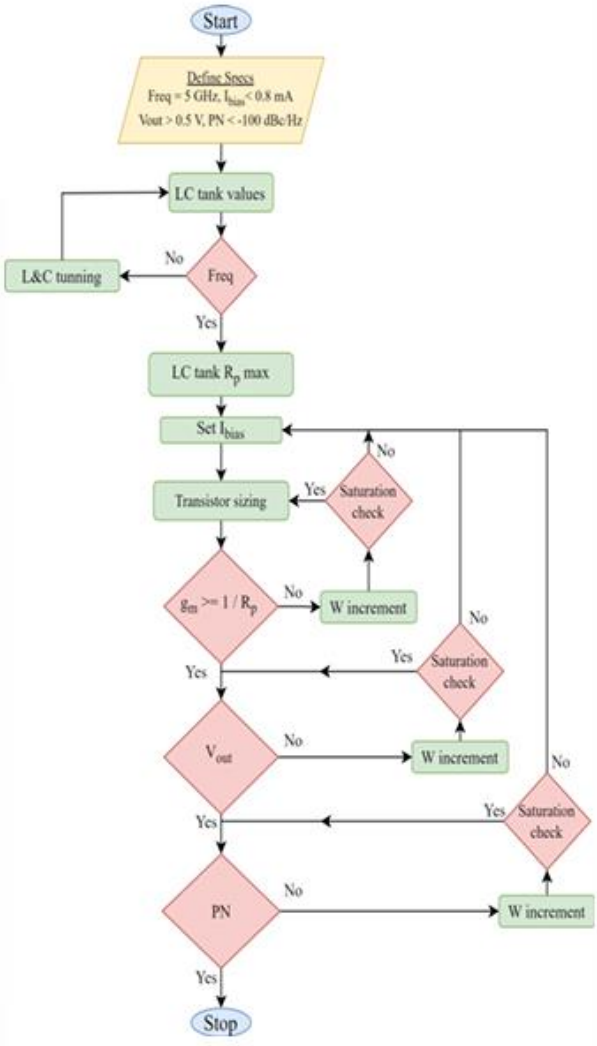
$$g_m R_p \geq 1 \tag{4}$$

In the initial step of the LC VCO design, a high – Q factor inductor must be chosen as it is the dominant parameter that affects the oscillator's phase noise [3]. The capacitance is implemented with MOS varactors, so they can contribute to the frequency tuning of the VCO. Parasitic capacitances of the cross-coupled transistor pair add up to create the total capacitance of the LC tank [1], [2], [7].

The condition that needs to be met for a guaranteed oscillation states that the resistance Rp of the tank must be equal to or greater than the inverse of the transistor's transconductance [6]. Using the LC tank as a single-port network [2], the Z parameters of the network can be extracted, hence the resistance Rp of the tank. Specifically, the proposed methodology is based on the Z_{11} parameter, as

it describes the ratio of the flowing current into the tank to the voltage across the tank. In other words $Z_{11}$is the impedance of the tank which for the resonance frequency becomes resistive and the imaginary part vanishes.

Turning back to the LC tank values, the designer's task is the tuning of the LC tank according to the desired frequency of oscillation and, exploiting Z parameter analysis, obtaining the impedance of the tank and the maximum value (Rp) at the resonant. The LC tank, designed for the verification of the proposed methodology, was tuned to 5 GHz and had an Rp resistance equal to 320 Ohms. From (4), the minimum gm value for which oscillation can occur can be calculated as $gm \geq 3.125\ mS$. Below, a design methodology flowchart is provided with a detailed explanation, along with a pseudocode that describes it.



VCO Design Methodology Algorithm

```
Define VCO Specifications and parameters:
    Targeted frequency (freq) → f0
    Bias Current → Ibias
    Output Amplitude (Vout) → V0
    Phase Noise (PN) → PN0
Define max L (Lmax) (PCELL limit)
Define min L (Lmin) (PCELL limit)
Define max varactor Cmax (arbitrarily selected)
Define min varactor Cmin (arbitrarily selected)
Define width (W) of nMOS and pMOS for equal gm

    for (L = Lmin : Lmax) loop{
      for (C = Cmin : Cmax) loop{
        if(1/(2π√LC) = f0)
          {return L and C values;
          break;}}}
// Z parameters analysis
    return Rp = max(Z11) at f = f0;

Ibias_LOOP:
    define Ibias;
    while(TRUE){
      if(gm < 1/Rp){
        W increment;
        if(saturation_check = FALSE)
          goto Ibias_LOOP;
      }
      else{break;}
    }
    while(TRUE){
      if(Vout < V0){
        W increment;
        if(saturation_check = FALSE)
          goto Ibias_LOOP;
      }
      else{break;}
    }
    while(TRUE){
      if(PN > PN0){
        W increment;
        if(saturation_check = FALSE)
          goto Ibias_LOOP;
      }
      else{break;}
    }
    return L, C, Ibias, W; end;
```

Fig. 2: Design methodology flowchart with the equivalent pseudocode.

It is obvious that with the minimum gm value the VCO cannot meet strict specifications of phase noise, output voltage but also the oscillator will present startup delay. To further improve the design, the methodology flowchart introduces some optimization loops for every important parameter. The

most dominant parameter that has a big influence in both output amplitude and phase noise is the bias current. With fixed bias current the degrees of freedom are reduced, hence the optimization loops focus only on transistor sizing. If any of the specifications cannot be met, while the transistors are in saturation, the algorithm changes the bias current. It is important to note that the width "W" increment referred in the algorithm, is a combinational width increment of both NMOS and PMOS transistors. The key point here is to keep the gm of both NMOS and PMOS transistors the same, that is keeping the ratio Wn/Wp about equal to 1/3.

3 Simulation Results

Using the proposed methodology, an LC – VCO, tuned at 5 GHz, was designed using a 65 nm RF CMOS process node. The schematic representation of the design VCO is illustrated in Fig. 1 (a). The proposed methodology was applied accordingly as to keep the power consumption of the circuit at 0.9 mW using a V_{DD} = 1.2 V supply. In Table 1, the parameters are presented as extracted from the proposed methodology. The spectrum and phase noise of the oscillator are depicted in Fig. 3, while in Table 2 a performance summary, derived from post-layout simulations, is presented. The phase noise of the design, extracted using the proposed methodology, is at adequate levels reaching -106.2 dBc/Hz while maintaining low power consumption. The tuning range is at 850 MHz, having a 17% deviation from the frequency of 5 GHz. Moreover, the circuit layout is also provided in Fig. 4.

Parameter	Value
Frequency	5 GHz
L (nH)	1.3
$C_1 = C_2 = C$ (pF)	1.5
M_1, M_2 (W/L) (μm)	14/0.08
M_3, M_4 (W/L) (μm)	42/0.08
R (Ohm)	600

Table 1: LC-VCO topology parameters.

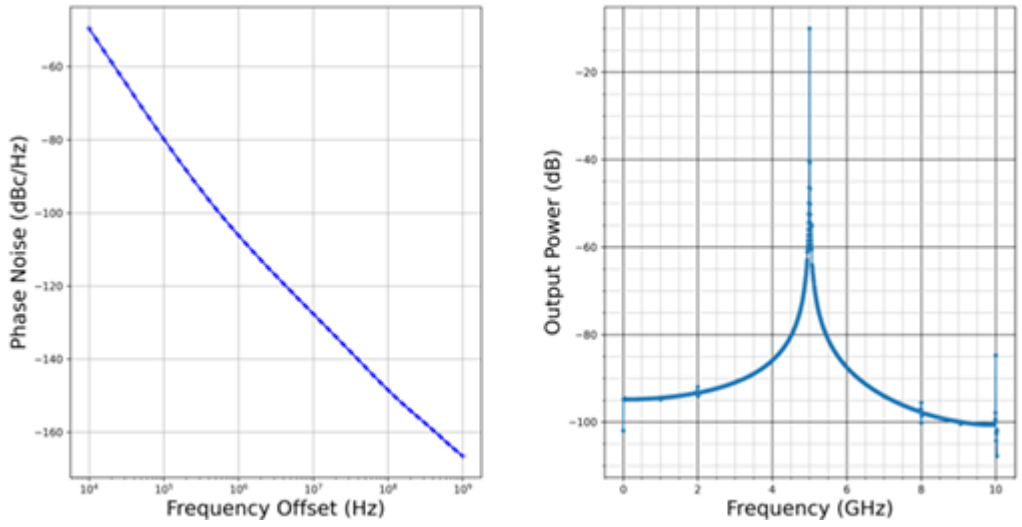


Fig. 3: (a) Simulated Phase Noise, (b) Simulated Output Spectrum.

Parameter	Value
Power Consumption (mW)	0.9
Output Swing (mV)	635
Phase Noise @ 1 MHz (dBc/Hz)	-106.2
Tuning Range (GHz)	4.75 – 5.60 (850 MHz span)
Silicon Area (mm^2)	0.12

Table 2: VCO performance summary extracted from post-layout simulations.

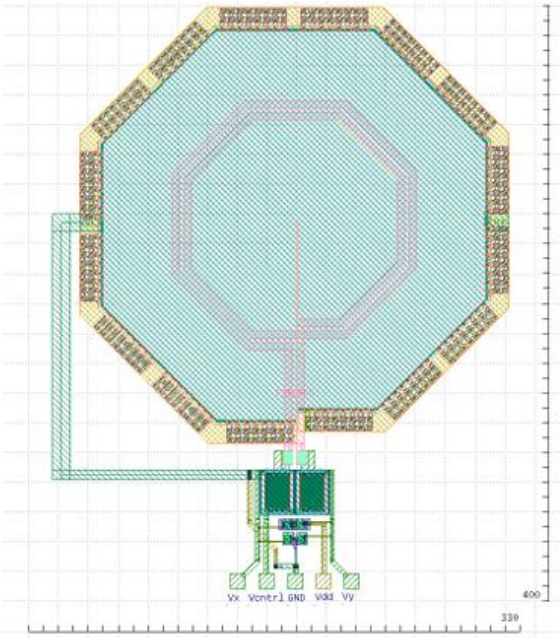


Fig. 4: Voltage Controlled Oscillator layout.

4 Conclusion and Future Work

In this paper, a design methodology for RF/mmWave VCOs is presented, mainly focused on the characterization of the LC tank and optimization loops for achieving the desired specifications dictated by the respected application. Using the proposed methodology, a 5 GHz LC – VCO was designed, using a 65 nm RF CMOS process. The developed VCO achieved a tuning range of 850 MHz, a Phase Noise performance of −106.2 dBc/Hz at 1 MHz offset from the carrier while keeping the power consumption at 0.9 mW. Future steps include a VCO or a DCO (Digitally Controlled- Oscillator) as the test – case vehicle in order to enable an automated design methodology to extremely reduce the respective design cycle time of such complex and frequently – used circuits. This future work implementation will be approached using the same design methodology, proposed in this paper, while exploiting Look-Up Tables (LUTs) for the well-defined parameters as to avoid multiple loops. The targeted design parameters will address both the carrier frequency and the phase noise specification. The focus will be paid on the accuracy improvement and acceleration as to enable a high range of queries in a time-conservative way. Enhanced interpolation techniques can be used versus the respective enlargement of the LUT grid spacing.

5 Acknowledgement

“This work was supported by the DAAD-Projekt 57647733 Radiation-Hard Integrated Circuits Educational platform ”RADHARD”, with funds from the Federal Foreign Office and German Academic Exchange Service (Deutscher Akademischer Austauschdienst) in the framework of the ”Hochschulpartnerschaften mit Griechenland 2023 2025” action.”

References

[1] R. Fiorelli, E. J. Peralias and F. Silveira, "LC-VCO Design Optimization Methodology Based on the gm/ID Ratio for Nanometer CMOS Technologies," in IEEE Transactions on Microwave Theory and Techniques, vol. 59, no. 7, pp. 1822-1831, July 2011, doi: 10.1109/TMTT.2011.2132735

[2] B. Razavi, “RF Microelectronics,” 2nd edition, Prentice Hall, 2011

[3] L. Yadi, W. Zhigong and T. Lu, "A 1.9–3 GHz broadband LC VCO with low phase noise for wireless communications," 2016 International Conference on Integrated Circuits and Microsystems (ICICM), Chengdu, China, 2016, pp. 108-113, doi: 10.1109/ICAM.2016.7813573

[4] Y. Jin, J. Bae and C. Nguyen, "A 4.6–5.9 GHz fully integrated 0.25-μm CMOS complementary LC VCO with buffer," 2016 21st International Conference on Microwave, Radar and Wireless Communications (MIKON), Krakow, Poland, 2016, pp. 1-3, doi: 10.1109/MIKON.2016.7492051

[5] L. Bin, F. Xianging and W. Zhigong, “A wideband LC-VCO with small VCO gain variation and adaptive power control,” Journal of Semiconductors, 2012, 33(10), doi: 101088/1674-4926/33/10/105008

[6] W. -C. Lai, J. -F. Huang, P. -G. Yang and W. -T. Lay, "Chip design of a high performance LC-VCO and mixer charge-injection signal processing for WiMAX communication and internet application," 2014 12th International Conference on Signal Processing (ICSP), Hangzhou, China, 2014, pp. 1713-1716, doi: 10.1109/ICOSP.2014.7015287

[7] E. B. Ortega-Rosales, F. Sandoval-Ibarra and E. Becerra-Alvarez, "A silicon-based 2.4GHz fully-differential LC-VCO: A design methodology proposal," 2015 International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), Istanbul, Turkey, 2015, pp. 1-4, doi: 10.1109/SMACD.2015.7301688

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 160 – 165

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Automated Low-Noise Amplifier Design Methodology for Wireless Sensor Nodes

A.P. Tsimpou, A. Michailidis, V. Gogolou and T. Noulis

Electronics Laboratory, Physics Department, Aristotle University of Thessaloniki Thessaloniki Greece

ptsimpou@physics.auth.gr, anamicha@physics.auth.gr, vgogolou@physics.auth.gr, tnoul@physics.auth.gr

Abstract

This paper outlines a detailed design methodology for Low-Noise Amplifiers (LNAs), using a 65nm CMOS process node. The main target of this work is to simplify and accelerate the LNA design cycle without compromising on performance. The provided simulation results confirm the efficacy of the proposed methodology, deriving an LNA with S11<-30dB, S21>15dB, S12<-30dB, S22<-10dB, NF<3dB, at the specified frequency of 5 GHz.

1 Introduction

High-speed demands of mobile communication applications require high performance from the sub blocks of transceivers. Since the Low-Noise Amplifier (LNA) is usually the first block of the receiver chain, it dominates the Noise Figure (NF) of the system, hence it has the most impact on the overall sensitivity. LNAs can be found in receiver blocks of mobile communication applications SoC [1-2] and they are important to suppress the noise in the receiver and maintain an adequate signal-to-noise ratio. This attribute is derived from the input impedance matching of the LNA at the desired frequency of interest, which is a crucial specification with respect to the targeted application field [3].

The frequency that the LNA input impedance matches the source impedance (antenna) is the driving factor of the design flow and leads to specific design parameters of the LNA topology using the same node, e.g., the sizing of the input transistor and the integrated spiral inductor values [3]. Designing an LNA involves several critical challenges that impact its performance. These challenges include Noise Figure minimization and maximum power transfer, by achieving proper input and output impedance matching and high linearity to avoid distortion and ensuring stability across various frequencies and operating conditions [4].

2 LNA Analysis and the Proposed Design Methodology

2.1. Input Impedance Matching

The input impedance matching of an LNA is crucial for maximizing power transfer between the source and the amplifier while minimizing wave reflections back to the antenna. This is particularly important in Radio Frequency (RF) circuits, where any impedance mismatch can lead to significant power loss

and degrade the overall system performance. The characteristic impedance of an antenna, which is typically connected to the input of the LNA block, is 50 Ohm. Therefore, it is essential to match the input impedance Zin of the LNA to 50 Ohm. To obtain the input impedance matching criteria, small signal analysis was applied.

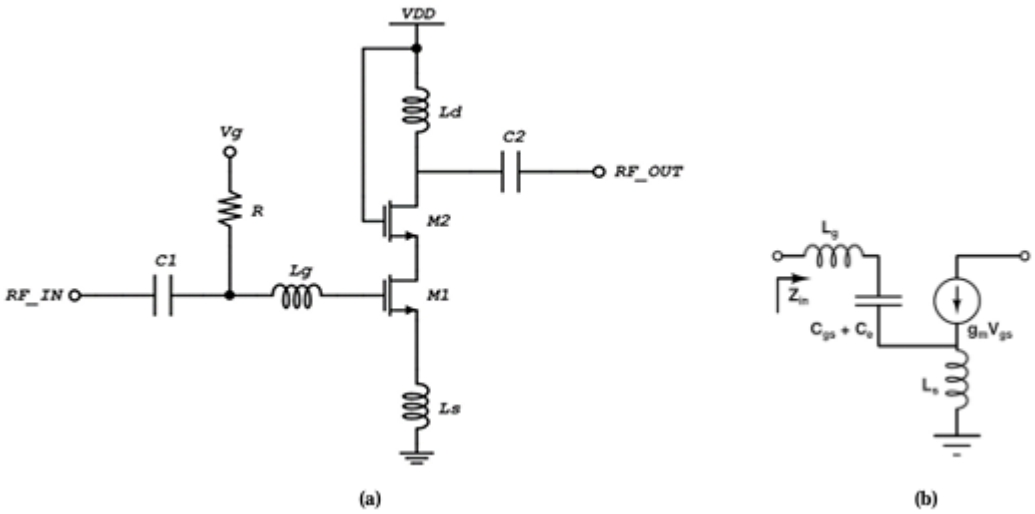


Figure 1: (a) Common Source LNA circuit topology (b) Small signal equivalent of the input.

The input impedance of the LNA with respect to the gate and degeneration inductors is extracted as follows [3],

$$Z_{in} = \frac{u_{in}}{i_{in}} = j\omega L_g + \frac{u_s + u_{gs}}{i_{in}}, \quad u_{gs} = \frac{i_{in}}{j\omega C_{gs}} \tag{1}$$

Where us is the source voltage and u_{gs} is the gate-source voltage of input transistor, while iin is the input current. The source current can be obtained as [3]:

$$i_s = i_{in} + g_m u_{gs} = i_{in}\left(1 + \frac{g_m}{j\omega C_{gs}}\right) \tag{2}$$

Where g_m is the transconductance of the input transistor M1. From the above equation the source voltage u_s is derived as [3]:

$$u_s = i_{in}\left(1 + \frac{g_m}{i\omega C_{gs}}\right) \cdot j\omega L_s \tag{3}$$

Combining (1), (2) and (3) we can get the following expression for the input impedance [3]:

$$Z_{in} = \frac{g_m \cdot L_s}{c_{gs}} + j\left[\omega(L_s + L_g) - \frac{1}{\omega C_{gs}}\right] \tag{4}$$

$$Re\{Z_{in}\} = \frac{g_m \cdot L_s}{c_{gs}} \tag{5}$$

To effectively match the input impedance of the LNA to 50 Ohm, the real part of Zin should be equal to 50 Ohm, while the imaginary part should be zero. Therefore, from (4) occurs the following equation

for the frequency [3]:

$$Im\{Z_{in}\} = 0 \implies \omega(L_s + L_g) - \frac{1}{\omega C_{gs}} = 0 \implies \omega = \sqrt{\frac{1}{C_{gs}(L_s+L_g)}} \quad (6)$$

2.2. Noise Figure (NF)

Noise Factor is defined as the input signal-to-noise ratio divided by the output signal-to-noise ratio. For an amplifier, it can also be interpreted as the amount of noise introduced by the amplifier seen at the output, besides that which is caused by the noise of the input signal [5].

$$NF = \frac{SNR_{in}}{SNR_{our}} \quad (7)$$

For ideal transmission the LNA does not introduce any noise ($SNR_{in} = SNR_{our}$) and NF=1. For real systems:

$$NF = 1 + \frac{Y}{2} + \frac{Z_{11}}{Z_{11}+\omega^2 L_s^2 g_m} \quad (8)$$

Where γ is the factor related to the channel thermal noise of the MOSFET, gm is the transconductance and Z_{11} is the input impedance of the LNA. Hence, the noise figure of the LNA also depends on how well the input impedance of the LNA matches the source impedance, thus achieving its minimum value for a specific input transistor sizing and biasing point.

2.3. Stability Factor (Kf)

The stability factor (K_f) is a parameter, particularly used in RF and microwave designs, to determine whether an amplifier will oscillate under specific conditions. The LNA must remain stable for all source impedances at all frequencies. If the LNA begins to oscillate at any frequency, it becomes highly non linear, and its gain is very heavily compromised. The Kf for the LNA is obtained as follows [6] ,

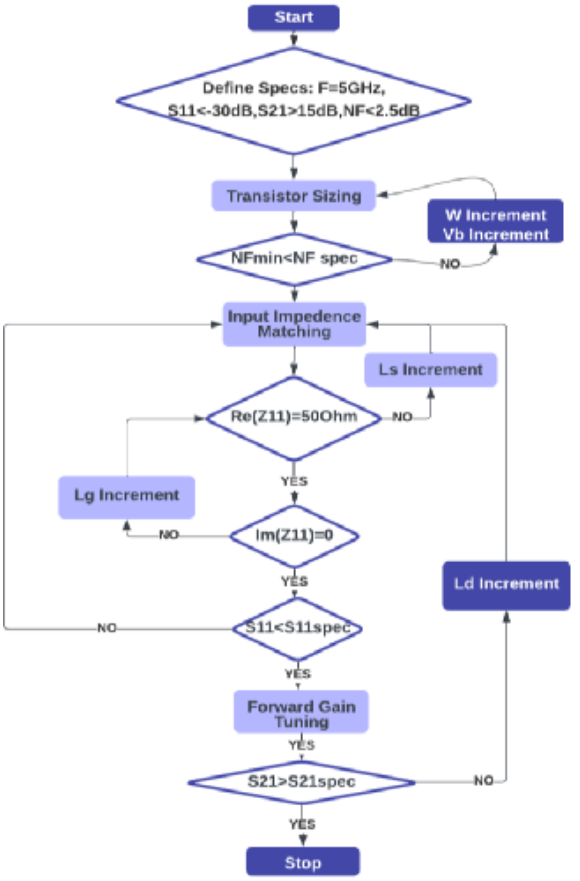
$$K_f = \frac{1-|S_{11}|^2-|S_{22}|^2+|\Delta|^2}{2|S_{12}||S_{21}|}, \quad \Delta = S_{11}S_{22} - S_{12}S_{21} \quad (9)$$

If $K_f > 1$, the amplifier is unconditionally stable, meaning it won't oscillate regardless of the source or load impedance.

2.4. LNA Design Methodology

The proposed design methodology is described via a flowchart illustrated in Figure 1. Also a pseudocode is provided to outline the design process. The workflow begins with defining the specifications for the LNA including the operating frequency (f), the input reflection coefficient (S_{11spec}), the forward gain (S_{21spec}) and the noise figure (NF_{spec}). The next step invloves transistor sizing: if the NF_{min} value is grater than NF_{spec}, there should be an increment at the transistor width (W) or the bias voltage Vb until the condition is satisfied. Once the NF specification is met, the process checks if the real part of the input impedance is equal to 50 Ohms. If not, the inductor Ls is incremented until Re(Z11) = 50. Then, the imaginary part of the input impedance is set to zero by increasing the value of inductor L_g. The input impedance matching procedures is repeated until $S_{11} \leq S_{11spec}$. The forward gain

tuning process follows: if $S_{21} < S_{21spec}$, the value of the inductor Ld should be increased and the input impedance should be rematched to 50 Ohms. Finally, the forward gain tuning process is repeated until the specified forward gain is achieved



Algorithm 1: LNA Design Methodology

1: **Define LNA Specifications and parameters:**
2: Frequency → f
3: Input impedance → S_{11spec}
4: Forward Gain → S_{21spec}
5: Noise Figure → NF_{spec}
6: **while** ($NF_{min} > NF_{spec}$)
7: {increase W; increase Vb;}
8: **if** ($NFmin \leq NF_{spec}$)
9: {**return** W and Vb values; **break**;}
10: **Input Matching Procedure:**
11: **while** ($Re(Z_{11}) \neq 50$)
12: {increase L_s;}
13: **if** ($Re(Z_{11}) = 50$)
14: {**return** L_s value; **break**;}
15: **while** ($Im(Z_{11}) \neq 0$)
16: {increase L_g;}
17: **if** ($Im(Z_{11}) = 0$)
18: {**return** L_g value; **break**;}
19: **if** ($S_{11} > S_{11spec}$)
20: {**goto** Input Matching Procedure;}
21: **else if** ($S_{21} < S_{21spec}$)
22: {increase L_d;
23: **goto** Input Matching Procedure;}
24: **else** {return L_d value;}
25: **end;**

Figure 2: Flowchart of the proposed design methodology with its respective pseudocode representation.

3 Methodology Verification and Simulation Results

The common source LNA topology that was designed using a 65nm CMOS process node is illustrated in figure 1(a). This LNA topology is commonly used in RF systems due to its ability to provide good input matching, gain and noise performance. The transistor M1 is the main amplifying device, and the RF signal is applied to its gate. The transistor M2 provides a stable bias current to M1, and it is used to improve the gain and the output impedance of the circuit, providing also isolation between input and output ports. Inductors Ls and Lg are essential for input impedance matching and the Ld is used as the output load and helps in maximizing the gain at the desired frequency. Finally, the capacitors C1, C2 are blocking the DC signal, and the resistor R is used for biasing the gate of M1, setting the operating point of the transistor. The device sizing, the supply and bias voltages and the inductor values, after applying the proposed design methodology at the targeted frequency of f = 5 GHz, are

shown in Table 1. The physical design of the LNA is illustrated in Figure 3. The S-parameters, noise figure and stability factor post-layout simulations of the circuit are shown in Figures 4(a) and 4(b) respectively. The post-layout simulation performance metrics of the LNA are summarized in Table 2. Using the proposed process the defined specifications were met:

Parameter	Value
Frequency	5 GHz
VDD	1.2 V
Vg	550 mV
W/L (M1)	275 um / 0.06 um
W/L (M2)	390 um / 0.08 um
Lg	3 nH
Ls	170 pH
Ld	2.4 pH

Table 1: Device sizing, supply and bias voltages and inductor values of the LNA design derived from the proposed methodology

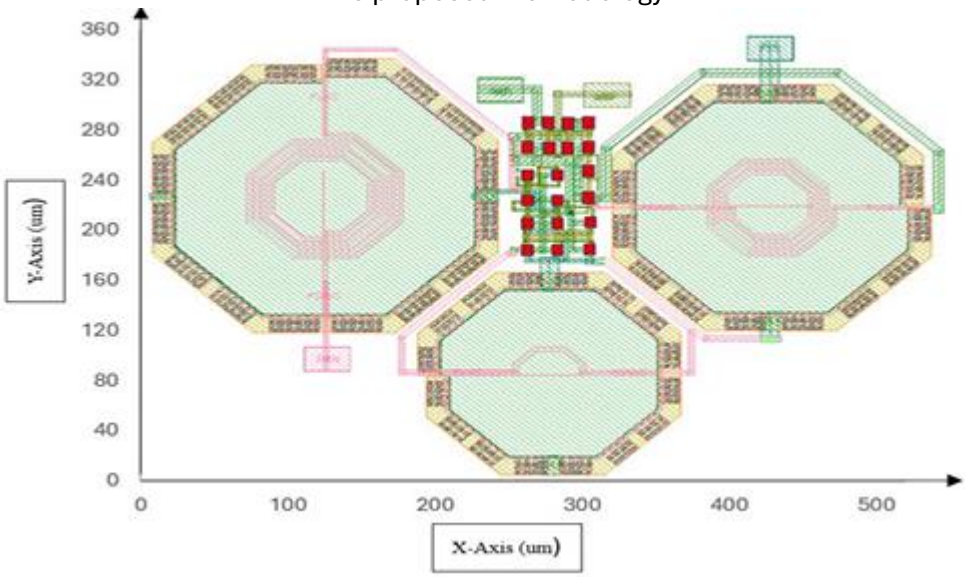


Figure 3: Physical Design of the designed LNA.

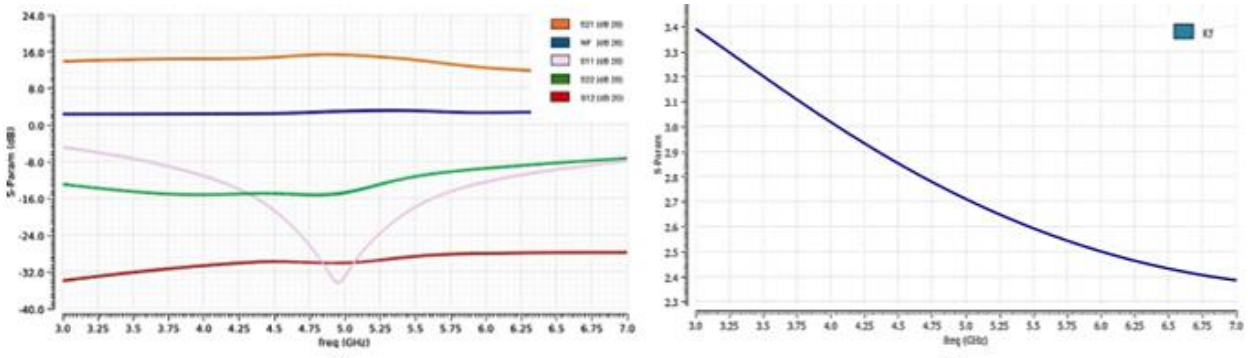


Figure 4: (a) S-Parameters and Noise Figure post-layout simulation (b)Stability Factor KF post-layout stimulation.

LNA Performance Metrics @ 5GHz	Value
S11	-32 dB
S21	15.2 dB
S12	-30.5 dB
S22	-15 dB
NF	2.6 dB
Power Consumption	22 mW
Silicon Area	0.19 mm2

Table 2: Post-layout simulation results of the designed LNA using the proposed methodology.

4 Conclusion

The common source LNA topology circuit, that was designed using the proposed methodology met successfully the given specifications, proving its efficiency while providing competitive LNA designs in an accelerated manner. The approach provides a structured framework for designers, making the design of LNAs more efficient and accessible without compromising on key performance metrics.

5 Acknowledgement

“This work was supported by the DAAD-Projekt 57647733 Radiation-Hard Integrated Circuits Educational platform ”RADHARD”, with funds from the Federal Foreign Office and German Academic Exchange Service (Deutscher Akademischer Austauschdienst) in the framework of the ”Hochschulpartnerschaften mit Griechenland 2023 2025” action.”

References

[1] A. Madan, M. McPartlin, C. Masse, W. Vaillancourt, and J. Cressler, "A 5 GHz 0.95 dB NF Highly Linear Cascode Floating-Body LNA in 180 nm SOI CMOS Technology," IEEE Microwave and Wireless Components Letters, vol. 22, 2012.

[2] C. Li, O. El-Aassar, A. Kumar, M. Boenke, and G. M. Rebeiz, "LNA Design with CMOS SOI Process 1.4dB NF K/Ka band LNA," in Proceedings of the 2018 IEEE/MTT-S International Microwave Symposium-IMS, Philadelphia, PA, USA, 10–15 June 2018, pp. 1484–1486.

[3] C. Sad, A. Michailidis, T. Noulis, and K. Siozios, "A Hybrid GA/ML-Based End-to-End Automated Methodology for Design Acceleration of Wireless Communications CMOS LNAs," Electronics, vol. 12, no. 2428, 2023. doi: 10.3390/electronics12112428.

[4] H. A. Eshghabadi and F. Eshghabadi, "Design Considerations for Low Noise Amplifier," in IEEE International Conference on Circuits and Systems (ICCAS), 2012.

[5] A. Dao, "Integrated LNA and Mixer Basics," National Semiconductor Application Note AN-884, April 1993.

[6] S. Kassim and F. Malek, "Microwave FET amplifier stability analysis using Geometrically-Derived Stability Factors," 2010 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, Malaysia, 2010, pp. 1-5, doi: 10.1109/ICIAS.2010.5716171.

Posters

Frequency and Procedure of Formative Assessment: Impact on Student Experience

Iris Gertner

Towards a vision system for olive fruit selection for enhanced olive oil production

Dimitrios Kosmopoulos, Erion-Vasilis Pikoulis and Kostas Blekos

Preliminary Hazard Analysis and Functional Safety Concept for a 2nd Life Battery Energy Storage System

Rigas Frangoudis, Dimitra Spanoudaki, Sotiris Athanasiou and Elisavet Elvanoglou

SAND5G - Security Assessments for Networks and Services in 5G Networks: From 5G to Edge

Kostas Lampropoulos, Kostas Pournaras, Christos Tranoris, Odysseas Koufopavlou, Spyros Denazis and Paris Kitsos

Nanotechnology processes with atomic-scale precision: Access to new tools and clean-room infrastructure for researchers and the Industry

Chloi Zormpa, Sotiris Mouchtouris, Vasilios Vamvakas and Evangelos Gogolides

PerCV.ai Platform: Leveraging Cloud and Edge Computing to Build Vision AI Solutions at Scale

Thomas Charisis, Dimitris Kastaniotis and Christos Theocharatos

Exploiting the triboelectric phenomena for energy harvesting and sensing

Christos Tsamis

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 166-167

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Frequency and Procedure of Formative Assessment: Correlation with Student Learning Experience

Iris Gertner Moryossef

School of Management Hadassah Academic College Israel
irisge@edu.hac.ac.il

Abstract

Formative Assessment (FA) is one of the learning tools in higher education that supports teaching and education in the classroom for the benefit of both teachers and students. This study examines the correlation between the frequency and procedure of formative assessment and Students' Learning Experience (SLA). The study was conducted by comparing the experience of students in their first-year classes at Hadassah Academic College in Jerusalem.

Class I studied based on the approach known as Assessment as Learning while Class II used Assessment for Learning. The assessment in the classes differed on two vectors: its frequency (up to two or more than two assessments during the semester) and its procedure—whether it included explanations and demonstrations by the lecturer.

Assessment for Learning (AfL), which was employed in Class I, engages both the teacher and students. This type of assessment involves formal and informal activities that are considered part of the learning and are also used to inform the planning of future learning. Assessment for Learning is part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning (Klenowski, 2009, p. 264).In contrast, Class II used Assessment of Learning (AoL) which in general was used to judge performance and measuring outcomes after a formal learning activity aimed for grades and evaluation. In AoL, the teacher administers the assessment and determines the students' progress, measuring their application of knowledge or skills against a standard.

In this study we examined the correlation between the frequency and procedure of the assessment and the SLA, determined via the parameters of enjoyment of learning, understanding of the material, satisfaction with the learning, and perceived usefulness.

Results

The results indicate high positive correlation between the frequency and procedure of the formative assessment and each of the four parameters of SLA.

Results show that the more the tutor explained, demonstrated, and illustrated the assessment, the more enjoyment, satisfaction, understanding, and perceived usefulness were reported by the students. The level of correlation between the procedure and SLA depended on the way the tutor

guided the students. The more the tutor illustrated the assessment, gave direct tutor, and managed students' expectations, the better the students understood the tasks.

Conclusion

The main conclusion drawn from this research refers to the important role of personal interaction between the service provider—i.e. the tutor, and the customer—i.e., the student. In Assessment for Learning (AfL), the tutor mentors the students and encourages them to think on their own, using the assessment for shaping the learning process. This kind of assessment enhances students' satisfaction, enjoyment, and understanding.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 168-170

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Towards a vision system for olive fruit selection for enhanced olive oil production

Dimitrios Kosmopoulos[1], Erion – Vasilis Pikoulis[1], Kostas Blekos[1]

[1]*Computer Engineering and Informatics Department, University of Patras, Greece*
dkosmo@upatras.gr, pikoulis@ceid.upatras.gr, mplekos@upatras.gr

Abstract

The olive oil industry plays a key role in the global agricultural economy, with olive quality and ripeness directly affecting oil quality. Efficient sorting and classification of olives are essential for optimizing yield and quality. Here we present our research objectives and ongoing research regarding an automated system using computer vision, to classify olives by ripeness and quality. The system also handles foreign objects (e.g., leaves, twigs) that could otherwise cause misclassifications.

1 Introduction

Extra virgin olive oil (EVOO) is a key ingredient in the Mediterranean diet due to its taste and health benefits. Around 82% of the world's EVOO production comes from European Mediterranean countries, where the climate is ideal for olive cultivation. EVOO is produced through mechanical processes such as washing, decantation, centrifugation, and filtration. Its quality depends on the olive fruit's condition, which is influenced by ripeness, climate, and agricultural practices. Ensuring the health and quality of the olives is crucial for high-quality EVOO. Modern sorting machines using real-time image analysis allow for fast, objective selection of olives, a growing yet promising field in EVOO production.

The simplest methods rely on color segmentation to classify olive fruit. For instance, (Violino et al., 2019) used a color-based classifier to sort the olives into green and black categories. Their analysis revealed significant chemical differences between the oils produced from each category, as well as differences identified by expert panel evaluations. DNN-based classifiers and object detectors have also been applied to olive fruit classification. (Figorilli et al., 2022) used a CNN with AlexNet as a backbone to classify olives into five quality classes using RGB images. (Saedi et al., 2023) employed transfer learning with a pre-trained Xception model to classify olives into five ripening stages. (Hayajneh et al., 2023) proposed a custom CNN for resource-constrained devices, classifying three olive varieties and two quality classes. For object detection, the YOLO has been used to quickly classify olive lots for oil production and table olives, as seen in the work by (Salvucci et al., 2022) using YOLO v7 for counting.

2 Objectives and proposed work

The goal is to develop a low-cost portable system, able to be deployed easily at a very short time and that will be able to separate olives according to color or quality very fast. The separation will be done

using visual information, while the separation will be done using an actuator. The machine features a conveyor belt with specialized scoops, each carrying a single olive at a time (see Fig. 1 for a depiction of the sorting machine under development). An Arduino camera and LED lighting setup is being installed on the conveyor belt, while the final prototype will include an electronic control unit, a processing unit, and servo-based actuators (one per olive compartment) to separate the olives based on the automated quality control system's decisions using a lightweight YOLO architecture. During operation, the camera captures each scoop as it passes, and the image is processed by a DNN-based object detector to classify the contents according to color or according to their quality status. The actuators will then be configured to allow healthy olives to proceed for further processing, while damaged ones will be redirected to a separate container.

The 3-class detector demonstrates a generally satisfactory performance reaching an accuracy level of 86% for the class of “good” olives, while falling slightly behind on the “bad” (olive) and “leaf” classes, with an accuracy of 77% and 71%, respectively.

In our next steps we will implement the full processing chain to allow processing of large numbers of fruits, so that we can have more data and evaluate the system in a lab environment.

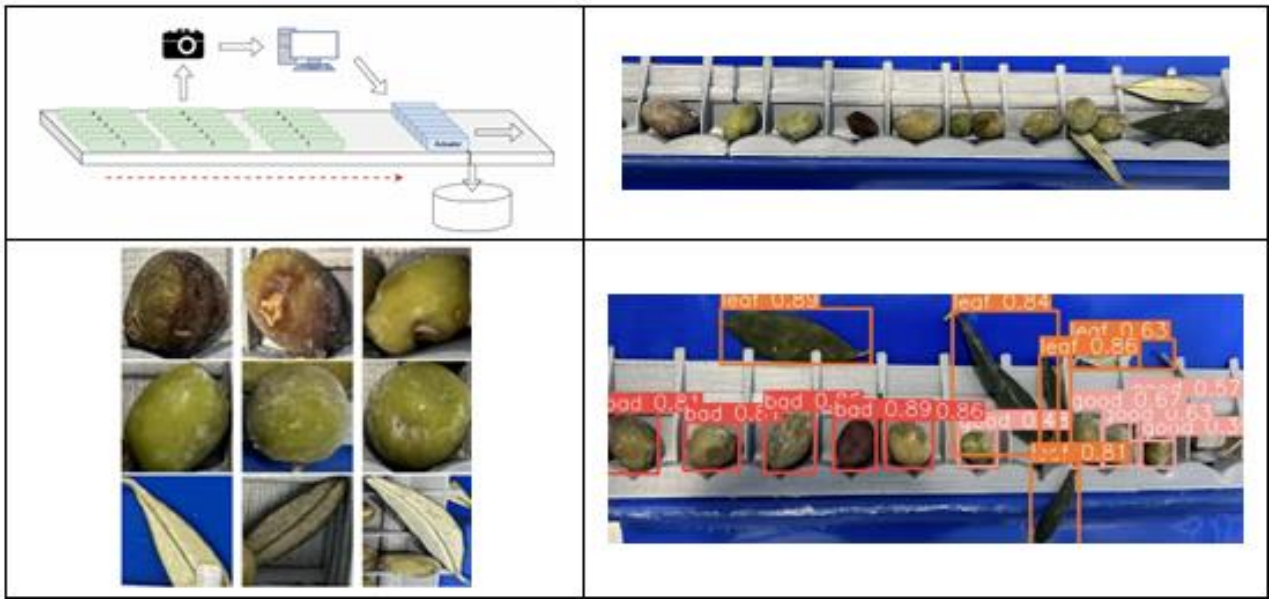


Figure 1: System configuration, example of heathy/damaged olive fruit and random leaves, olive scoops and classification results

Acknowledgement: This research was funded by the Greek Ministry of Agricultural Development and Food, under Sub-Measure 16.1 – 16.2 “Establishment and Operation of Business Groups of the European Innovation Partnership for Productivity and the Sustainability of Agriculture”, Action 2, project number M16SYN2-00229.

References

Figorilli, et al., (2022), “Olive fruit selection through ai algorithms and rgb imaging,” Foods, 11(21), 3391.

Hayajneh, A. M., Batayneh, S., Alzoubi, E., and Alwedyan, M., (2023) , "Tinyml olive fruit variety classification by means of convolutional neural networks on iot edge devices," AgriEngineering 5(4), 2266–2283

Saedi, S. I. and Rezaei, M., (2023) "A modified xception deep learning model for automatic sorting of olives based on ripening stages," Inventions 9(1), 6.

Salvucci, et al., (2022), "Fast olive quality assessment through rgb images and advanced convolutional neural network modeling," European Food Research and Technology 248(5), 1395–1405

Violino, S., et al. (2019), "Are the innovative electronic labels for extra virgin olive oil sustainable, traceable, and accepted by consumers?," Foods 8(11), 529

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 171-176

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

Preliminary Hazard Analysis and Functional Safety Concept for a 2nd Life Battery Energy Storage System

Rigas Frangoudis[1], Dimitra Spanoudaki[1] , Elisavet Elvanoglou[1] and Sotiris Athanasiou[2]

[1] *Sunlight Group Energy Storage Systems, 67200, Xanthi, Greece*
[2] *Sunlight Group Energy Storage Systems,14564, Attica, Greece*
d.spanoudaki@sunlight.gr

Abstract

The anticipated increase in retired electric vehicle (EV) batteries by 2030 offers both challenges and opportunities for their reuse in stationary energy storage systems (ESS). This paper presents a comprehensive Preliminary Hazard Analysis (PHA) and Functional Safety Concept (FSC) for a 2nd life Battery Energy Storage System (BESS). The analysis focuses on both theoretical and practical aspects, identifying key risks and mitigation strategies for the system. Particular attention is given to system analysis, which incorporates Electrochemical Impedance Spectroscopy (EIS) and a wireless Battery Management System (BMS) to monitor and safeguard against potential hazards. The paper provides a detailed breakdown of potential hazards, discusses the implementation of safeguards, and ensures compliance with international safety standards. The work includes both qualitative and quantitative evaluations, providing a basis for further pilot implementations.

Acronym	Definition
BS	Battery System
BMS	Battery Management System
EV	Electric Vehicle
EIS	Electrochemical Impedance Spectroscopy
PHA	Preliminary Hazard Analysis
FSC	Functional Safety Concept
ESS	Energy Storage Systems
BESS	Battery Energy Storage System

1 Introduction

The reuse of electric vehicle (EV) batteries for stationary applications is becoming a vital consideration as the volume of retired batteries grows significantly. By 2030, it is estimated that more than 5 million tons of EV batteries will no longer be suitable for mobility purposes but will still retain 70-80% of their original capacity. These "2nd life" batteries can be repurposed for energy storage

systems, providing an environmentally friendly alternative to disposal. This paper presents a Preliminary Hazard Analysis (PHA) and a Functional Safety Concept (FSC) for a 2nd life Battery Energy Storage System (BS), which is intended for grid balancing purposes. We highlight the importance of managing the specific hazards associated with 2nd life BS and detail the necessary safety protocols, such as a Battery Management System (BMS) capable of managing both natural and technical risks.

2 Hazard Identification

2.1. Topography, Land-use and Natural Hazards

The main hazards associated with the function of a BS are the natural hazards, the lithium-ion cell hazards, other chemical hazards and the hazards associated with the BMS..

The 2nd life BS will be installed in Kantza, Attica at the Innovation Hub of the PPC, with geographical coordinates 37°59'37"N 23°51'08"E. The geographical coordinates are needed for the assessment of natural hazards that could lead the BS to a failure mode.

Figure 1 shows the earthquake hazard map of Greece and the expected ground shaking at the specific location of PPC hub due to future potential earthquakes that might occur locally or at a greater distance. Ground shaking is expressed as Peak Ground Acceleration (PGA), normally given in the percentage of "g", the Earth's gravitational acceleration.

From the seismic map, we observe that Greece is one of the most cosmogenic areas in Europe (Fig. 1A). Specifically, for the area of the PPC innovation hub, is coloured orange (moderate hazard) the acceleration is estimated at g = 0.229. This information is important for the stability analysis we are going to perform to the 2nd life BS in the designing phase.

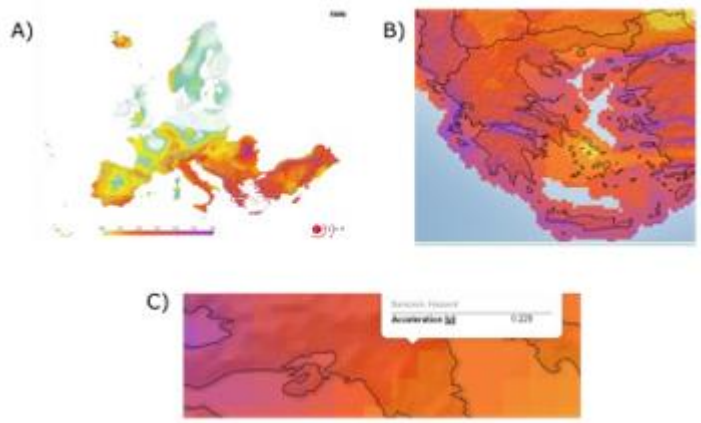


Figure 1: Earthquake hazard maps of: A) Europe, B) Greece, C) PPC's Innovation Hub in Attica

External temperatures are of a tantamount interest for designing and development of BS, due to the strict range of temperatures lithium-ion cells can function. In general, lithium-ion cells can function from -20 oC to 55 oC (discharging), and any further increase of the temperature could lead to thermal runaway phenomena, posing danger for the BS, humans close to the BS and other natural and human properties.

The average mean surface air annual temperature in Greece is 10 o C. The extreme temperatures in Greece are considered the summer months from June – August, where the average mean surface air

temperature in those months is 30 o C. The last years due to the climatic change conditions extreme permanent heatwaves are observed with maximum surface air temperatures at 43 – 44 oC in Attica specifically. This fact consequently leads to the investigation of an efficient cooling system in the 2nd life BS, and of an additional protective construction on the PPC site.

2.2. Lithium-ion Hazards and other chemical Hazards

The primary consideration for lithium-ion batteries is for adequate cooling and management of temperature excursions. In most cases, the temperature of lithium-ion batteries should not be higher than 60 o C to prevent thermal runaway, which causes propagation of increasing temperatures, pressures, and fire towards neighboring cells. Similarly, the temperature of the lithium-ion batteries in most cases should not drop below 0 o C to prevent lithium plating around the anode during charging which can cause internal short-circuits. Although the choice of anode – cathode chemistry can result in different thermal stability and volatility, all lithium-ion batteries are flammable when exposed to fire. Lithium – ion fires are a unique class of fire and may result in emissions of large volumes of toxic or combustible gases, which need to be managed accordingly.

Hazard Type	Hazard Description	Causes	Hazards
Chemical Hazards	Electrolyte Leakage	Leakage of lithium salts and organic solvents	Toxicity, flammability, corrosiveness
Chemical Hazards	Thermal Runaway	Overcharging, short circuits, external heat, mechanical damage	Release of flammable gases, explosion, fire
Chemical Hazards	Decomposition Products	Degradation of battery materials	Release of HF, CO, toxic gases

The chemical hazards associated with BS primarily revolve around the dangerous properties of the electrolyte and the hazardous products released during thermal events or material degradation. These hazards underscore the importance of robust safety measures, including proper monitoring, containment, and emergency response protocols, to mitigate the risks associated with chemical incidents in BS.

By integrating these preventive measures, detection systems, and response strategies, the risks associated with chemical hazards in BESS can be significantly reduced, ensuring safer operation and handling of large-scale energy storage systems.

2.3. Battery Management System Hazards

When repurposing 2nd life batteries for grid applications, ensuring robust technical safeguards in the Battery Management System (BMS) is crucial to manage the aged batteries safely and efficiently. BMS must monitor the volage, temperature and current of the 2nd life BS to take the necessary action to protect the battery, the humans, the environment and other property from a possible failure of the BS. However, there are several hazards associated with BMS that need to be carefully managed:

- **Overvoltage**: It is a primary BMS failure and can be result of the conductor's failure to open. The consequences from this failure are thermal runaway of the cells, battery swelling due to the breakdown of the electrolyte and the formation of gas inside the battery, reduced lifespan of the battery, leakage of the electrolyte, capacity loss.

- **Undervoltage**: The BMS controls the discharge process and when the battery voltage drops below a certain level (specified by the cell manufacturer), it prevents damage to the battery cells. When a BMS fails to control the undervoltage then the risk of internal short circuits increases, that leads to overheating or even fires.
- **Over temperature**: One of primary functions of the BMS is to monitor the temperature of the battery to shut down the BS to prevent overheating and potential damage. If a BMS fails to control the temperature of the BS, then a thermal runaway incident is unavoidable.
- **Under temperature**: If the BMS fails to monitor and control the temperature, then in case of low temperatures, lithium plating will occur in the cell increasing the risk of thermal runaway due to internal short circuits. Also, the lifespan of the battery decreases and its capacity as well.
- **Overcurrent**: The BMS can monitor and control the current flow to prevent overcurrent situations. The overcurrent can increase the risks of fires or explosions, especially if the BMS fails and damage connected to the BS devices.

The BMS is the primary safeguard for preventing hazards in 2nd life batteries. However, several BMS related risks exist, such as failure to monitor voltage, temperature, or current accurately. Overvoltage and undervoltage conditions can lead to thermal runaway or battery degradation, while overheating poses a direct fire risk. Therefore, the BMS must include redundant systems to monitor these parameters and take corrective action when needed.

3 Technical Safeguards and Functional Safety Concept

The Functional Safety Concept (FSC) of the 2nd life Battery Energy Storage System is centered around the BMS, which acts as the primary technical safeguard. The FSC ensures the reliable operation of the system by mitigating the specific hazards associated with 2nd life batteries.

3.1 Thermal Management: Thermal management is a critical function of the BMS, especially in high-temperature conditions like those found in Greece. The system must be equipped with redundant temperature sensors and an active cooling mechanism to prevent overheating. Failure to maintain appropriate thermal conditions can lead to thermal runaway, one of the most significant risks in battery systems.

3.2 Electrical Safeguards: The system must be protected against electrical failures such as short circuits and overcurrent. Electrical safeguards, including fuses and circuit breakers, ensure that any fault in the system is quickly isolated to prevent larger system failures.

3.3 Structural and Environmental Safeguards: In addition to internal electrical and thermal protections, the system must be designed to withstand environmental factors such as earthquakes and extreme temperatures. Structural reinforcements and housing will be designed to protect the system from seismic activity while preventing external heat from compromising the battery system's safety.

3.4 Wireless Communication Safeguards: Given the reliance on wireless communication for the Electrochemical Impedance Spectroscopy (EIS) used in estimating the battery's State-of-Charge and State-of-Temperature, robust communication protocols must be in place. Redundant communication channels ensure that even if one path fails, control over the system is maintained.

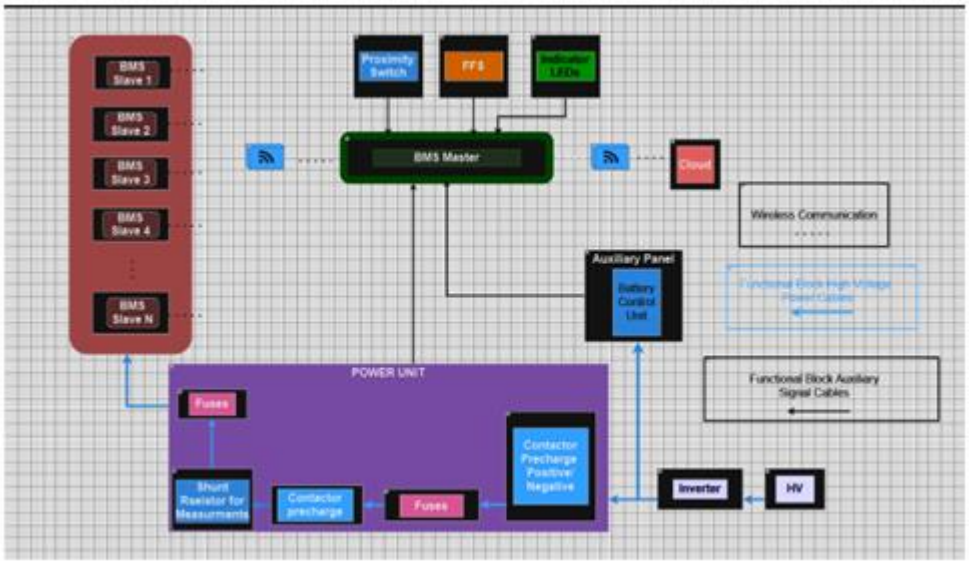


Figure 2: Preliminary Functional Safety Concept for the 2nd life BS.

BMS Slaves (BMS Slave 1, BMS Slave 2, etc.)	Function: BMS (Battery Management System) slaves are responsible for monitoring individual cells or groups of cells in a battery pack. They collect data such as voltage, temperature, and current, then send this information to the BMS Master.
BMS Master	Function: The BMS Master receives the data from the BMS slaves and manages the overall battery system. It ensures that the battery operates within safe limits, balancing cells and protecting them from overcharging or overdischarging.
Proximity Switch:	Function: This is likely used to detect the physical presence or position of components in the system, such as the connection of battery modules or other key components.
FFS(FireFighting System):	Function: The FFS may refer to a fire prevention or protection system, designed to mitigate fire risks, especially important in battery systems where thermal runaway is a concern.
Indicator LEDs:	Function: These LEDs provide visual feedback or status indicators for different parts of the system, such as system health, charging status, or fault conditions.
Shunt Resistor for Measurements:	Function: A shunt resistor measures the current flowing through the system. By measuring the voltage drop across the shunt, the system can calculate the current, helping in monitoring power flow and ensuring safe operation.
Contactor Precharge:	Function: Precharge contactors limit the initial inrush of current when the battery system is first connected to the rest of the circuit. This protects sensitive components, such as capacitors and inverters, from being damaged by a sudden surge of power.

4 System Analysis and Results:

The system analysis, as presented in Figure 2, shows the comprehensive approach to ensuring functional safety. The PHA (Preliminary Hazzard Analysis) and FSC (Functional Safety Concept) are designed to mitigate the primary hazards identified in Sections 2 and 3. However, a more detailed quantitative analysis is needed to fully validate the effectiveness of the proposed safety measures. Future work will focus on testing the system under real-world conditions through pilot projects to assess the feasibility of these safeguards.

Conclusion:

This paper presented the key hazards associated with 2nd life Battery Energy Storage Systems and outlined a Functional Safety Concept to mitigate those risks. The BMS, combined with structural and environmental safeguards, ensures that the system can operate reliably and safely in the face of natural and technical hazards. Further validation is required through pilot projects to refine these safety protocols and ensure the system's long-term viability.

Acknowledgement:

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor CINEA can be held responsible for them.

References

1 Hu, X., et al. "A Review of Second-Life Lithium-Ion Batteries for Stationary Energy Storage Applications." Proceedings of the IEEE, vol. 110, no. 6, 2022, pp. 1040-1059.

2 Hassan, A., et al. "A Review on Power Grid Applications, Degradation Mechanisms, and Power Electronics Interface Architectures." Batteries, vol. 9, no. 571, 2023.

3 European Seismic Hazard Model 2020 (ESHM2020). "Seismic Hazard Assessment for Europe." European Facilities for Earthquake Hazard and Risk. Available at: http://www.hazard.efehr.org/en/Documentation/specific-hazard-models/europe/eshm2020 overview/, accessed 2024.

4 World Bank Group. "Climate Change Knowledge Portal: Historical Climate Data for Greece." World Bank Climate Change Knowledge Portal. Available at https://climateknowledgeportal.worldbank.org/country/greece/climate-data-historical, accessed 2024.

5 See, K. W., et al. "Critical Review and Functional Safety of Battery Management Systems for Large-Scale Lithium-Ion Battery Pack Technologies." International Journal of Coal Science & Technology, vol. 9, 2022, pp. 36-50.

Co-funded by the European Union

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 177-178

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

SAND5G - Security Assessments for Networks and Services in 5G Networks: From 5G to Edge

Kostas Lampropoulos[1], Kostas Pournaras[2], Christos Tranoris[2], Odysseas Koufopavlou[1], Spyros Denazis[1], and Paris Kitsos[1, 3]

[1] University of Patras Patras, Greece ce, [2] p-net Patras, Greece, [3] Univ. of the Peloponnese Patras, Greece

klamprop@ece.upatras.gr, kpournaras@p-net.gr, ctranoris@p-net.gr, odysseas@upatras.gr, sdena@upatras.gr, pkitsos@upatras.gr

Abstract

In the framework of SAND5G, funded by Digital Europe Programme (DIGITAL) under the Cybersecurity and Trust (DIGITAL-ECCC-2022-CYBER-03) call, an efficient architecture proposed and integrated to deliver a risk assessment platform to help 5G stakeholders secure their systems, enable national authorities to oversee security measures, and align with European cybersecurity policies and proposed EU toolbox for 5G security. Since 5G can support a millions of devices which means the transferred data will be extended compared to old-fashioned networks, 5G is significantly more vulnerable to cyberattacks. On the other hand, because 5G increases the speed the data travels at, the edge computing reduces the distance it travels before it is processed. So, in this poster a security extension of 5G at the network edge is proposed.

1. Architecture of SAND5G

The SAND5G project seeks to create a platform capable of producing both technical and financial risk assessments for stakeholders in the 5G ecosystem, ensuring their infrastructures and services align with national cybersecurity strategies and telecommunications security frameworks. Furthermore, the platform will offer continuous monitoring of 5G systems and integrate automated Incident Response (SOAR - Security Operations and Automated Response) capabilities for proactive threat defense. Figu re 1 presents the proposed architecture, showcasing its main components and the services it provides (Bantouna A. et all, Security Assessments for Networks and Services in 5G Networks. 2024).

This article describes work undertaken in the context of the SAND5G project, "Security Assessments for Networks anD services in 5G" which has received funding from the European Union's Digital Europe programme under grant agreement No 101127979 and is supported by European Cybersecurity Competence Center. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

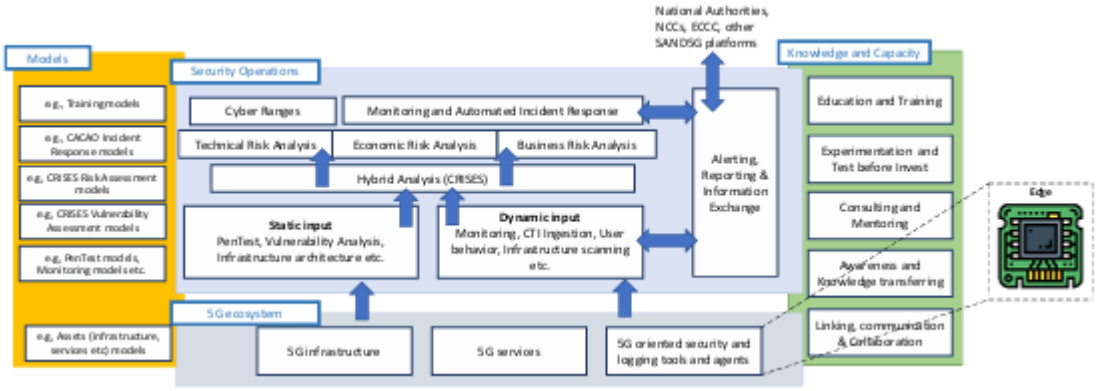


Figure 1: SAND5G Architecture

The Security Operation layer provides two main functions: (a) actively protecting the 5G system and (b) performing real-time technical and financial risk assessments. To achieve this, the platform collects and analyzes both static data, like architecture models and vulnerability assessments, and dynamic data, such as real-time network monitoring, Cyber Threat Intelligence (CTI), user behavior, and infrastructure changes.

The Models enable a customizable security platform for 5G and future networks with interoperable, open solutions. In the SAND5G architecture, functions like vulnerability analysis, risk assessments, and incident response are based on predefined models, allowing seamless integration of changes without reconfiguration.

The Knowledge and Capacity layer provides supportive services that extend beyond individual 5G systems or infrastructures, aiming to enhance the skills, awareness, and collaboration within the broader EU 5G ecosystem.

Considering the 5G security tools that the UPAT team is responsible for, one of the major tasks is the efficient implementation of an Intrusion Detection System (IDS) adaptable to Edge computing. This is because emerging 5G networks enable lower latency, higher capacity, and increased bandwidth compared to previous networks. This means that an IDS (especially the distributed type) in edge computing shares workloads across multiple components. This allows not only high scalability and reliability but also fast processing and quick response times (Alsubhi, K., A Secured Intrusion Detection System for Mobile Edge Computing. 2024). Therefore, an extension of the SAND5G research will focus on an efficient hardware acceleration of an IDS platform. For this purpose, many hardware accelerators will be examined (e.g., Google Edge TPU, Raspberry Pi 3, FPGA).

2. References

[1] Bantouna A., Kitsos P., Lampropoulos K., Poulios K., Qaise O., Raptis L., Stamoulis A., Tranoris C. (2024). Security Assessments for Networks and Services in 5G Networks. In Proceedings of the 27th Euromicro Conference Series on Digital System Design (DSD). IEEE.

[2] Alsubhi, K.(2024). A Secured Intrusion Detection System for Mobile Edge Computing. Appl. Sci. 2024, 14, 1432. https://doi.org/10.3390/app14041432.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 179-180

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Nanotechnology processes with atomic-scale precision: Access to new tools and clean-room infrastructure for researchers and the Industry

C. Zormpa[1], S. Mouchtouris[1,2], V. E. Vamvakas[1] and E. Gogolides[1,*]

[1] *Institute of Nanoscience and Nanotechnology, NCSR "Demokritos", Agia Paraskevi, 15341, Greece*
[2] *School of Chemical Engineering, National Technical University of Athens, Athens, 15780, Greece*
e.gogolides@inn.demokritos.gr

The accelerated research and advancements in the field of nanotechnology, nanodevices and novel materials, require progressively more elaborate processes and precise dimensional control over the materials handling, reaching to the atomic-scale. For fabricating and integrating inorganic nanosystems, deposition equipment, lithography, and etching need to be available at the same place, and be able to handle the same material palette, and the same sample size. No such capability to its full extent exists today in the research infrastructure of Greece. To this end, under the "PlasmaForNano" project[*****], a new clean-room infrastructure and tools were integrated, to extend the capacity of the existent facility of the Institute of Nanoscience & Nanotechnology (INN) of the National Centre for Scientific Research "Demokritos" (NCSR). The infrastructure will be available for access to research and industrial users free of charge, from December 2024 to March 2025, and further on, via the innovation-el[†††††] network with a no-profit access cost.

Targeting metallic and inorganic material etching of large samples with precise dimensional control at the atomic level, the laboratory invested to a cutting edge ICP etcher. This etching tool has the ability to remove and pattern metals, oxides, and polymers for micro/nano system fabrication under the appropriate chlorinated and fluorinated gas chemistries, in inductively coupled plasma (ICP), reactive ion etching (RIE) and atomic layer etching (ALE) modes, all in one system. The technologically advanced technique of ALE, functions under sequential steps of self-limiting reactions, resulting in layer-by-layer material subtraction instead of addition. An embedded optical end-point allows for automatic control of the etching process.

For the case of thin film metal deposition, a new electron-beam (e-beam) evaporation system was attained, with the ability of thermal evaporation as well. The evaporation system is capable of co deposition (alloys) and single or multi-layer metal deposition, with the advantage of e-beam's lower degree of contamination from the crucible and capability of melting high boiling point metals (refractory). A quartz crystal microbalance (QCM) sensor mounted over each source, enables careful monitoring of the deposition rate and final thickness of the layer, in situ. This tool is ideal for lift-off processing, a very important technique to the fabrication of intricate systems due its non-conformality.

[*****] PlasmaforNano information: https://www.plasmafornano.eu/
[†††††] Innovation-el information: https://innovation-el.net/

Another evaporation tool was integrated to complement the existing deposition abilities of the clean-room facilities. The second evaporator system contains a thermal source for metal deposition, and a low-temperature thermal source. The low-temperature evaporation (LTE) system is temperature controlled by an embedded thermocouple, and its heating source is conveniently designed to receive powders and fine material sources, capable of depositing temperature-sensitive, non-toxic materials. The system adds a unique opportunity to work with novel materials and exploit their capabilities in patterned structures and nanosystems integration.

The National Centre of Scientific Research “Demokritos” is the host Institution, and the infrastructure was funded by the Hellenic Foundation for Research and Innovation (HFRI).

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 181-182

Proceedings of Emerging Tech Conference: Edge Intelligence 2024

PerCV.ai Platform: Leveraging Cloud and Edge Computing to Build Vision AI Solutions at Scale

Thomas Charisis, Dimitris Kastaniotis, and Christos Theocharatos

Irida Labs, Kastritsiou 4, 26504 Patras, Greece
{charisis.thomas,kastaniotis,htheohar}@iridalabs.com

Abstract

Edge Vision AI – which processes images to produce analytics and trigger actions – has enormous potential for sectors such as manufacturing, logistics and smart cities. The significant technical challenges associated with Vision AI have made it hard, if not impossible, for many systems integrators and resellers to deliver. Using PerCV.AI, a platform that streamlines the development lifecycle for vision AI-based solutions, cameras can be transformed into smart, AI-powered devices, in easy and quick steps.

1 Introduction

Vision AI is a pioneering technology, in which a video is processed by an onboard AI application that analyzes the visual data and produces an output in the form of analytics data or triggers. In this way, Vision AI uses AI to transform a camera into an AIoT or vision sensor. This is a paradigm shift in how cameras work and what they can be used for. Instead of a human watching a video, an onboard AI application can monitor it, interpreting the images, making decisions, and acting accordingly. Without the need to constantly watch the video, the AI application replaces human effort by discarding nearly all unnecessary data before leaving the camera. However, Vision AI technology comes with a unique product development complexity. Adopters soon have to navigate through a fragmented and multi-disciplinary ecosystem that involves system design, data campaigns, AI/ML models, edge SW optimisation, edge device management and more. According to McKinsey & Company, 74% of Vision AI projects fail to scale and do not meet expectations. For that reason, PerCV.ai is introduced, an end to-end Vision AI platform that addresses the challenge of simplifying Vision AI solution development.

2 PerCV.ai Vision AI Platform

PerCV.ai (pronounced "Perceive AI"), Irida Labs' flagship product, is a groundbreaking Computer Vision AI platform that simplifies the entire computer vision development process. The platform offers seamless end-to-end support, bringing together all the necessary building blocks for building and scaling up Vision AI products. It is a comprehensive Vision AI platform designed to streamline the development and deployment of advanced vision-based applications, allowing developers to create a virtual representation of physical environments (vision twin), simulate use cases, and select appropriate camera modules and sensors, all without requiring extensive data. As a centralized hub,

it facilitates the management and processing of large datasets, enabling tasks like annotation, segmentation, detection, classification, labeling, 2D/3D data synthesis, and analysis. The platform is dynamic, supporting continuous adaptation to scene and object changes by offering a complete Vision AI software pipeline for building, combining, testing AI models, and implementing rapid iterations and over-the-air updates. It also enables swift validation of business cases and supports small-scale deployments with quick feedback loops and direct connections to edge devices. Additionally, PerCV.ai scales efficiently to accommodate thousands of deployments, ensuring broad and scalable solutions for diverse applications.

Figure 1: Left: PerCV.ai platform building blocks. Right: Workflow for building vision AI solutions.

PerCV.ai operates as a robust Cloud SaaS platform, offering a wide range of functionalities to streamline the Vision AI lifecycle. In the cloud, the platform excels in data management, AI model optimization, and device management, including performance assessment. It supports over-the-air (OTA) updates, API integrations, and provides comprehensive metadata and analytics capabilities. These features ensure that the system remains agile and can rapidly adapt to new requirements or updates. The Edge Vision AI SW runs directly on edge devices, ensuring that metadata communication and API functionalities are executed locally. This setup allows for real-time processing and immediate response to environmental changes while maintaining connectivity with the cloud for higher-level data management and model optimization. The platform comes also with Industry 4.0 and Smart City Suites, which include off-the-shelf vision AI solutions for fast and easy application deployments. Populat solutions comprise both suites, such as packaging quality inspection, warehouse monitoring, QR scanning, product counting, PPE monitoring, vehicle tracking and recognition, people flow monitoring, ANPR, smart parking and more. Of course, PerCV.ai also provides a seamless and streamlined toolset for the deployment of custom vision AI solutions that address specific requirements.

Figure 2. User Interface of the simplified version of PerCV.ai platform.

Cutting through the need for investing heavily in R&D across a wide interdisciplinary range of expertise, PerCV.ai helps Irida Labs partners leverage the Vision AI adoption and transform image and video data into valuable, actionable insights that address their real-life business challenges.

EMERGING TECH CONFERENCE – Edge Intelligence

Volume 03, 2024, pages 183-184

Proceedings of Emerging Tech Conference:
Edge Intelligence 2024

Exploiting the triboelectric phenomena for energy harvesting and sensing

Christos Tsamis

Institute of Nanoscience and Nanotechnology, National Center for Scientific Research "Demokritos"
c.tsamis@inn.demokritos.gr (For Poster presentation)

Abstract

Since the pioneering work by Fan et al. [1] on triboelectric nanogenerators (TENG) for the conversion of mechanical energy to electrical energy, triboelectric phenomena have been identified as an attractive mechanism not only for the scavenging of ambient energy but also for sensing, for the manipulation of electronic devices, and for neuromorphic electronics. The power of triboelectric phenomena stem from the fact that their underlying mechanisms, of contact electrification and electrostatic induction between two surfaces that are in relative motion, can appear in all materials if they have appropriate properties while their triboelectric signal depends on their separation in the triboelectric series. Furthermore, with the coupling of triboelectricity and semiconductors, the exciting era of tribotronics has emerged. In this work we will present different applications that exploit the triboelectric effect for energy harvesting and sensing. A triboelectric tile (Fig 1a) as well as a Rotational-To-Linear TENG [2] will be presented for harvesting energy from walking and rotational energy, such as the wind energy due to the rotation of an anemometer. In addition, a flexible tactile sensor (Fig 1b) that can identify different materials as well as surface roughness's, suitable for artificial haptic sensation will be presented [3].

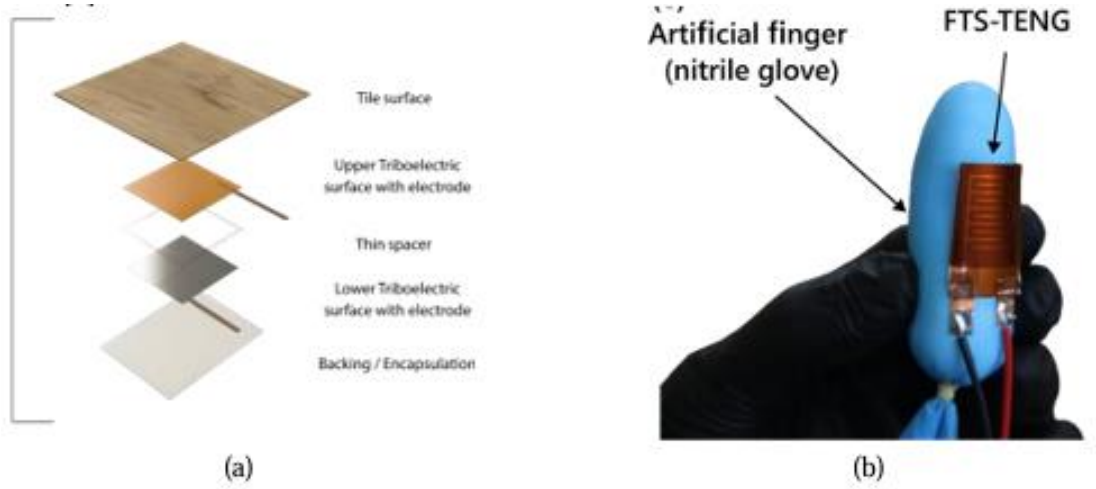


Figure 1: (a) Triboelectric tile for harvesting walking energy and (b) Tactile sensor for material recognition and roughness identification

[1] F. R. Fan et al. Flexible Triboelectric Generator. Nano Energy 2012, 1, 328 334, doi:10.1016/j.nanoen.2012.01.004.

[2] A. Bardakas at al. Zinc Oxide-Based Rotational–Linear Triboelectric Nanogenerator, Appl. Sci. 2024, 14, 2396. https://doi.org/10.3390/app14062396

[3] V. Zacharia et al. Development of a Flexible Tactile Sensor for Material and Texture Identification Based on the Triboelectric Effect, Nano Energy 2024,127, 1–25. https://doi.org/10.1016/j.nanoen.2024.109702

Author Index